IMPERIAL
Dyson School of Design Engineering

Auditory Model-Based Similarity Metric for Head-Related Transfer Functions

Rapolas Daugintis
2025

Thesis submitted for the degree of Doctor of Philosophy (PhD)

Abstract

The head-related transfer function (HRTF) captures direction-dependent acoustic features shaped by the listener's anatomy that give rise to personal spatial auditory cues. While obtaining individual HRTFs at scale is impractical, studies on HRTF personalisation or user adaptation within binaural spatial audio applications require a meaningful metric to assess HRTF similarity. This is a challenging task due to the multimodal nature of auditory perception.

To address this, the thesis proposes an auditory model-based metric for evaluating non-individual HRTFs based on predicted localisation performance. The metric employs an existing Bayesian sound localisation model to compare non-individual HRTFs against the individual ones by matching their auditory cues. The thesis first details the model calibration and metric analysis using prior data from human localisation and HRTF rating studies, followed by perceptual validation through a series of listening tests.

A static sound localisation experiment demonstrates that the metric can reliably identify well-matched non-individual HRTFs that provide localisation performance similar to that of individual HRTFs, as well as poorly matched ones that significantly impair localisation. A dynamic spatial audio quality assessment further reveals that degraded localisation with non-individual HRTF correlates with perceived difference in overall rendering quality and tone colour, but not with externalisation or naturalness. Final studies evaluate the metric in a masked speech comprehension task using binaural rendering in reverberant conditions. While HRTF matching does not directly influence intelligibility, performance varies with reverberation rendering type and depends on acoustic HRTF features.

Overall, the validated similarity metric provides a perceptually grounded tool for assessing system-to-user and user-to-system HRTF adaptation strategies. However, perceptual evaluations suggest that the importance of HRTF personalisation depends on the listening scenario. These findings offer insights into potential optimisation of binaural audio personalisation methods.

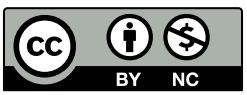
Statement of Originality

The thesis is my own original work, except where otherwise stated, produced under supervision of Prof. Lorenzo Picinali and Prof. Michele Geronazzo. Parts of this thesis have been adapted from previously published journal and conference publications, in accordance with relevant copyright agreements. These instances are indicated at the beginning of the respective chapters or sections. My contributions to each publication are listed in the List of Publications. Generative AI tools were used solely to proofread and revise the text without influencing its underlying content.

Rapolas Daugintis

September 2025

Copyright Statement



Acknowledgements

First, I would like to thank my supervisors, Prof. Lorenzo Picinali and Prof. Michele Geronazzo, for their unwavering support throughout the PhD process. While guiding me through different academic paths and allowing me to find my own, you were always present and ready to help. I would also like to thank Prof. Christoph Pörschmann and Dr Rebecca Stewart for agreeing to be my thesis examiners.

Thank you, Roberto, for showing me the world of auditory modelling, Kat, for all the statistics lessons, Julie and Pedro, for following me from Helsinki to London, Ollie, for all the lab support, Kathi, for your enthusiasm for DTFs, Neel, for sharing the PhD journey, Aidan, Becky, Fulvio, Isaac, Jack, Johan, Kevin, Ludo, Nicola, Nils, Oscar, Pierre, Rahul, Rifqi, Tianxiao, Thibault, Vincent, Yushan, Zyque, the SAP group, and everyone else who I've crossed paths with at the Audio Experience Design team. I am also grateful to have met so many great researchers through the SONICOM project, and I thank members of the Imperial Research Impact Management Office for keeping good company during the annual meetings.

I would like to thank Olivier Warusfel for hosting me at IRCAM and Benoît for the collaboration and conversations about everything from anisotropic reverb to tea. I am also thankful to the researchers at L'Institut ∂'Alembert, at Sorbonne University, especially Brian F. G. Katz and David Poirier-Quinot, for very valuable discussions during my time in Paris, and Sarabeth, Nolan, Gonzalo, and Peter for that short-lived Wednesday Social Club. I am grateful to the Imperial Global Fellows Fund, supported by the UK's Turing Scheme, for sponsoring my stay there.

I am thankful to the past and present members of Aalto Akulab for being a great network of colleagues across the globe. I am also happy to have become an accidental founding member of the Dead Acousticians Society, together with Fabian, Pedro, Nils, Antonio, Marco, Annika, Kevin, and Felix, and I hope that tradition stays on.

Thank you, Karolina, Ieva, and Calvin, for being the best friends I could possibly ask for during my journey from Edinburgh to London, and Lachlan, Magdalena, Raminta, and Francesco for joining our family on the way. Also, Ieva, I am glad we are doing this together!

Noriu padėkoti savo šeimai: Simonai, Mariui, sesėms Emilei ir Agnei – už betarpišką palaikymą, močiutėms Jonei ir Daivai – už tai, kad užaugau kažkur tarp mokslo ir meno, seneliui Stasiui, kuris nebesulaukė mano apsigintos doktorantūros, bet dėl kurio išugdyto smalsumo aš čia ir esu, ir seneliui Rimantui, kurio niekada nesutikau, bet kurio kultūrinis palikimas mane seka ligi šiol. Na, ir ačiū visai likusiai Daugінčių bei Vaičiūnų giminei už mūsų nuostabiai tvirtą ryšį.

Finally, thank you, Matt, for witnessing the ups and downs of my PhD journey and for being my kindest supporter in every step.

List of Publications

I. **Engel, I., Daugintis, R., Vicente, T., Hogg, A. O. T., Pauwels, J., Tournier, A. J. and Picinali, L. (2023). 'The SONICOM HRTF Dataset'.** ***J. Audio Eng. Soc. (AES)*** **71.5, pp. 241–253. DOI:** 10.17743/jaes.2022.0066

Parts of the publication authored by me are adapted in § 4. I helped build the HRTF measurement setup and measured over 50 subjects. I performed HRTF measurement consistency analysis using a computational sound localisation model and wrote the respective part of the paper. © 2023 AES, reproduced with permission (see Appendix A).

II. **Daugintis, R., Barumerli, R., Picinali, L. and Geronazzo, M. (2023b). 'Classifying Non-Individual Head-Related Transfer Functions with a Computational Auditory Model: Calibration and Metrics'. In:** ***Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*****. Rhodes, Greece. DOI:** 10.1109/ICASSP49357.2023.10095152

The publication is reproduced in § 6.1–§ 6.4. I conceptualised the metric together with other coauthors and performed the auditory model calibration based on the code provided by Roberto Barumerli. I conducted the metric analysis and wrote the paper under the guidance from the coauthors. © 2023 IEEE, reproduced with permission (see Appendix A).

III. **Daugintis, R., Barumerli, R., Geronazzo, M. and Picinali, L. (2023a). 'Initial Evaluation of an Auditory-Model-Aided Selection Procedure for Non-Individual HRTFs'. In:** ***Proc. Conv. EAA Forum Acusticum*****. Turin, Italy. DOI:** 10.61782/fa.2023.0489

The publication is not reproduced in the thesis because it presents a pilot study, which was later improved and presented in publication **VIII**. This is discussed in § 7. I developed and conducted the listening test and wrote the paper under the guidance from the coauthors.

IV. **Daugintis, R., Alary, B., Geronazzo, M. and Picinali, L. (2024). 'Effects of Binaural Rendering Personalisation and Reverberation on Speech-on-Speech Masking'. In: *Proc. Audio Eng. Soc. (AES) Conf. Audio Virtual & Augmented Reality*. Redmond, WA, USA. URL:** https://aes2.org/publications/elibrary-page/?id=22671

The publication is reproduced in § 8. I conceptualised the experiment with assistance from Benoît Alary during a research visit to IRCAM, Paris, hosted by Olivier Warusfel. I designed and performed the listening tests and wrote the paper under the guidance from the coauthors. © 2023 AES, reproduced with permission (see Appendix A).

V. **Daugintis, R., Geronazzo, M. and Picinali, L. (2025b). 'On Comparing Auditory Models and Perceptual Assessment When Rating Head-Related Transfer Functions'. In: *Proc. Conv. EAA Forum Acusticum EuroNoise*. Málaga, Spain, pp. 3421–3424. DOI:** 10.61782/fa.2025.0643

The publication is reproduced in § 6.6. I conducted the analysis and wrote the paper under the guidance from the coauthors. © 2023 Rapolas Daugintis et al., under CC BY 3.0.

VI. **Hogg, A. O. T., Barumerli, R., Daugintis, R., Poole, K. C., Brinkmann, F., Picinali, L. and Geronazzo, M. (2025). 'Listener Acoustic Personalisation Challenge - LAP24: Head-Related Transfer Function Upsampling'. *IEEE Open J. Signal Process.* 6, pp. 926–941. DOI:** 10.1109/OJSP.2025.3588776

The publication is summarised in § 5.3. I helped organise the challenge, prepared the challenge dataset, and conducted some of the analysis of the winning solution. © 2025 The Authors, under CC BY 4.0.

VII. **Daugintis, R., Barumerli, R., Geronazzo, M., Pauwels, J., Picinali, L. and Poole, K. C. (2025a). 'Listener Acoustic Personalisation Challenge—LAP24: Head-Related Transfer Function Dataset Harmonization'. *IEEE Open J. Signal Process.* 6, pp. 950–964. DOI:** 10.1109/OJSP.2025.3592601

The parts of the publication authored by me are adapted in § 3.2, § 3.3, and § 5.2. I helped develop the challenge evaluation and led the writing of the paper. I wrote most parts of the paper and incorporated contributions from the coauthors. © 2025 The Authors, under CC BY 4.0.

VIII. **Daugintis, R., Geronazzo, M., Poole, K. C. and Picinali, L. (2026). 'Perceptual evaluation of an auditory model–based similarity metric for head-related transfer functions'. Under review**

The text from the publication is adapted in § 1, § 2.3.1, and § 7. I designed and conducted the listening tests, analysed the results, and wrote the manuscript under the guidance from the coauthors. The training protocol, used in the sound localisation test, was developed together with Katarina C. Poole. The paper was submitted to the Journal of the Acoustical Society of America and is currently undergoing a peer review process.

IX. **Poole, K. C., Meyer, J., Martin, V., Daugintis, R., Marggraf-Turley, N., Webb, J., Pirard, L., Magna, N. L., Turvey, O. and Picinali, L. (2025). 'The Extended SONICOM HRTF Dataset and Spatial Audio Metrics Toolbox'. In: *Proc. Conv. EAA Forum Acusticum EuroNoise*. Málaga, Spain, pp. 6483–6486. DOI:** 10.61782/fa.2025.0864

The text from the publication is not included in the thesis but the extension of the SONICOM HRTF dataset is discussed in § 4. I contributed to the extended dataset by measuring a subset of HRTFs, maintaining the measurement facility, and training new personnel to perform measurements.

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

AIC Akaike Information Criterion

AMT Auditory Modelling Toolbox

AR augmented reality

BEM boundary element method

BRIR binaural room impulse response

CI confidence interval

CRM coordinate response measure

CTF common transfer function

DCN dorsal cochlear nucleus

df degrees of freedom

DRR direct-to-reverberant ratio

DTF directional transfer function

ERB equivalent rectangular bandwidth

FDTD finite-difference time-domain

FEM finite element method

HRIR head-related impulse response

HRTF head-related transfer function

IACC interaural cross-correlation

IC inferior colliculus

ICC intraclass correlation

ILD interaural level difference

IPD interaural phase difference

IQR interquartile range

IR impulse response

ITD interaural time difference

JND just noticeable difference

KEMAR Knowles Electronics Manikin for Acoustic Research

LAP Listener Acoustic Personalisation

LE RMS lateral error

LL log-likelihood

LRT likelihood-ratio test statistic

LSD logarithmic spectral distance/distortion

LSO lateral superior olive

MFCC mel-frequency cepstral coefficients

ML machine learning

MS mean square

MSE mean squared error

MSO medial superior olive

PAE RMS polar angle error

PE RMS polar error

pp percentage point

QE quadrant error rate

RM ANOVA repeated measures analysis of variance

RMS root mean square

RT reverberation time

SAQI Spatial Audio Quality Inventory

SC superior colliculus

SD standard deviation

SE standard error

SNR signal-to-noise ratio

SOC superior olivary complex

SOFA Spatially Oriented Format for Acoustics

SRIR spatial room impulse response

SRM spatial release from masking

SRT speech reception threshold

SS sum of squares

VR virtual reality

Chapter 1

Introduction

1.1 Motivation

Sounds in the physical world are inherently spatial. They originate from sources with physical dimensions and specific positions, and they propagate through space according to physical constraints. Consequently, the human auditory system perceives physical sounds as spatial, detecting incoming sounds via two ears and interpreting spatial characteristics of their sources from acoustic cues present in the incoming sound to form a corresponding mental auditory scene (Blauert 2013).

Advances in audio technologies over the last century have enabled the reproduction and synthesis of arbitrary sounds using signal processing techniques and electroacoustic transducers, such as wearable headphones, without requiring equivalent acoustic sources. However, the sounds produced by headphones, which are placed directly over the ears or in the ear canals, bypass the natural acoustic path of real-world sources, lacking the directional cues essential for spatial perception. Traditional stereo panning can provide limited horizontal spatialisation, but headphone listening is often associated with an internalised (inside-the-head) sound image, which can reduce immersion compared to real-world listening (Roginska and Geluso 2018).

Recently, binaural (headphone-based) spatial audio reproduction has gained significant attention, driven by applications in multi-talker teleconferencing, spatial music and movie audio streaming, as well as virtual and augmented reality (VR/AR). In these systems, a spatial sound scene is created virtually by convolving audio signals with a direction-dependent head-related impulse response (HRIR). HRIR—more commonly referred to by its frequency-domain equivalent head-related transfer function (HRTF)—is formed of filters encapsulating the missing directional acoustic cues, that would arise from the interaction of the incoming sound with the human ears,

head, and torso (Zhong and Xie 2014).

Because these cues depend on an individual anatomy, each auditory system learns to interpret its unique set of cues for accurate spatial perception (Schnupp et al. 2010). Therefore, individual HRTFs are generally required to achieve perceptually authentic virtual sound scenes (Brinkmann et al. 2017). Using non-individualised HRTFs can lead to more ambiguous perceived sound source location (Middlebrooks 1999b; Møller et al. 1996; Romigh and Simpson 2014; Wenzel et al. 1993) and reduced realism of rendered scenes (Jenny and Reuter 2021).

Although individual HRTFs can be acquired acoustically (Li and Peissig 2020) or computed from high-resolution 3D head scans using numerical wave equation solvers (Brinkmann et al. 2023), these approaches are not feasible on a mass scale due to time and resource requirements. Instead, various HRTF personalisation techniques have been proposed, some of which match the listener with an HRTF from an already measured dataset (Fantini et al. 2025; Guezenoc and Seguier 2018). For example, these may be based on a perceptual listening test (Katz and Parseihian 2012; Seeber and Fastl 2003; Zagala et al. 2020), or an anthropometric ear feature matching (Arbel et al. 2024; Geronazzo et al. 2019; Warnecke et al. 2022; Zhi et al. 2022; Zotkin et al. 2003).

However, when large HRTF datasets are considered, behavioural personalisation approaches can be time-consuming (Katz and Parseihian 2012), and are known to produce inconsistent results, especially for naïve listeners (Andreopoulou and Katz 2016; Kim et al. 2020). Numerical approaches offer an alternative by estimating the best-matching non-individual or simulated HRTF based on a similarity metric. However, defining such a metric is a non-trivial, multidimensional problem because it must reflect the perceptual sensitivity of the human auditory system (Ananthabhotla et al. 2021; Doma et al. 2023b; Yao et al. 2024).

At the same time, research has shown the ability of the human auditory system to adapt to non-individual HRTF cues using various sound localisation training routines (Picinali and Katz 2023), suggesting that non-individual HRTFs could be used in various applications to promote long-term user-to-system adaptation and perceptual improvement of binaural audio rendering (Berger et al. 2018; Lladó et al. 2024b). Nonetheless, previous studies have indicated that while the localisation performance with perceptually well-fitting HRTF can approach that of individual HRTFs after a few short training sessions (Parseihian and Katz 2012), the adaptation to a perceptually poorly-fitting HRTF is slower and might not be successful for some listeners (Stitt et al. 2019). Consistent with these findings, Trapeau et al. (2016) observed a correlation between the degree of spectral distortion introduced by physical ear moulds that altered pinna geometry and the rate of adaptation. Conversely, other studies have shown that the success of short

adaptation sessions might not be specific to the HRTF used and thus generalisable to other HRTFs not used during the training, although the results vary across subjects (Meyer and Picinali 2025b; Steadman et al. 2019). This variability highlights the need for further investigation into the relationship between initial HRTF fit and adaptation success.

An increase in mainstream applications for HRTF-based binaural audio rendering technology, where individual HRTFs are not available, demands a more systematic investigation of system-to-user (HRTF personalisation) and user-to-system (listener training) HRTF adaptation procedures (Picinali and Katz 2023) and their perceptual impact on a variety of tasks. Such investigation requires a robust, perceptually motivated numerical similarity metric that incorporates our current understanding of human auditory processing with the flexibility of the numerical analysis. Numerous efforts have already been made to model human spatial auditory processing numerically (Majdak et al. 2022). The proposed computational auditory models aim to consolidate empirical knowledge and theoretical hypotheses about the auditory system, providing a quantitative and adaptable framework for studying perception (Meddis et al. 2010). These models are typically developed and calibrated to reproduce psychoacoustic findings, such as sound-source localisation performance (Middlebrooks 1992). When applied to HRTF personalisation, they can serve as proxies for human listeners in evaluating numerical HRTF matching quality (Geronazzo et al. 2018), enabling more efficient iterative algorithm development. However, current models cannot fully capture the full complexity of human auditory perception. Therefore, their validity in predicting non-individual HRTF fit has to be investigated through perceptual listening tests before they can be reliably used in further HRTF personalisation research.

Recently, a new Bayesian sound localisation model has been developed with the aim of combining observations of auditory features, extracted from given HRTFs, with the prior beliefs of the auditory system to obtain a human-like localisation performance for both horizontal and vertical dimensions (Barumerli et al. 2023). This recent methodological advancement in modelling spatial auditory perception, combined with the increasing availability of large HRTF datasets, driven by the data demand of machine learning (ML) personalisation algorithms (Engel et al. 2023; Fantini et al. 2025), presents a timely opportunity for a more detailed investigation into the perceptual HRTF-based spatial audio rendering quality. Such analysis could help better understand the limitations of HRTF training as well as inform non-individual HRTF selection procedures.

1.2 Aims and Objectives

Bayesian sound localisation model, proposed by Barumerli et al. (2023)[1], provides the groundwork for this thesis to investigate the perceptual HRTF similarity. Using this model, the thesis aims to establish a link between numerical and perceptual analyses of HRTFs and examine how numerical differences in non-individual HRTFs influence perceptual experience across various spatial audio scenarios.

This thesis addresses the following research question: 'Can a Bayesian auditory model be used to predict the perceptual fit of non-individual HRTFs, and does the predicted HRTF fit correlate with performance in spatial audio tasks?'. The methodological foundation of this thesis is the development of an HRTF similarity metric derived from the proposed sound localisation model, which enables the classification of non-individual HRTFs into perceptual categories based on their similarity to an individual HRTF and selection of one well-fitting and one poorly-fitting HRTF. The two perceptually distant non-individual HRTFs then form the basis for investigating the impact of HRTF features across various spatial listening scenarios. Given the complexity of defining a continuous perceptual scale for non-individual HRTFs, focusing on perceptual extremes simplifies experimental design and reduces uncertainty, while still capturing the general effects of HRTF filtering.

In summary, the thesis is structured around three main objectives:

Objective 1: Investigate the use of the proposed Bayesian auditory model for evaluating perceptual HRTF quality in binaural synthesis. This objective is addressed by presenting two applications of the model: in § 4 the model is applied to assess the consistency of the HRTF measurements and in § 5 it forms part of the metric to assess the quality of HRTF harmonisation and upsampling.

Objective 2: Building on the use of the Bayesian auditory model in assessing HRTFs, develop a metric capable of distinguishing between perceptually well-fitting and poorly-fitting non-individual HRTFs by comparing them to an individual HRTF. The metric development is presented in § 6.

Objective 3: Design and conduct perceptual listening tests using individual, well-fitting, and poorly-fitting non-individual HRTFs to evaluate the proposed metric and assess the perceptual impact of HRTF fit in various spatial audio scenarios. This objective

[1]The model was first realeased in the Auditory Modelling Toolbox (AMT) version 1.0.0 in May 2021. However, the associated journal article was released in 2023 after the peer-review process. The thesis refers to the model as *barumerli2023* following the latest AMT version. Nonetheless, the model, originally available as *barumerli2021*, was used from the beginning of this doctoral project.

is addressed in § 7, where HRTFs are evaluated using sound localisation and spatial audio quality assessment, and in § 8, where they are evaluated in a speech intelligibility task.

1.3 Thesis Contributions

This section outlines the primary contributions of this thesis to the fields of spatial audio and psychoacoustics. Each contribution is accompanied by the relevant publications produced during the course of this research. The publication numbers in Roman numerals—used here and throughout the thesis—refer to the entries in the List of Publications where the author's specific contributions to each publication are detailed.

Contribution 1: It presents measurements, validation analysis, and release of the SONICOM HRTF dataset, currently comprising over 300 measured HRTFs, 3D head scans, and headphone equalisation filters. This dataset is among the largest publicly available collections of acoustically measured HRTFs and serves as a resource for ongoing research in HRTF personalisation. *Contribution presented in § 4. Associated publications: I and IX.*

Contribution 2: It reports on the development of a comprehensive evaluation framework for the inaugural Listener Acoustic Personalisation (LAP) challenge that assessed the perceptual validity of data-driven HRTF harmonisation and upsampling methods through auditory model-based objective metrics. It establishes methodological guidelines for evaluating HRTF manipulation techniques and demonstrates the importance of community-driven benchmarking in advancing perceptually valid HRTF processing methods. *Contribution presented in § 5. Associated publications: VI and VII.*

Contribution 3: It introduces a novel similarity metric for non-individual HRTFs based on a recently developed Bayesian auditory model for sound localisation. The metric's predictions are compared against prior perceptual data and evaluated through listening tests. *Contribution presented in § 6 and § 7. Associated publications: II, V, and VIII.*

Contribution 4: It quantifies HRTF personalisation effects across multiple spatial audio tasks through controlled perceptual experiments. This work establishes the differential impact of HRTF fit on sound localisation, spatial audio quality perception,

and speech intelligibility, informing practical decisions about when individualised HRTFs are necessary. *Contribution presented in § 7 and § 8. Associated publications: III, IV, and VIII.*

1.4 Thesis Structure

This thesis is organised as follows:

Chapter 2 provides background on the human spatial hearing mechanism. It introduces the head-related transfer function (HRTF) and other spatial hearing concepts relevant to this thesis from the psychoacoustics perspective.

Chapter 3 reviews HRTFs from the perspective of acoustics and audio signal processing. It highlights the dichotomy between numerical and perceptual HRTF analyses and presents case studies addressing this gap.

Chapter 4 reports the release of the SONICOM HRTF dataset. It overviews the measurement setup and presents the analysis conducted to validate the consistency of the HRTF measurements. This chapter includes content from publication I.

Chapter 5 discusses some of the current challenges in HRTF personalisation research. Specifically, it summarises the outcomes of the first Listener Acoustic Personalisation (LAP) challenge. This chapter includes content from publications VI and VII.

Chapter 6 introduces a new auditory model-based similarity metric for non-individual HRTFs. It details the calibration of the Bayesian auditory localisation model and evaluates the similarity metric against prior perceptual data from localisation and HRTF rating studies. This chapter is based on publications II and V.

Chapter 7 presents the validation of the proposed similarity metric through perceptual listening tests. It describes the design and results of two experiments for the following tasks: static sound localisation and dynamic spatial audio quality assessment. This chapter is based on publication VIII.

Chapter 8 presents the evaluation of the proposed similarity metric in a speech-on-speech masking task using binaural rendering with reverberation. It outlines the experimental design and results of the listening test. This chapter is based on publication IV.

Chapter 9 summarises the main findings of the thesis and discusses their implications for future HRTF research.

Chapter 2

Human Spatial Hearing

2.1 Introduction

In the physical world, an external sound event from an acoustic source is detected by the human auditory system and gives rise to an internal auditory event (Pulkki and Karjalainen 2015). Humans do not respond to the sound itself, but to the perceptual experience it evokes (Blauert and Braasch 2020). Virtual audio technologies exploit this principle: the physical source of a sound need not be present, as long as the reproduced sound creates an equivalent auditory event for the listener. Through appropriate sound manipulation, it is possible to induce a plausible (Lindau and Weinzierl 2012) or even authentic (Brinkmann et al. 2017) auditory illusion. However, the success of such illusions depends on a combination of multiple factors, including multisensory integration, room acoustics, body motion, and individual listener experience (Brandenburg et al. 2020). While these aspects are crucial for binaural spatial audio, each represents a separate research field, some beyond the scope of this thesis.

Fundamentally, spatial audio technology relies on a solid understanding of the human spatial hearing mechanism. This chapter introduces its key concepts, focussing on human sound localisation and its role in complex everyday scenarios that include listening in reverberant spaces and understanding speech. It also reviews current approaches to modelling spatial hearing that help understand the auditory system and inform the development of virtual audio technologies.

2.2 Human Sound Localisation

The sound that reaches human ears is acoustically shaped by its interactions with the torso, head, and pinnae. Acoustically, this direction-dependent filtering is known as head-related impulse response (HRIR) in the time domain and head-related transfer function (HRTF) in the frequency domain. It gives rise to auditory cues that the auditory system uses to interpret the spatial auditory scene and infer the positions of sound sources around the listener (Blauert 1997; Middlebrooks and Green 1991). This mechanism differs from other senses such as vision, where the negligible diffraction of light reaching the eye (due to its very short wavelengths) allows spatial information to be derived from the topographic mapping of light receptors on the retina (Schnupp et al. 2010, Chapter 5). In contrast, the cochlea maps the incoming sound tonotopically, i.e. by frequency, and sound location is resolved in the central auditory system (Middlebrooks 2015). It is an ill-posed problem for the auditory system, as various combinations of source spectra and positions can produce identical binaural signals at the ear canal. Therefore, the auditory system must rely on certain assumptions to resolve these ambiguities and enable perceptual inference of the sound source location (Majdak et al. 2020).

2.2.1 Binaural Cues

The primary cues for localising sound along the lateral (left–right) axis arise from differences in the acoustic signals received at the two ears, and are therefore referred to as binaural, or interaural, cues. The difference in sound pressure level between the ipsilateral and contralateral ears caused by the acoustic shadowing of the head is known as the interaural level difference (ILD). The cue is predominantly encoded by the lateral superior olive (LSO), located within the superior olivary complex (SOC) of the brainstem, which combines excitatory input from the ipsilateral ear with the inhibitory signal from the contralateral ear (Tollin 2003).

ILDs are most effective for high-frequency sounds above approximately 4 kHz where the level difference reaches 20 dB and more for lateral source positions (Middlebrooks 2015). Correspondingly, neurons in the LSO are biased towards responding to high-frequency stimuli (Tollin 2003). In this frequency range, the auditory system can detect ILDs as small as 0.5 dB (Mills 1960). However, at lower frequencies, particularly when the wavelength exceeds the size of the head, diffraction reduces the ILD, making it an unreliable cue for sound localisation.

Lord Rayleigh (1907) was among the first to document that humans can discriminate between left and right sound source positions for very low-frequency pure tones (below 300 Hz), despite

physically negligible ILDs at these frequencies. He proposed that listeners are sensitive to interaural phase differences (IPDs) at low frequencies, which led to the formulation of the 'duplex theory' of sound localisation. IPDs—more generally referred to as interaural time differences (ITDs)—are effective cues for tonal sounds up to approximately 1.4 kHz (Mills 1958; Sandel et al. 1955). ITD encoding is thought to primarily occur in the medial superior olive (MSO) within the SOC but its precise mechanism remains debated (Middlebrooks 2015). The influential place theory proposed by Jeffress (1948) suggested that the encoding occurs via a series of neural delay lines, each tuned to a specific ITD value. While the theory garnered experimental support over the subsequent years, more recent studies found evidence for diverse ITD encoding strategies across animal species (Ashida and Carr 2011; Joris et al. 1998).

In humans, the maximum ITD for lateral sound sources is around 700 µs, depending on head size. Just noticeable differences (JNDs) for ITDs can be as small as 10–20 µs for pure tones between 700 Hz to 1000 Hz but they increase for lower and higher frequencies (Brughera et al. 2013). At higher frequencies, ITDs become less effective due to phase ambiguity and the inability of auditory nerve fibers to phase-lock to the temporal fine structure of the waveform (Schnupp et al. 2010, Chapters 2 and 5).

While pure tones around 2 kHz are the most difficult to localise (Mills 1958; Sandel et al. 1955), most real-world sound localisation scenarios involve complex broadband sounds. For high-frequency complex sounds, the auditory system can track ITDs in the amplitude envelopes, though with reduced sensitivity (Bernstein and Trahiotis 1994), making the envelope ITD a less dominant cue (Macpherson and Middlebrooks 2002). When localising full-range broadband stimuli that include low-frequency content, the auditory system appears to rely more heavily on ITDs than on ILDs (Folkerts and Stecker 2022; Macpherson and Middlebrooks 2002; Wightman and Kistler 1992). Nonetheless, the plasticity of the auditory system allows for a varied perceptual weighting of these cues depending on their reliability for an individual (Klingel et al. 2025).

2.2.2 Monaural Cues

While many studies on sensitivity to binaural cues are limited to dichotic left–right lateralisation scenarios, realistic localisation tasks involve front–back discrimination and the perception of elevation. In his experiments, Lord Rayleigh (1907) observed that, although left–right discrimination of pure tones was generally straightforward, front and back positions were often indistinguishable to listeners. Binaural cues can only resolve sound source direction along the 'cone of confusion' around the interaural axis: all positions on this cone produce nearly identical

ILDs and ITDs due to the symmetry of the head (Woodworth 1938). To break this ambiguity, the auditory system relies on monaural spectral cues.

Spectral cues arise from interference between the direct sound wave entering the ear canal and multiple reflections—primarily from the pinna—creating direction-dependent frequency patterns. Acoustic measurements and simulations reveal notches whose centre frequencies tend to rise with increasing source elevation, as well as peaks associated with resonant modes of the pinna cavities, typically in the range of 4 kHz to 16 kHz and dependent on individual ear morphology (Middlebrooks 1999a; Raykar et al. 2005; Takemoto et al. 2012). In addition, reflections from torso might provide secondary cues in the lower frequencies above 700 Hz (Algazi et al. 2001b). However, the precise mechanism by which the auditory system analyses these spectral patterns remains uncertain.

Perceptual studies suggest that monaural cues are processed independently from binaural cues (Hofman and Van Opstal 1998; Macpherson and Middlebrooks 2002) but monaural cue contributions from each ear are weighted based on the binaural cues (Hofman and Van Opstal 2003; Macpherson and Sabin 2007). When processing spectral cues, the auditory system must disentangle two unknowns: the original source spectrum and the directional filtering imposed by the outer ear. Elevation perception is therefore most robust with broadband, spectrally flat sources. Narrowband sources don't have enough spectral information and tend to be localised according to their centre frequency (Blauert 1969), which suggests that the auditory system performs a correlation-based comparison with an internal spectral map to estimate the source elevation (Middlebrooks 1992).

Neurophysiological evidence indicates that the dorsal cochlear nucleus (DCN), which is sensitive to spectral notches via an inhibitory mechanism, is responsible for encoding these cues (Schnupp et al. 2010, Chapter 5). Studies with cats have found evidence that the DCN is sensitive to rising spectral edges of HRTF notches or peaks (Reiss and Young 2005), suggesting an equivalent monaural feature encoding mechanism is present in the human auditory system, too (Baumgartner et al. 2014). Nonetheless, while experiments that removed or altered spectral cues generally support the importance of the relative levels of the notches and peaks for robust elevation perception, they disagree on the exact features needed for robust localisation (Asano et al. 1990; Balachandar and Carlile 2019; Iida and Oota 2018; Langendijk and Bronkhorst 2002; Macpherson and Sabin 2013). Perceptual spectral weighting analyses have identified the main notch region around 8 kHz as particularly influential, but also reveal large inter-individual differences (Lladó et al. 2025; Zonooz et al. 2019).

2.2.3 Sound Localisation in Motion

Due to uncertainties in spectral cues, static sound localisation tasks can still lead to ambiguous estimates, particularly in cases of front–back and top–down confusions. To resolve these ambiguities, the auditory system leverages motion to refine the perceived direction of the sound source. Dynamic cues may be passive, arising from movements within the sound scene relative to the listener, or active, generated by deliberate head movements (McLachlan et al. 2021). Despite the near-constant motion of both the sound scene and the listener in real-world environments, the ability to maintain a stable auditory scene relies on a continuous perceptual process. This involves prior beliefs, predictive mechanisms, and ongoing updates to the perceptual representation of the auditory scene (Carlile and Leung 2016).

Changes in ITD and ILD due to head yaw help distinguish front from back positions, as the nature of cue variation depends on the location of the sound source (Thurlow and Runge 1967; Wallach 1940). While dynamic ITD is typically a more dominant cue, variations in ILD can also be informative in certain scenarios (Pöntynen and Salminen 2019). Additionally, head-roll-induced dynamic binaural cues assist in resolving up–down ambiguities (Jiang et al. 2019; Thurlow and Runge 1967). However, small changes in monaural spectral cues resulting from head pitch do not appear to be utilised by the auditory system as dynamic localisation cues (McLachlan et al. 2023).

2.2.4 Neural Mapping of Localisation Cues

Neural signals encoding localisation cues from various regions of the brainstem converge in the midbrain: specifically in the inferior colliculus (IC) (Schnupp et al. 2010, Chapter 5), and subsequently in the superior colliculus (SC), which is involved in sensorimotor integration (Middlebrooks 2015). Animal studies have demonstrated a topographic organisation of neurons in the SC, each maximally sensitive to auditory cues from specific spatial locations (Middlebrooks 2015). In contrast, there is limited evidence for such topographic mapping in the auditory cortex (Middlebrooks 2021). These findings suggest that the spatial map of auditory space is formed primarily in the midbrain, although the auditory cortex also contributes to the successful formation and adaptation of this map (Mendonça 2014).

Auditory localisation cues vary across individuals due to anatomical differences, such as the unique morphology of the outer ear. Consequently, the auditory system learns to interpret these cues during early development and continues to adapt as anatomical changes occur, for example, through ear growth (Schnupp et al. 2010, Chapter 7). Experimental manipulations such as the use

of ear moulds to alter the shape of the pinnae have shown that while localisation performance initially declines, subjects can adapt to the new cues, particularly when active training is provided (Carlile et al. 2014; Carlile and Blackman 2014; Friedrich and Schönwiesner 2025; Hofman et al. 1998; Trapeau et al. 2016; Van Wanrooij and Opstal 2005). However, it remains unclear whether a successful adaptation in these studies indicates a perceptual remapping of the auditory space or procedural learning specific to the task (Mendonça 2014; Meyer and Picinali 2025b).

2.3 Assessing Human Sound Localisation Performance

The ability of humans to localise sound sources is typically evaluated through behavioural listening tests in controlled acoustic environments (Carlini et al. 2024). Various test paradigms have been used to measure sound localisation performance, differing in the task performed by the listener, methods for collecting and analysing the responses, types of stimuli used, and spatial arrangements of sound sources (Blauert 1997, § 1). This section overviews some common approaches, placing emphasis on the methods and metrics adopted in this thesis.

A widely used test paradigm is a static localisation test, where the listener remains stationary while a short sound is presented and the listener's task is to indicate the perceived direction of each sound source (e.g., Brungart et al. 1999; Hofman and Van Opstal 1998; Macpherson and Middlebrooks 2000; Makous and Middlebrooks 1990). Previous studies have employed a variety of methods to capture the perceived direction of the sound source. Many used an egocentric pointing method to register the perceived direction: either pointing with the head (Geronazzo et al. 2019; Majdak et al. 2010; Makous and Middlebrooks 1990; Middlebrooks 1999b; Zagala et al. 2020), with a hand-held pointer (Brungart et al. 1999; Daugintis et al. 2023a; Majdak et al. 2010), or even eye movements (Hofman and Van Opstal 2003; Hofman and Van Opstal 1998). While some studies rely on exocentric techniques, where the direction is reported on a 2D user interface (Begault 2001) or a 3D sphere (Bolia et al. 2001; Gilkey et al. 1995), verbally (Wenzel et al. 1993), or using hybrid approaches (Meyer and Picinali 2025b), egocentric methods are seen as the most intuitive and reliable (Bahu et al. 2016). Different egocentric pointing methods may still have their own biases (Bahu et al. 2016). However, the overall differences may not significantly affect the outcome of the study (Majdak et al. 2010). Once the responses are collected, the performance in static localisation tests is typically quantified using localisation error metrics that measure the deviation between the target sound source directions and the listener's responses. Those, relevant to this thesis are summarised below (§ 2.3.1).

Static sound localisation tests usually employ short bursts of broadband noise, no longer than a few hundred milliseconds, to provide sufficient localisation cues without allowing for head movements (Bahu et al. 2016; Blauert 1997; Vliegen and Van Opstal 2004). Furthermore, a sequence of short burst have been found to improve the localisation accuracy over one burst of the same length (Macé et al. 2012). Other stimuli, for example speech, have also been used (Begault 2001; Meyer and Picinali 2025b). The spatial arrangement of sound sources can vary from a few discrete positions along one dimension (azimuth or elevation) to dense grids covering the entire sphere around the listener (Middlebrooks and Green 1991). The choice of stimuls and spatial arrangement is often guided by the research hypothesis and practical limitations such as the number of loudspeakers available or the time required for testing.

Although seen as the most direct way to assess localisation performance, static sound localisation tests have their limitations. Importantly, untrained participants often have poor static sound localisation skills, leading to high localisation errors and limiting the test sensitivity (Daugintis et al. 2023a). The blind static sound localisation task is not a natural task for human listeners, who often rely on dynamic acoustic and visual cues to determine a precise position of an object. To account for this, the static localisation scenarios often include extensive training sessions to reduce the errors and improve the test sensitivity (Carlile et al. 2014; Carlini et al. 2024; Friedrich and Schönwiesner 2025; Trapeau et al. 2016).

Real-world localisation is significantly enhanced by continuous head movements, which introduce dynamic listening cues, as discussed in § 2.2.3. However, assessing dynamic sound localisation presents greater challenges, requiring careful consideration of appropriate experimental protocols and analytical metrics. In particular, unrestricted dynamic localisation using continuous sound stimuli can generally result in minimal localisation error, as listeners can actively orient themselves towards the sound source. To address this, previous studies have implemented various constraints on head movements, such as limiting the duration of sound stimuli or restricting the range of head rotations (McLachlan et al. 2023, 2025; Thurlow and Runge 1967; Wallach 1940). Others opted for measuring either the response time or the angular distance travelled by the listener while searching for the sound source instead. For example, Perrott et al. (1990) introduced a paradigm in which participants were asked to locate and identify a visual target, examining the decrease in response times when a congruent auditory cue was presented. Subsequent studies have employed this paradigm to evaluate the impact of hearing protection (Simpson et al. 2005) and hear-through headphones (Lladó et al. 2024a) on the audiovisual search performance.

Instead of localisation accuracy, other studies have used discrimination tasks to evaluate

human spatial hearing. Such paradigms often rely on participants comparing the position of the target (probe) stimulus with the reference to, for example, measure the minimum audible angles (Aggius-Vella et al. 2020; Blauert 1997; Mills 1958; Perrott and Saberi 1990). While discrimination tasks provide insight into the sensitivity of the auditory system, they are more reflective of its ability to compare acoustic features of the stimuli rather than mentally spatialise the sound sources (Makous and Middlebrooks 1990).

In summary, the choice of the test paradigm and metrics for assessing human sound localisation performance depends on the specific research question, constraints, and the desired sensitivity of the test. Static sound localisation tests remain a widely used approach, particularly when combined with training protocols to ensure appropriate listener performance. The following section details the localisation error metrics employed in this thesis to quantify performance in a static sound localisation task.

2.3.1 Localisation Perfomance Metrics

Human static sound localisation performance is typically evaluated using localisation error metrics, which quantify the average deviation between target sound source directions and listener responses. Poirier-Quinot et al. (2022) reviews common approaches found across research studies for decomposing target and response directions into coordinate axes and aggregating errors across multiple trials to derive global performance measures. This section introduces the metrics specifically adopted in this thesis, along with their mathematical and algorithmic formulations.

The most general analysis approach is to calculate a great circle error (GCE), which corresponds to the central angle between each target ($\mathbf{r}$) and response ($\tilde{\mathbf{r}}$) directions:

$$\mathrm{GCE} = \arccos\left(\mathbf{r} \cdot \tilde{\mathbf{r}}\right). \tag{2.1}$$

The average GCE across all the positions tested for each participant indicates the global extent of the localisation performance (Poirier-Quinot et al. 2022). However, GCE offers little insight into the nature of the localisation error. It can therefore be useful to split the error based on the auditory cue regions. Following the proposal of Morimoto and Aokata (1984) and its modification described by Pollack et al. (2022), the egocentric unit sphere around the listener can be decomposed into a lateral and polar axis (Figure 2.1). This interaural-polar coordinate system is used to separate the axis along which binaural cues dominate direction perception (described

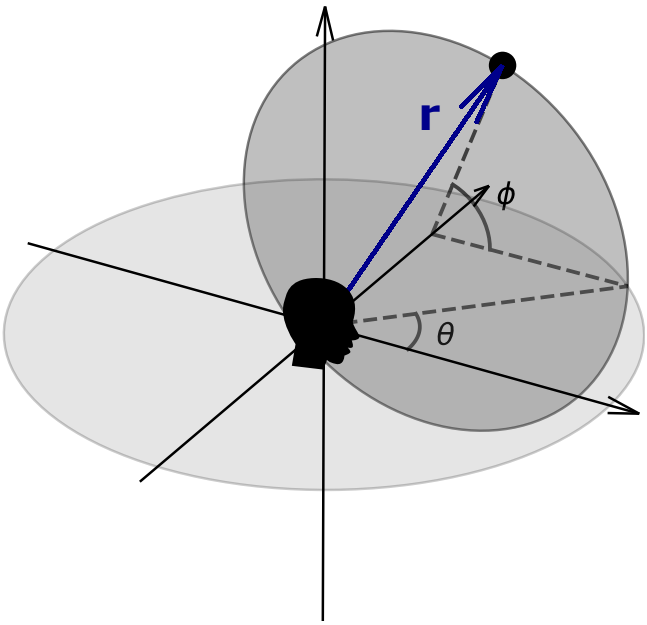


Figure 2.1: Interaural-polar coordinate system. $\mathbf{r}$ represents the direction to the sound source, θ the lateral angle and ϕ the polar angle.

by the lateral angle $\theta \in [-90°, 90°]$ from the median plane to the sagittal plane of interest) from the axis where monaural cues are required (described by the polar angle $\phi \in (-90°, 270°]$ around the sagittal plane).

Within this coordinate system, three error metrics are described by Middlebrooks (1999b): RMS lateral error (LE), RMS polar error (PE) and quadrant error rate (QE). Following the formulation used in Daugintis et al. 2023b, for a set $\mathcal{A} = \{i : \text{wrap}\,|\tilde{\phi}_i - \phi_i| < 90°\}$, which describes local responses (i.e. within 90° polar angle from the target), these errors are defined as follows:

$$\text{PE} = \sqrt{\frac{\sum_{i\in\mathcal{A}} \left(\text{wrap}\left(\tilde{\phi}_i - \phi_i\right)\right)^2}{|\mathcal{A}|}}, \tag{2.2}$$

$$\text{QE} = \left(1 - \frac{|\mathcal{A}|}{N}\right) \times 100\,\%, \tag{2.3}$$

$$\text{LE} = \sqrt{\frac{\sum_{i=1}^{N} \left(\text{wrap}\left(\tilde{\theta}_i - \theta_i\right)\right)^2}{N}}. \tag{2.4}$$

Here, 'wrap' represents angle wrapping to an appropriate range ($[-180°, 180°]$ in this case), N is the total number of trials, $|\mathcal{A}|$ is the number of responses classified as local, and $\tilde{\mathbf{r}} = \left(\tilde{\theta}, \tilde{\phi}\right)$ is the response direction on the unit sphere. LE is defined as the root mean square (RMS) difference in lateral angle between the target (real) direction and the response (simulated by the auditory model or indicated by the listener) direction. PE is then an RMS difference in the polar angle for responses that were within $90°$ of the target direction. QE corresponds to the percentage

of responses that were greater than $90°$ in polar angle. Middlebrooks (1999b) also defines two bias metrics - one on the lateral axis (LB) and one on the polar axis (PB). They represent a signed mean between the target and response positions.

The metrics along the polar axis suffer from a compression issue for directions near the poles on the sides, where a small difference in distance results in a high angle change. Therefore, Middlebrooks (1999b) limited the PE and QE analysis to points within $\pm 30°$ lateral angle (and LE to $\pm 60°$ to avoid angle folding artefacts), which omits a large area of the localisation sphere. To extend the analysis domain, a polar weighting could be used to compensate for angle compression when the source is closer to lateral positions. Hence, Carlile et al. (2014) defined a RMS polar angle error (PAE), by weighing the polar error by the cosine of the target lateral angle:

$$\mathrm{PAE} = \sqrt{\frac{\sum_{i \in \mathcal{A}} \left(\mathrm{wrap}\left(\tilde{\phi}_i - \phi_i\right) \cos\theta\right)^2}{|\mathcal{A}|}}, \tag{2.5}$$

PE and PAE can be seen as local error metrics because they account for responses within proximity of the target positions, while QE indicates the rate of large confusions in the perceived source position. However, the QE metric is limited in its ability to separate local errors from common types of confusion that listeners make. Instead, confusion classification, detailed in Poirier-Quinot et al. (2022), can be used to categorise each error into precision (GCE $< 45°$), front-back confusion (GCE $< 45°$ from the target position mirrored across the frontal plane), in-cone confusion (other responses with lateral angle within 45° from the target) and off-cone confusion (remaining responses). This classification is summarised in Algorithm 2.1. Confusion rates are then calculated as the percentage of trials classified within each category.

Algorithm 2.1 Confusion classification.

if $\mathrm{GCE}\left(\tilde{\mathbf{r}}, \mathbf{r}\right) < 45°$ **then**
    confusion category ← precision
else if $\mathrm{GCE}\left(\left(\tilde{\theta}, \mathrm{wrap}\left(\tilde{\phi} - 180°\right)\right), \mathbf{r}\right) < 45°$ **then**
    confusion category ← front-back confusion
else if $|\tilde{\theta} - \theta| < 45°$ **then**
    confusion category ← in-cone confusion
else
    confusion category ← off-cone confusion

In general, every chosen metric captures a specific aspect of localisation performance, depending on the coordinate system, the range of repsonses considered, and error aggregation method. While no single metric can fully characterise localisation performance, the set of metrics described above provides a comprehensive framework for analysing sound localisation

performance and comparing the results to previous literature.

2.4 Computational Auditory Models for Sound Localisation

Based on experimental sound localisation data with human subjects, as well as neurophysiological recordings obtained from other mammals, various computational auditory models have been developed to understand and predict the mechanisms of human spatial hearing. Accurate models provide a faster alternative to time-consuming listening tests for evaluating or validating new audio processing and rendering algorithms. To support knowledge sharing and reproducibility, the Auditory Modelling Toolbox (AMT) framework was developed to host various auditory models in a unified environment (Majdak et al. 2022). Within this toolbox, several models estimate human sound localisation ability by simulating a virtual 'listener' presented with sounds from specific directions—typically represented by HRTFs—and predicting the perceived source location. These virtual listeners can perform localisation tasks across many directions in a fraction of the time required by human participants. When human localisation inaccuracies are realistically modelled, such models can replace or complement listening tests in perceptual evaluations of spatial audio technologies.

Following the early work of Middlebrooks (1992), many subsequent *functional* models predicting sagittal-plane localisation from spectral features adopted a template-matching principle, mimicking the internal auditory map. Various features have been proposed for comparing the spectral characteristics of the target sound and the internal template, including spectral cross-correlation (Hofman and Van Opstal 1998; Middlebrooks 1992), first- and second-order spectral derivatives (Zakarauskas and Cynader 1993), standard deviation of inter-spectral difference (Baumgartner et al. 2013; Langendijk and Bronkhorst 2002), and positive spectral gradients (Baumgartner et al. 2014). Baumgartner et al. (2013, 2014) and Langendijk and Bronkhorst (2002) also introduced probabilistic components with tunable parameters to replicate human response data. These models have been evaluated and used to assess localisation performance with non-individual HRTFs (Barumerli et al. 2018a; Warnecke et al. 2022).

Building on this, Reijniers et al. (2014) combined template matching with Bayesian inference to propose a full-sphere ideal observer model, exploring the upper limits of localisation from binaural signals. However, this model produces a 'super-human' listener that exceeds realistic human performance (Barumerli et al. 2020). Extending the Bayesian framework, Barumerli et al. (2023) developed a model (referred to as *barumerli2023* in this thesis) that incorporates

human auditory limitations and prior biases to estimate more realistic full-sphere localisation performance based on static binaural and monaural cue matching. This model has been used extensively throughout this thesis and is therefore described in more detail in § 2.4.1. In parallel, McLachlan et al. (2025) extended the static ideal-observer model to account for active sound localisation.

Beyond *functional* models, machine learning (ML) approaches have also been proposed (Jin et al. 2000; Saddler and McDermott 2024). While these models can reproduce human-like responses at the group level, they often struggle to generalise to individual listeners outside their training set, limiting their applicability beyond the original scope.

2.4.1 Barumerli2023 localisation model

Barumerli2023 is a Bayesian model based on the template-matching procedure, in which auditory features of an individual template are compared to noisy features of the target input to provide a human-like localisation response (Barumerli et al. 2023). It is available from the Auditory Modelling Toolbox (AMT) (Majdak et al. 2022). The structure of the model is presented in Figure 2.2 and summarised below.

When provided with a binaural sound as a target input, the model extracts its binaural and monaural features. It uses the maximum of the interaural cross-correlation of a low-passed binaural input to extract ITD, according to Andreopoulou and Katz (2017). The target feature x_{itd} is obtained by transforming the extracted ITD in microseconds to a dimensionless quantity, comparable to other features, according to Reijniers et al. (2014). Interaural level difference (x_{ild}) is calculated from the RMS levels of time-averaged envelopes. Monaural spectral feature arrays ($\mathbf{x}_{mon,L}$ and $\mathbf{x}_{mon,R}$) are extracted by passing the left and right ear signals through a gammatone filter bank with 27 equivalent rectangular bandwidth (ERB) bands (Glasberg and Moore 1990) from 700 Hz, where torso reflections are present (Algazi et al. 2001b), to 18 kHz,

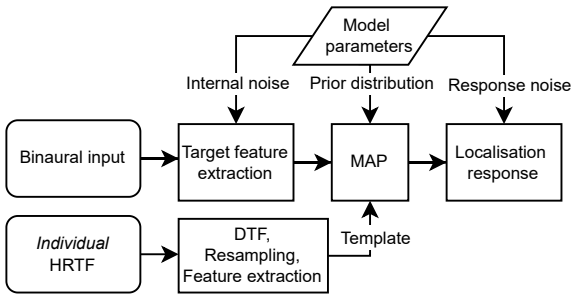


Figure 2.2: *Barumerli2023* auditory model structure. Adapted from Figure 1 in Barumerli et al. 2023.

which corresponds to the upper limit of the hearing range (Baumgartner et al. 2013). A half-wave rectification and compression is used to mimic hair cell activity and obtain an excitation value for each band, similarly to previous models (Baumgartner et al. 2013; Dietz et al. 2011; Roman et al. 2003). While alternative monaural features, based on positive spectral gradient extraction (Baumgartner et al. 2014), are available within *barumerli2023* model setup, spectral magnitudes were used throughout this thesis to stay consistent with the past literature (Barumerli et al. 2018a; Middlebrooks 1992). The obtained input feature vector:

$$\mathbf{t} = \left[x_{itd}, x_{ild}, \mathbf{x}_{mon,L}, \mathbf{x}_{mon,R}\right] + \delta \tag{2.6}$$

is corrupted by multivariate Gaussian noise $\delta \sim \mathcal{N}(0, \mathbf{\Sigma})$ with a diagonal covariance matrix $\mathbf{\Sigma}$, controlled via three independent parameters σ_{itd}, σ_{ild}, and σ_{mon}, to simulate realistic human performances when inferring the sound source direction.

Similarly, a template feature matrix ($\mathbf{T}(\mathbf{r})$) is constructed from the individual HRTF, containing the extracted feature arrays for all possible sound source directions on a dense uniform grid. For this template, individual directional transfer function (DTF) is first obtained from the individual HRTF by filtering it with the inverse of the complex minimum-phase averaged log-magnitude spectra across all its measured directions (known as the common transfer function, or CTF), according to Majdak et al. (2010). This step emphasises direction-dependent high-frequency HRTF features and can reduce direction-independent measurement bias (Baumgartner et al. 2013). While the filtering is limited in its ability to completely remove systemic biases (Daugintis et al. 2025a), it is implemented to provide consistency with previous auditory modelling and sound localisation research (Baumgartner et al. 2013; Majdak et al. 2010; Middlebrooks 1992). Then, individual DTF is resampled to a uniform spherical grid using regularised 15$^{\text{th}}$-order spherical harmonics interpolation, according to Zotkin et al. (2009), to ensure that the template evenly covers the full sphere. Finally, the same features as the target are extracted for every direction of the resampled individual DTF.

The Maximum a-posteriori estimation (MAP) is then employed by the model to compare the target features to the template and determine the most likely direction ($\mathbf{r}$) of the sound source:

$$\tilde{\mathbf{r}} = \arg\max_{\mathbf{r}} p\left(\mathbf{t}|\mathbf{T}(\mathbf{r})\right) p(\mathbf{r}) + \mathbf{m}, \tag{2.7}$$

where the likelihood $p\left(\mathbf{t}|\mathbf{T}(\mathbf{r})\right)$ is weighted by a prior $p(\mathbf{r})$, modelled uniformly across the azimuth but normally distributed for the elevation (with zero mean and standard deviation σ_{prior}) to

represent human localisation bias towards the eye level, according to the findings by Ege et al. (2018). Additionally, the estimate is corrupted by response noise $\mathbf{m}$, representing the uncertainty in the action of pointing towards a sound source during a localisation test, which is modelled as a von Mises-Fisher distribution with a standard deviation of σ_{motor}.

The quantities σ_{itd}, σ_{ild}, σ_{mon}, σ_{prior}, and σ_{motor} can be tuned individually using data from real localisation tests to produce accurate human-like localisation responses. Due to the stochastic nature of the model, a single estimation may not represent the expected human performance. Instead, the estimation for each binaural input must be iterated multiple times and simulated response errors averaged to obtain stable model estimates, comparable to averaged human responses. Barumerli et al. (2023) found 300 model repetitions to produce converging averaged outcomes.

A well calibrated *barumerli2023* model can as a proxy for a human listener, and help understand the localisability of different natural and processed sounds. As demonstrated in the following chapters, by replacing a single binaural input with another (target) HRTF, the model can be used to simulate localisation responses for all its available directions in a fraction of the time needed to run such localisation experiment with real human listeners. The predicted localisation performance can then be used to quantify the difference between the features of the target HRTF and the individual (template) HRTF.

2.5 Spatial Hearing and Room Acoustics

In the previous sections of this chapter, human sound localisation has been discussed primarily in the context of a free-field anechoic environment. However, in everyday life, most sounds interact with the surrounding space. This interaction affects the acoustic properties of the sound reaching our ears and, consequently, influences our spatial hearing abilities in realistic environments. The effect of virtual room acoustics rendering on spatial hearing is investigated further in § 8, while this section provides a literature background to the topic.

In enclosed spaces, the sound reaching our ear canals comprises a direct sound, a sparse set of early reflections from nearby surfaces, and a dense late reverberation tail formed by a multitude of higher-order reflections. This is illustrated in a schematic diagram of impulsive sound decay over time in Figure 2.3, known as the room impulse response (IR). The acoustics of a room are typically characterised by the reverberation time (RT)—the time it takes for the sound energy to decay by 60 dB. Since measuring the length of the full 60 dB drop (notated as

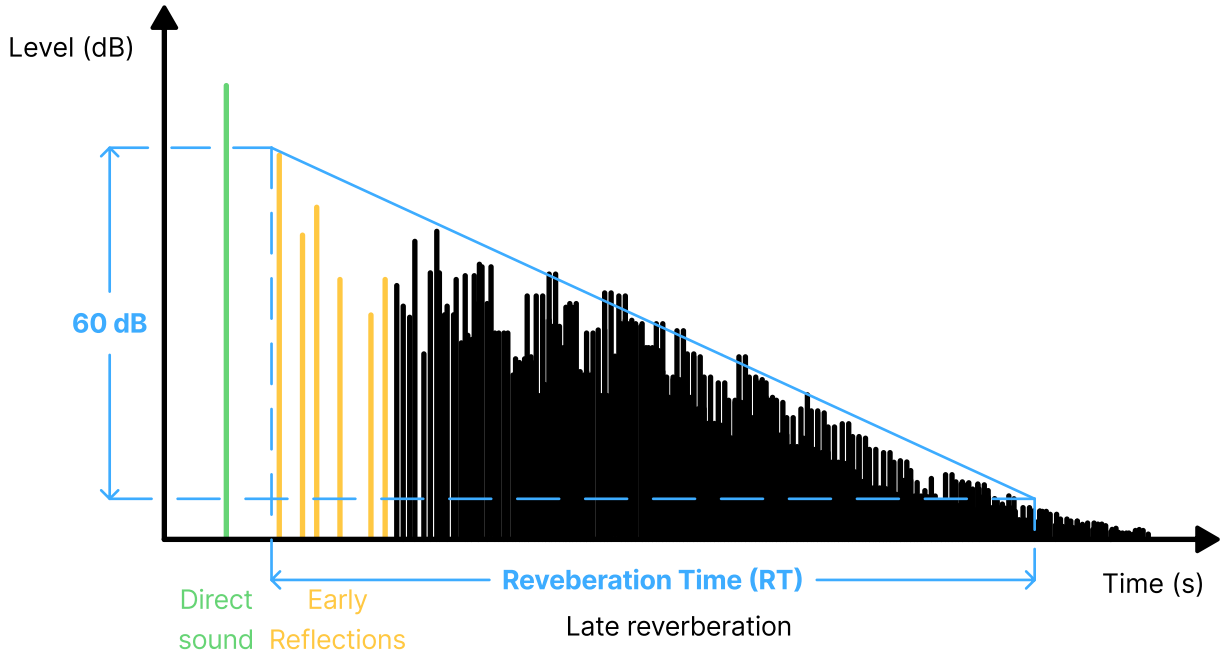


Figure 2.3: A schematic representation of sound decay over time in a room after an impulsive sound source. The direct sound is followed by early reflections and late reverberation.

T_{60}) can be impractical in realistic noisy conditions, shorter decay times are often used as proxy measurements and time extrapolated (e.g., T_{30} refers to the measurement of 30 dB decay time multiplied by two). The relative strength of the direct sound compared to the reverberation is described by the direct-to-reverberant ratio (DRR)—the ratio of the direct sound energy to the reverberant energy. However, no standardised definition of the direct sound time window exists, since the arrival of the first reflection varies depending on the room geometry and the position of the source and the listener.

RT and DRR depend on the room geometry and volume, surface materials, and the position of the sound source and the listener within the room (Kuttruff 2016). While these parameters provide a general overview of room acoustics, they are often measured from omnidirectional IRs, which do not capture the spatial characteristics of the sound field, crucial for spatial hearing. Therefore, realistic measurements and reproduction of spatial room acoustics requires the use of spatial room impulse responses (SRIRs) measured with multi-channel microphone arrays or binaural microphones (Alary et al. 2021; Engel et al. 2019; Tervo et al. 2013).

Although early reflections are highly directional, they typically do not interfere with the perceived location of the sound source. This is because the auditory system localises sound based on the earliest arriving wavefront—a phenomenon known as the ‘precedence effect’ (Blauert 1997). Despite this effect, early reflections can alter the spectral characteristics of the direct sound due to comb-filtering caused by interference between the direct and reflected sound waves. While these spectral changes are audible, listeners generally require training

to use early reflections effectively for understanding their spatial orientation within a room (Meyer-Kahlen et al. 2022; Neidhardt et al. 2016; Shinn-Cunningham and Ram 2003).

The transition from sparse, directional early reflections to a dense, spatially diffuse reverberant field is referred to as the mixing time. In typical room geometries, the late reverberation is perceived as diffuse and indistinguishable across different positions in the room (Lindau et al. 2012). However, in more complex spaces, such as coupled volumes, the reverberant tail may exhibit spatial anisotropy, meaning it has an uneven spatial distribution of energy, which can be perceived by listeners (Alary et al. 2021).

Room acoustics can influence auditory perception in two ways. On the one hand, they can enrich the auditory scene: reverberation often enhances the sense of externalisation, which means that sound sources are perceived outside the listener's head (Best et al. 2020). Early reflections in well-designed performance venues can also contribute to a sense of envelopment and enhance the dynamic qualities of sound (Lokki and Pätynen 2020). On the other hand, excessive reverberation can impair speech intelligibility (see § 2.6), negatively impact concentration and increase listener mental effort (Mercugliano et al. 2025). Therefore, thoughtful acoustic design in physical as well as virtual spaces is essential for delivering optimal auditory experience.

2.6 Spatial Hearing and Speech

Beyond direct sound localisation, spatial hearing plays a crucial role in speech perception, particularly in complex auditory scenes often referred to as the *cocktail party* scenarios (Cherry 1953). The ability of the human auditory system to analyse the auditory scene—grouping the sound components into auditory streams based on their spatial and temporal characteristics and segregating different auditory objects—is essential for understanding the attended speech in the presence of background noise and surrounding babble (Bregman 1990).

The overlap of sound energy in the same frequency bands between the signal of interest (for example, target speech) and noise causes *energetic masking* within the peripheral auditory system (Culling and Stone 2017). When the target speech and the noise are spatially separated, the auditory system exploits binaural cues to help unmask the useful signal. ILD allows the auditory system to attend to the ear with higher signal-to-noise ratio (SNR) (known as the 'better ear listening') while differences in ITD and interaural cross-correlation (IACC) provide additional binaural unmasking via binaural decorrelation mechanism (Bronkhorst and Plomp 1988; Bronkhorst 2015; Culling et al. 2004). The combined benefit from these cues is known as

the spatial release from masking (SRM). It can be quantified as a difference in speech reception threshold (SRT) between the colocated and the separated scenarios. Here, SRT is often defined as the SNR at which the subject understands 50 % of the speech.

In addition to *energetic masking*, the target speech is also masked by competing auditory streams in the central auditory system, such as semantically meaningful background conversations. This additional masking process is known as *informational masking* (Kidd and Colburn 2017). Various cues, such as differences between talker timbre, sound level, and temporal modulations (known as 'dip listening'), are used by the auditory system to segregate the target speech from the masking speech (Brungart 2001b). Still, in such scenarios, perceived spatial separation between the speakers plays a crucial role in unmasking the target speech (Freyman et al. 1999). Studies that tested speech intelligibility with spatially separated target and masker sources on the median plane (i.e., varying elevation) have found substantial SRM benefits, that cannot be explained by interaural cue asymmetries alone, suggesting the importance of informational unmasking mechanisms for vertically separated sources (González-Toledo et al. 2024; Martin et al. 2012).

Besides background noise and interfering conversations, real-world conversations often happen in reverberant environments. Reverberation blurs the target speech and smooths envelope modulations of the masking speech, impairing dip listening. It also reduces ILD and IACC, diminishing binaural listening benefits (Lavandier and Best 2020). The auditory system appears to be more sensitive to reverberation on the masking noise because of reduced binaural coherence than reverberation on the target speech that leads to amplitude modulation smoothing—a monaural effect (Lavandier and Culling 2008). Nonetheless, the auditory system utilises binaural hearing for de-reverberation, providing a small benefit to speech intelligibility (Lavandier and Best 2020; Lavandier and Culling 2008).

Our current understanding of binaural speech intelligibility has been used to develop multiple models that predict the binaural benefit in speech understanding (Lavandier and Best 2020). Some of them predict the SRM by estimating better-ear listening and binaural unmasking effects from binaural target and the masker signals or equivalent binaural room impulse responses (BRIRs) (Lavandier et al. 2022). These models can account for anechoic (Jelfs et al. 2011) or reverberant (Leclère et al. 2015) conditions, non-stationary maskers (Vicente and Lavandier 2020) and hearing impairments (Vicente et al. 2020). The models are useful in predicting the amount of *energetic masking* in various real and virtual spatial audio scenarios and can help design better architectural spaces (Culling et al. 2013). However, our ability to estimate *informational* masking components is more limited due to its cognitive basis and contextual

nature (Lavandier and Best 2020).

Overall, spatial hearing mechanism is an important communication tool. By using binaural cues and spatial source separation, the auditory system can mitigate both energetic and informational masking and enhance our ability to understand speech in complex everyday situations.

2.7 Chapter Summary and Conclusions

This chapter provided an overview of human spatial hearing, focusing on sound localisation mechanisms and its assessment methods. It discussed the acoustic features of the incoming sound used by the human auditory system for localising sound sources, including binaural and monaural cues, and highlighted the importance of dynamic listening with head movements. The chapter also reviewed common experimental considerations for assessing sound localisation performance and introduced key localisation error metrics used throughout this thesis to quantify human localisation performance.

Furthermore, the chapter overviewed the current state of the art in computational auditory modelling for sound localisation, including a detailed description of the *barumerli2023* model, which forms the basis for the perceptually-motivated analysis of HRTFs, presented in the subsequent chapters of this thesis. Finally, it explored the relationship between spatial hearing, room acoustics, and speech perception, providing the background literature review for the investigation on the intelligibility of binaurally rendered reverberant speech in Chapter 8. This background in basic auditory science lays the groundwork for the next chapter that reviews HRTFs and their individual nature from the technical perspective of spatial audio processing and rendering.

Chapter 3

Head-Related Transfer Function (HRTF)

3.1 Introduction

Chapter 2 introduced the concepts of the HRIR and HRTF as representations of spatial hearing cues employed by the human auditory system to construct spatial auditory scenes, localise sound sources, and support speech comprehension. This chapter approaches HRTF from the perspective of acoustics, audio technology, and signal processing.

The scattering of sound around the human head can be understood as a linear time-invariant system, allowing it to be measured and represented as a set of audio filters. These filters can then be convolved with audio signals to simulate virtual auditory scenes. Although this process is fundamentally numerical, its effectiveness is ultimately judged through perceptual experience. This duality between technical implementation and perceptual evaluation poses challenges in assessing the limitations of HRTF acquisition methods and the quality of binaural rendering algorithms.

Therefore, this chapter provides an overview of the contrasting perspectives on HRTFs as both measurable audio filters and perceptually meaningful constructs, in the context of binaural audio rendering.

3.2 Acoustic Definition and Practical Limitations

Section adapted from publication **VII** (Daugintis et al. 2025a).

Acoustically, the HRTF describes how a sound wave originating from a source in the free field is transformed at the entrance to the ear canal, accounting for the acoustic effects of the listener's head, torso, and pinnae (Xie 2013). Mathematically, it can be defined as follows (assuming

far-field condition):

$$H_{L,R}(\theta, \phi, f) = \frac{P_{L,R}(\theta, \phi, f)}{P_{ref}(f)} \tag{3.1}$$

where $H_{L,R}$ denotes the HRTF for the left or the right ear, $P_{L,R}(\theta, \phi, f)$ represents the complex sound pressure at the left or the right ear canal as a function of frequency f and incident direction (θ, ϕ) of the sound wave, while $P_{ref}(f)$ denotes the reference sound pressure at the center of the head without the listener present[1].

Although this theoretical definition provides a precise formulation of HRTFs, practical implementations rely on acoustic measurements conducted in specialised facilities. These measurements introduce variations due to factors such as internal equipment noise, deviations from an ideal free-field environment, uncertainty of microphone placement, and other measurement constraints (Møller et al. 1995). To enhance the quality and consistency of measured HRTFs, various post-processing methods have been identified. For example, some HRTF datasets are diffuse-field compensated by removing the average transfer function (i.e. the common transfer function (CTF)) from the HRTF (Møller 1992). The compensated transfer functions are also known as DTFs (Middlebrooks 1999b). Some may also use methods to compensate for the poor low-frequency response of the loudspeakers used in the measurements (Armstrong et al. 2018; Brinkmann et al. 2019b). However, procedural differences across measurement facilities (Li and Peissig 2020) result in dataset-specific artifacts (Andreopoulou et al. 2015; Pauwels and Picinali 2023) with unknown perceptual impact on the binaural rendering quality.

3.3 Perceptual Considerations

Section adapted from publication **VII** (Daugintis et al. 2025a).

Inherent practical limitations in acquiring an HRTF that would perfectly correspond to its theoretical definition (Equation 3.1), necessitate the discussion on the perceptual validity of the measured or modelled HRTFs. To address this, Geronazzo (2025) proposed a framework of *strong* and *weak* HRTF definitions. It mirrors a discussion on artificial intelligence (AI), where *strong/general AI* aims to replicate human cognition entirely and *weak/narrow AI* focuses on specific task-oriented capabilities (Goertzel 2014). Here, an aspect of this framework is discussed, focussing on the dichotomy between physically authentic and perceptually plausible HRTF definitions.

[1] Here, a singular HRTF refers to a set of direction-dependent filters, corresponding to one subject and an HRTF dataset is a collection of HRTFs measured in the same facility.

Physically authentic HRTF aspires to encapsulate the full physical reality of acoustic transformations through rigorous and highly detailed measurements, emphasising fidelity to physical properties. This definition characterises the HRTF as an exhaustive acoustic measurement of the individual's body, accounting for every anatomical detail that affects sound propagation, such as the shape and position of the ears, head, and torso. It emphasises the *in vivo* measurement of HRTFs, aiming at a representation as close to physical reality as possible. However, it inherently assumes a static model, disregarding dynamic variables such as transient movements, skin elasticity, or even hair growth. Physically authentic HRTF aims for the ideal but practically unattainable goal of achieving zero dissimilarity with negligible validation uncertainty.

Although, in general, measurements are considered a better indicator of reality (Oberkampf and Trucano 2002), precise HRTF measurements are still impractical and physically unvalidated at any standardised level, despite their proven perceptual validity (Langendijk and Bronkhorst 1999; Morimoto and Ando 1980; Wightman and Kistler 1989b) and their potentially short acquisition times (Dietrich et al. 2013; Enzner 2009; Majdak 2007). Moreover, significant discrepancies revealed by previously discussed cross-evaluation studies (Andreopoulou et al. 2015; Brinkmann et al. 2019b) reflect a common challenge in establishing a universally accepted *ground truth* (Oberkampf and Roy 2010; Roache 1998). These challenges illustrate the gap between theoretical aspirations and practical limitations, which requires complementary strategies such as incorporating perceptual validation metrics alongside physical measurements.

In contrast, perceptually plausible HRTFs can have broader applicability due to their adaptability and lower computational demands. By focusing on perceptual metrics, such as JNDs, such HRTFs can be effectively validated within specific auditory dimensions. For example, left-right localisation tasks rely on minimal ILDs and ITDs, which can be approximated without exhaustive physical accuracy. Studies have shown that JNDs for ILD and ITD are higher in realistic reverberant environments compared to anechoic ones (Klockgether and Van De Par 2016) and that JNDs for ITDs increase for more lateralised positions (Mossop and Culling 1998), suggesting context-driven approximation strategies. Similarly, research on spectral peak and notch detection thresholds provides an indication of the spectral resolution requirements for pinna-related spectral cues (Moore et al. 1989) and highlights the importance of specific peaks and notches for sagittal plane localisation (Spagnol et al. 2013). In virtual auditory displays, approximate HRTF data might suffice to create convincing spatial auditory experiences without the exhaustive detail required to capture physically authentic HRTF (Berger et al. 2018). The perceptual definition acknowledges the inherent trade-offs in scalability, reproducibility, and context-specific validation of HRTF measurements and simulations.

The dichotomy between the physical and the perceptual HRTF descriptions underscores the interplay between theoretical rigour and practical utility, the former an aspirational benchmark and the latter a viable framework for advancing HRTF research and applications within the constraints of current technology.

3.4 On the Individual Nature of HRTFs

Using individually measured HRTFs for binaural rendering is generally considered to provide the highest localisation accuracy (Middlebrooks 1999b; Møller et al. 1996; Romigh and Simpson 2014; Wenzel et al. 1993) and the most realistic virtual audio experience (Jenny and Reuter 2020). However, large-scale acquisition of individual measurements is impractical due to time and resource constraints. Consequently, most binaural audio applications rely on non-individual HRTFs, sometimes referred to as *generic* HRTFs when no specific adaptation to the listener is applied.

Non-individual HRTFs are commonly based on measurements of mannequin heads designed to approximate the average human anatomy. A well-known example is Knowles Electronics Manikin for Acoustic Research (KEMAR), developed with the aim of representing typical adult upper body dimensions (Burkhard and Sachs 1975). Its HRTF has been widely measured and included in multiple datasets (Armstrong et al. 2018; Braren and Fels 2020a; Burkhard and Sachs 1975; Engel et al. 2023; Gardner and Martin 1995). Another commonly measured mannequin head is the Neumann KU 100 (Andreopoulou et al. 2015; Armstrong et al. 2018; Brinkmann et al. 2019b). However, because these mannequins are based on assumed average dimensions, their use can lead to variability in user experience, particularly when the individual anatomy deviates significantly from the assumed average.

3.4.1 HRTF Personalisation

Although the use of non-individual HRTFs may be sufficient to achieve a sufficient perceptual rendering quality in dynamic audiovisual VR scenarios (Roßkopf et al. 2024; Rummukainen et al. 2021), the use of matching spatial cues might benefit some listeners in more critical listening tasks (Poirier-Quinot and Katz 2020). Therefore, various indirect HRTF personalisation techniques have been investigated to improve rendering quality without direct acoustic measurements.

Among the most accurate HRTF personalisation approaches is the synthesis of HRTFs from

high-resolution 3D head scans (Pollack et al. 2022) using numerical wave equation solvers, such as the boundary element method (BEM) (Brinkmann et al. 2023), finite element method (FEM) (Farahikia and Su 2017) or finite-difference time-domain (FDTD) methods (Prepeliță et al. 2019, 2020). Synthesised HRTFs can achieve quality comparable to the measured ones, with only minor deviations in spectral and directional cues (Brinkmann et al. 2019b). However, this method requires highly detailed scans to capture fine pinna structures (Dinakaran et al. 2018), and obtaining such scans typically demands specialised equipment, limiting its scalability.

More user-friendly personalisation methods generally involve either selecting the best-fitting HRTF from a database or adapting a non-individual HRTF to better match the listener. These methods can be based on perceptual feedback from the listener or numerical metrics (Guezenoc and Seguier 2018). Various techniques have been proposed, as reviewed by Fantini et al. (2025), Guezenoc and Seguier (2018) and Picinali and Katz (2023). However, no universally accepted and standardised approach exists, while multiple challenges remain in developing and validating perceptually robust HRTF personalisation techniques.

Perceptual personalisation approaches typically rely on listening tests to evaluate specific aspects of binaural audio rendering using non-individual HRTFs. These tests may, for example, ask listeners to rank or rate the quality of the sounds rendered along predefined spatial trajectories using different HRTFs (Iwaya 2006; Parseihian and Katz 2012; Seeber and Fastl 2003; Yamamoto and Igarashi 2017). Alternatively, listeners may assess modified versions of a non-individual HRTF, for example, its variants scaled in frequency domain (Middlebrooks et al. 2000). Although comprehensive listening tests can be time-consuming, several optimisations have been proposed, including reducing the HRTF dataset (Katz and Parseihian 2012) or adopting tournament-style designs (Iwaya 2006; Voong and Oehler 2019).

While most perceptual HRTF personalisation methods aim at minimising the localisation error, other subjective aspects of binaural audio rendering can be used (Simon et al. 2016). For example, perceived colouration and externalisation have been considered as metrics for fitting non-individual HRTFs (McMullen et al. 2012; Pelzer et al. 2020; Seeber and Fastl 2003; Voong and Oehler 2019). However, the relationship between qualitative subjective ratings and quantitative localisation performance remains unclear. Although some studies report correlations between localisation accuracy and overall perceived quality (Zagala et al. 2020), results often depend on listeners' prior critical listening experience (Kim et al. 2020). While perceptual HRTF personalisation approaches give listeners a sense of control, some perceptual approaches exhibit low repeatability rates, particularly among naïve listeners (Andreopoulou and Katz 2016), raising concerns about their suitability for large-scale deployment.

Numerical approaches instead match or adapt non-individual HRTFs based on an indirect measurement of the listener. Most commonly, anthropometric measurements of the head and pinna are used. Given the dominant role of ITDs in lateral localisation of broadband sounds, a partial personalisation strategy is adjusting the ITDs of a non-individual HRTF to match individual head dimensions. Individual ITDs can be estimated from head circumference measurements using simple geometric models, such as the Woodworth (1938) spherical head model, or by data-driven approaches such as linear regression across existing HRTFs (Katz and Parseihian 2012). Commercial solutions for personalising ITDs based on acoustic measurements by microphones placed in earphones have also been proposed (Satongar et al. 2021).

While ITD personalisation may improve lateral localisation accuracy, dominated by interaural cues, humans are more sensitive to deviations in pinna cue deviations, which reduce localisation performance along the cone of confusion (Romigh and Simpson 2014). Therefore, comprehensive HRTF personalisation methods must also account for anthropometric pinna features. The proposed techniques for selecting or adapting non-individual HRTFs based on ear shapes differ in how they capture the pinna morphology and map it to the acoustic characteristics. Some approaches involve manually measuring predefined distances within the pinna and selecting the HRTF that corresponds to the closest matching pinna shape (Zotkin et al. 2003). Others estimate the frequency of the main spectral notch from the distance between the ear canal and the pinna contour using a ray-tracing model and compare it to the main notches of non-individual HRTFs (Geronazzo et al. 2019). Automated ML techniques have also been explored to detect pinna landmarks from images (Zhi et al. 2022). In addition, data-driven approaches have been used to learn more complex relationship between pinna features and HRTF characteristics, such as main HRTF notch frequencies (Arbel et al. 2024), predicted binaural colouration (Marggraf-Turley et al. 2024) and perceptual HRTF characteristics (Pelzer et al. 2020). Some of these techniques have been patented and used commercially in wearable audio devices (Brimijoin et al. 2021; Johansson et al. 2020; Joyner et al. 2023).

In general, recent advances in ML have enabled data-driven personalisation methods that learn complex relationships between morphological, visual, acoustic, or perceptual input data and HRTFs (Fantini et al. 2025). Despite these developments, several key challenges remain, including the lack of large unified datasets for model training (Daugintis et al. 2025a) and the difficulty of defining objective metrics that reliably predict the perceived quality of binaural rendering for use as loss functions (Ananthabhotla et al. 2021; Doma et al. 2023a). Some of these challenges are discussed further in Chapter 5.

3.4.2 Human Adaptation to Non-Individual HRTFs

Most HRTF personalisation approaches focus on adapting the HRTF to the listener (i.e. system-to-user HRTF adaptation). However, the ability of the auditory system to adapt to altered hearing cues, discussed in § 2.2.4, allows for the investigation into user-to-system HRTF adaptation approaches (Picinali and Katz 2023). Studies have shown that listeners can adapt to non-individual or altered HRTFs through appropriate training (Berger et al. 2018; Majdak et al. 2013; Mendonça et al. 2013, 2012; Parseihian and Katz 2012; Poirier-Quinot and Katz 2021; Steadman et al. 2019; Stitt et al. 2019; Zahorik et al. 2006). This evidence has motivated proposals for the adoption of a standardised non-individual HRTF across binaural applications to promote long-term adaptation and improve perceptual quality (Lladó et al. 2024b).

Despite these generally positive findings, the effectiveness of adaptation appears to depend on the initial match between the HRTF and the listener. For example, Parseihian and Katz (2012) reported faster improvement when training with well-matched HRTFs, whereas Stitt et al. (2019) observed limited and inconsistent adaptation with poorly matched HRTFs. Consequently, while a standard non-individual HRTF may benefit some users, it may be suboptimal for others, leading to variability in user experience. These findings highlight the need for a more holistic personalisation strategy that combines system-to-user and user-to-system HRTF adaptation approaches.

3.5 HRTF Evaluation Approaches

The uncertainties associated with the acoustic and numerical HRTF acquisition methods, as well as HRTF personalisation techniques, require robust approaches to evaluate the validity of the resulting HRTFs. Such evaluation is inherently complex due to the multimodal nature of spatial hearing. Depending on the application, an appropriate perceptual HRTF plausibility criterion must be established, as discussed in § 3.3. This section reviews commonly used techniques for assessing HRTF similarity and perceptual quality.

3.5.1 Numerical analysis

A straightforward approach to comparing two HRTF or assessing the similarity of a non-individual HRTF to the individual one is through numerical distance metrics. These provide an efficient, quantitative measure of similarity without requiring extensive listening tests but establishing their relationship to perceived differences can sometimes be difficult.

Spectral difference metrics are widely used to evaluate HRTF similarity and as loss functions in ML-based personalisation algorithms (Fantini et al. 2025). They can be calculated by averaging squared magnitude differences in a linear scale (known as a mean squared error (MSE)) or logarithm of magnitude ratios in a decibel scale (known as a logarithmic spectral distance/distortion (LSD)) across all frequencies and directions. Alternative approaches use differences in spectral variance rather than mean differences to reduce sensitivity to overall gain variations (Middlebrooks 1999a). While computationally simple, these methods do not account for perceptual sensitivity to spectral cues. To address this, some studies use perceptually motivated frequency bands (e.g. ERB or mel-frequency cepstral coefficients (MFCC)) or weighting schemes when calculating spectral differences (Doma et al. 2023a). However, as discussed in § 2.2.2, determining appropriate weights remains challenging.

More specialised spectral metrics have been proposed for specific applications. For example, Geronazzo et al. (2018) proposed a distance metric between the main HRTF notches to assess the fit of non-individual HRTFs and used it to select a better fitting non-individual HRTF in terms of sound localisation (Geronazzo et al. 2019). However, the metric assumes the importance of the most prominent notch but does not consider other spectral features. On the other hand, limiting the scope to perceived spectral colouration, McKenzie et al. (2022) developed a perceptually weighted spectral metric for predicting timbral differences, later refined with binaural weighting (McKenzie and Brinkmann 2025). While the metric is able to predict perceived differences in colouration, its relationship to other spatial perception aspects remains unclear.

Comparative analyses on spectrally altered HRTFs reveal that different metrics vary in their sensitivity to numerical manipulations (Doma et al. 2023a), motivating composite models that integrate multiple metrics (Doma et al. 2023b). Yao et al. (2024) further demonstrated that identical LSD values can yield different perceptual ratings depending on degradation method and spatial position. They observed that degradations producing LSD of up to 2 dB were generally perceived as similar to the original, regardless of the method used, though further research is needed to establish the JNDs. To improve prediction accuracy, Yao et al. (2024) proposed a perceptually enhanced metric combining mean and variance statistics, peripheral auditory system-inspired spectral analysis, LSD thresholding, and binaural weighting. While this metric outperforms others in predicting listening test results, its generalisability to other perceptual attributes, such as localisation, remains uncertain.

In addition to spectral metrics, binaural measures can be used to assess differences in ILDs and ITDs. ILD is typically computed as the broadband level difference between left and right HRTFs (Barumerli et al. 2023; Middlebrooks et al. 1989). On the other hand, ITDs estimation

methods vary. They can be based on IACC (Middlebrooks 1999a), group delays (Minnaar et al. 2000), or onset thresholds (Algazi et al. 2001a). Katz and Noisternig (2014) reported that using different methods can result in ITD differences exceeding JNDs. Listening tests by Andreopoulou and Katz (2017) identified that the threshold method using relative value of −30 dB from the low-passed (at 3 kHz) HRIR peak was the most perceptually relevant way to extract ITDs. Otherwise, a commonly used maximum of the IACC of the low-passed energy envelope (MaxIACCe) method also provided similar ITD estimates.

Generally, auditory models, discussed in § 2.4, can serve as numerical HRTF evaluation tools (Barumerli et al. 2018b). The models often integrate the discussed numerical metrics in a perceptually motivated manner. Predicted localisation performance can then be analysed using localisation error metrics (see § 2.3.1), providing an intuitive means to assess the impact of different HRTFs on human listeners. However, while these models are designed to replicate empirical perceptual data, their application beyond the scope for which they were validated should be considered with caution.

3.5.2 Perceptual Evaluation

Various perceptual evaluation methods have been employed to assess HRTF quality and personalisation techniques. These methods can be broadly categorised into quantitative tests, based on a measurable listener's performance within a given task, and qualitative assessments, where the listener is asked to subjectively evaluate the HRTF-based audio rendering.

Sound Localisation

A common strategy for quantitatively assessing perceptual HRTF quality is by performning a sound localisation experiment. Many performed tests were static (Begault 2001; Daugintis et al. 2023a; Djelani et al. 2000; Geronazzo et al. 2019; Majdak et al. 2010; Meyer and Picinali 2025b; Middlebrooks 1999b; Romigh and Simpson 2014; Wenzel et al. 1993; Wightman and Kistler 1989a; Zagala et al. 2020), but they varied in the pointing method, sound stimulus, and data analysis methodology, based on considerations, similar to the ones discussed in § 2.3. Previous studies generally found that with careful calibration, sound localisation performance using individual HRTF-based spatialisation can be generally comparable to the free-field listening condition (Djelani et al. 2000; Romigh et al. 2015; Wightman and Kistler 1989a). The use of non-individual HRTFs, however, typically resulted in degraded localisation performance, especially in reducing elevation perception and increasing the front-back confusion rates (Middlebrooks

1999b; Romigh and Simpson 2014; Wenzel et al. 1993). Using a succesfull HRTF personalisation strategy resulted in reduction of these errors (Geronazzo et al. 2019; Middlebrooks 1999b). However, some studies, especially those using speech stimuli instead of broadband noise bursts, reported no significant difference between individual and non-individual HRTFs (Begault 2001; Meyer and Picinali 2025b), possibly due to the limited spatial cues in such stimuli. These findings suggest that while HRTF individualisation is important in critical listening scenarios, it is often unclear how these findings generalise to dynamically rendered scenes with more ecologically valid sound stimuli.

In an attempt to have a more ecologically valid localisation-based scenario with quantitative outcomes, Poirier-Quinot and Katz (2020) employed a shooter-style game paradigm to test the effect of HRTF personalisation on game performance. In this study, participants engaged in a VR task where they were required to shoot approaching audiovisual enemies. As the game progressed and difficulty increased, several performance metrics were tracked, including mean game score, spawn-spot reaction time, and angular distance travelled. Participants were divided into groups based on their use of either the best- or the worst-matched HRTFs, determined through qualitative preference-based HRTF selection. The results showed no substantial improvement in performance metrics attributable to the specific HRTF used, except for some direction-dependent advantages of well-fitting HRTFs in elevated regions, which were limited to the early stages of gameplay. The authors discussed a complex interaction between learning the game mechanics and adapting to sound localisation with non-individualised HRTFs. This study highlights the challenges of designing ecologically valid sound localisation experiments for HRTF evaluation, particularly due to the influence of multimodal interactions in complex listening environments.

Authenticity and Plausibility

In the context of auditory VR/AR, several paradigms based on signal detection theory have been developed to assess the perceptual success of auditory illusions through quantitative frameworks. While these experiments also include realistic room acoustics rendering to match real sources, thus testing a combination of rendering aspects altogether, the accuracy of HRTF (or, more precisely, BRIR) personalisation also plays a critical role in the test outcomes. Among the proposed paradigms, the question of *authenticity*—which asks whether a virtual source is indistinguishable from a real one (Blauert 1997)—is considered the most sensitive. This is typically tested using an ABX setup, where listeners must identify which of two references (A or

B, randomly assigned to virtual and real sources) matches the target (X, which is either the virtual or real source) (Brinkmann et al. 2017). Due to the presence of a physical reference, this paradigm is highly sensitive to small differences in rendering quality. Even highly calibrated systems with individualised BRIRs can fail to achieve authenticity under this paradigm (Brinkmann et al. 2017), making it challenging to apply in practical scenarios (Meyer-Kahlen et al. 2024).

A more lenient approach is to assess the *plausibility* of the virtual source, defined as an 'agreement with the listener's expectation towards an equivalent real acoustic event' (Lindau and Weinzierl 2012). The evaluation is typically formulated as a yes/no task, where the listener decides whether the source is virtual or real, relying on their internal reference of an equivalent real source (Lindau and Weinzierl 2012). This paradigm has been used to evaluate the importance of near-field HRTF cues (Arend et al. 2021b) and the effect of simplified spherical-model HRTFs (Lübeck et al. 2024), showing a limited effect of HRTF manipulation on plausibility ratings.

In the context of audio AR, which may include concurrent virtual and real sources, a paradigm of transfer-plausibility has been proposed to assess whether a 'virtual sound source [...] is believed to be real in the presence of real sound sources' (Meyer-Kahlen et al. 2024). The task involves listening to three or more concurrent sound sources and identifying which one is virtual (e.g. a 3AFC paradigm). Test sensitivity has been found to depend on the complexity of the scene (Wirler et al. 2020), but this approach avoids the ceiling effect of the strict authenticity test and improves testing efficiency compared to the plausibility test (Meyer-Kahlen et al. 2024).

Speech Intelligibility

With the development of spatialised teleconferencing, creating a realistic virtual *cocktail party* scenario while ensuring adequate intelligibility of the speakers and keeping the natural conversation flow has become an important goal for binaural audio rendering technologies. Since spatial hearing cues help the auditory system unmask attended speech (see § 2.6), the choice of HRTF may have an affect on speech comprehension. Although several studies have explored the use of speech intelligibility as a quantitative assessment of HRTF-based rendering, the results remain inconclusive. This section reviews some of the key studies in this area, which provide background for the study, discussed in Chapter 8.

Kondo et al. (2010) evaluated speech intelligibility in noisy environments with the target speaker positioned in front and maskers placed at various azimuths and distances (i.e. different SNRs). The study compared real loudspeakers, individual HRTFs, and KEMAR HRTF. While differences were minor, intelligibility was highest with real loudspeakers, and slightly better with

individual HRTF than the KEMAR HRTF. The largest discrepancy occurred when the masker was directly behind the listener, likely due to front-back confusions. However, the study lacked statistical analysis and involved only five participants (some of whom varied across sessions), limiting the strength of its conclusions.

Orduña-Bustamante et al. (2018) investigated speech intelligibility in noisy and reverberant (dichotic and diotic) conditions using individual and non-individual HRTFs. They observed a benefit from individual HRTFs over non-individual ones. However, no correlation was found between the spectral distortion of non-individual HRTFs and intelligibility scores. Instead, a model based on modulation index theory was proposed to partially explain the findings.

Ahrens et al. (2021) used computational binaural speech intelligibility models to examine the effect of different HRTFs on intelligibility across varying target-masker separations. Their results showed differences in predicted SRM depending on the HRTF used, with greater variability at wider separations. However, the study could not assess the impact of HRTF individualisation due to its computational nature.

Cuevas-Rodriguez et al. (2021) investigated the effect of different HRTFs on understanding spatially separated speech, with the target speech presented in front and two speech-shaped noise maskers to the left and right. While differences in SRT across seven HRTFs from the LISTEN database and one spherical head model were generally minor (except for one HRTF and the spherical model), inter-subject variability allude to the individual nature of HRTF impact on SRT. However, the study did not investigate individual HRTFs.

Zenke and Rosen (2022) measured SRM using individual and non-individual HRTFs in adults and children. In adults, no significant difference was found between HRTF conditions when the target was in front and maskers at 90° left and right. The study used HRTFs representing the 'biggest' and the 'smallest' heads from the ARI database without compensating for ITD nor ILD. Among children, SRM increased with age but was similar across loudspeaker, individual, and KEMAR HRTF conditions. However, SRT was significantly higher for spherical head HRTF.

Most studies measured SRM on the horizontal plane, where binaural cues are the most relevant. However, researchers also explored the effect of perceived spatial separation on the median plane, particularly when using different HRTFs. Some of the release from masking appears to be driven by asymmetries between the ears, leading to interaural cues even on the median plane. Nonetheless, by using individual and manipulated symmetric HRTFs, Martin et al. (2012) demonstrated that monaural cues play an important role in differentiating competing speech sources. Building on these findings, González-Toledo et al. (2024) investigated the importance of individual HRTF for median-plane SRM, finding that individual HRTFs yielded

greater SRM than the KEMAR HRTF. Numerical differences between the HRTFs could not fully explain the results, suggesting that listener familiarity with specific HRTFs may contribute to the observed differences.

While some HRTF-individualisation-related effects on speech intelligibility have been reported, other HRTF-specific features often influence the results. Therefore, further research is needed to determine whether speech intelligibility can be reliably used as a metric for evaluating HRTF quality in virtual teleconferencing scenarios.

Neurophysiological Measures

Beyond behavioural assessments, neurophysiological methods have been employed to investigate the neural correlates of sound localisation performance and quantify HRTF-based spatial audio rendering quality. Callan et al. (2013) used functional magnetic resonance imaging (fMRI) to show that individualised binaural recordings elicited stronger activation of the posterior temporal gyri compared to internalised stereo recordings. Using electroencephalography (EEG) to measure event-related potentials (ERP) associated with detection of the change in virtual sound elevation, Wisniewski et al. (2016) found enhanced detection and a correlated increase in amplitudes of later ERP components (P3) when individualised HRTFs were used compared to non-individual ones. Marggraf-Turley et al. (2025b) compared sound localisation in the free-field and using non-individual (KEMAR) HRTFs using EEG and found reduced decoding accuracy with non-individual HRTFs. While these studies demonstrate the potential of using neurophysiological measurements to assess HRTF quality, the requirements for highly controlled experimental conditions, lengthy data collection, and complex data analysis due to high levels of noise currently limit their practical applicability beyond research settings.

Qualitative Approaches

Some studies have adopted a qualitative approach to assess binaural audio rendering and HRTF quality. For example, Katz and Parseihian (2012) proposed a method for rating HRTFs by rendering sounds along predefined trajectories and asking subjects to rate the rendering quality as ‘bad’, ‘ok’ or ‘excellent’. This HRTF evaluation was used to reduce the LISTEN HRTF dataset to seven ‘perceptually orthogonal’ HRTFs, later used for more rapid HRTF selection procedure (Parseihian and Katz 2012). However, while some correlation between the perceptual ratings and sound localisation performance could be found (Zagala et al. 2020), other studies have reported that this method can be inconsistent across repetitions, particularly when performed

by naïve listeners (Andreopoulou and Katz 2016; Kim et al. 2020).

More extensive qualitative tests can employ specific audio quality attributes, such as those defined in the Spatial Audio Quality Inventory (SAQI) (Lindau et al. 2014), or a more constrained set of attributes tailored for binaural rendering (Simon et al. 2016). Brinkmann and Weinzierl (2023, § 5.4.1.2) examined the effect of individual versus non-individual binaural rendering on perceived quality compared to a real loudspeaker. Results generally showed degradation across multiple attributes and increased response variability when non-individual rendering was used, as the goodness of fit for non-individual HRTFs depends on the listener. Furthermore, Jenny and Reuter (2020) found reduction in the sense of externalisation and realism when using non-individual HRTFs compared to the individual ones.

One of the key challenges with using qualitative measures for spatial audio assessment is the ambiguity and context-dependence of perceptual terms, such as immersion, presence, realism, and externalisation. These qualities can represent a varying combination of objective and subjective aspects, which can lead to multiple interpretations by researchers and study participants (Jenny and Reuter 2021). Similarly, Raake and Wierstorf (2020) discuss the challenges of assessing and predicting the *sound quality* and especially *quality of experience*, which represent a higher level of cognitive abstraction, in spatial audio contexts. In general, qualitative assessment can have high result variability and low statistical power, highlighting the importance of participant training to ensure consistent interpretation of the evaluation task.

3.6 Chapter Summary and Conclusions

Building on the theoretical foundations of the spatial hearing research presented in Chapter 2, this chapter provided an overview of HRTFs from the perspective of binaural spatial audio technologies. It highlighted the complexities involved in accurately capturing and utilising HRTFs for spatial audio rendering, emphasising the trade-offs between physical accuracy and perceptual relevance. The discussion on HRTF individualisation underscored the challenges and advancements in personalisation techniques, while the evaluation section reviewed various methods for assessing HRTF quality, both numerically and perceptually. This chapter provides a contextual background for the subsequent chapters that present original research contributions in measuring a new HRTF dataset (Chapter 4), developing perceptually-motivated numerical HRTF evaluation metrics (Chapters 5 and 6), and conducting perceptual validation experiments (Chapters 7 and 8).

Chapter 4

The SONICOM HRTF Dataset

This chapter presents the SONICOM HRTF dataset that was measured at Imperial College London as part of the SONICOM project[1] and to support the HRTF analysis and evaluation work, presented in this thesis. The original release of the dataset and its validation was presented in a journal publication **I** (Engel et al. 2023). This chapter is adapted from this publication, emphasising the work done by the thesis author in setting up the facility, measuring the subjects and performing measurement consistency evaluation. The extension of the dataset to over 300 subjects, together with a release of simulated HRTFs from 3D head scans is reported in a conference publication **IX** (Poole et al. 2025).

4.1 Introduction

Numerous HRTF datasets have been measured over the years, both within the academic community (Algazi et al. 2001c; Armstrong et al. 2018; Braren and Fels 2020b; Brinkmann et al. 2019b; Majdak et al. 2010; Sridhar et al. 2017; Warusfel 2003; Watanabe et al. 2014; Yu et al. 2018) and by commercial entities (Warnecke et al. 2024). However, most openly available datasets contain a limited number of subjects, typically fewer than 100. The available datasets also vary considerably in the measurement setup and the type of supplementary data they provide. Some include anthropometric measurements (Algazi et al. 2001c), high-resolution 3D scans of subjects' ears, heads, and torsos (Braren and Fels 2020b), or even simulated HRTFs obtained through wave-based computational methods (Brinkmann et al. 2019b). While such complementary data is valuable for developing ML-based HRTF personalisation methods, its sparsity across datasets, combined with quantifiable measurement inconsistencies (Andreopoulou et al. 2015;

[1] www.sonicom.eu

Pauwels and Picinali 2023), hinders effective cross-dataset data collation for training large-scale ML models (Daugintis et al. 2025a).

Beyond the need for larger amounts of consistent HRTF data for ML research, perceptual binaural rendering evaluation studies require in-house participants with individually measured HRTFs to serve as references in listening tests. This dual requirement motivated the measurement of a new, high-quality HRTF dataset containing a large number of subjects within the scope of the SONICOM project. The dataset thus supports both the SONICOM project's goal of advancing immersive audio in VR/AR environments through data-driven techniques and this research project's need for participants with their associated individual HRTFs for perceptual HRTF similarity evaluation experiments.

The dataset currently contains over 300 acoustic HRTF measurement. It also includes 360° photos and high-resolution 3D scans of subjects' heads as well as individual headphone equalisation filters. Recently, the dataset was extended to include simulated HRTFs for some subjects, obtained through the boundary element method using the Mesh2HRTF software (Poole et al. 2025). All data is made openly available to the research community to support further research in spatial audio and acoustics[2].

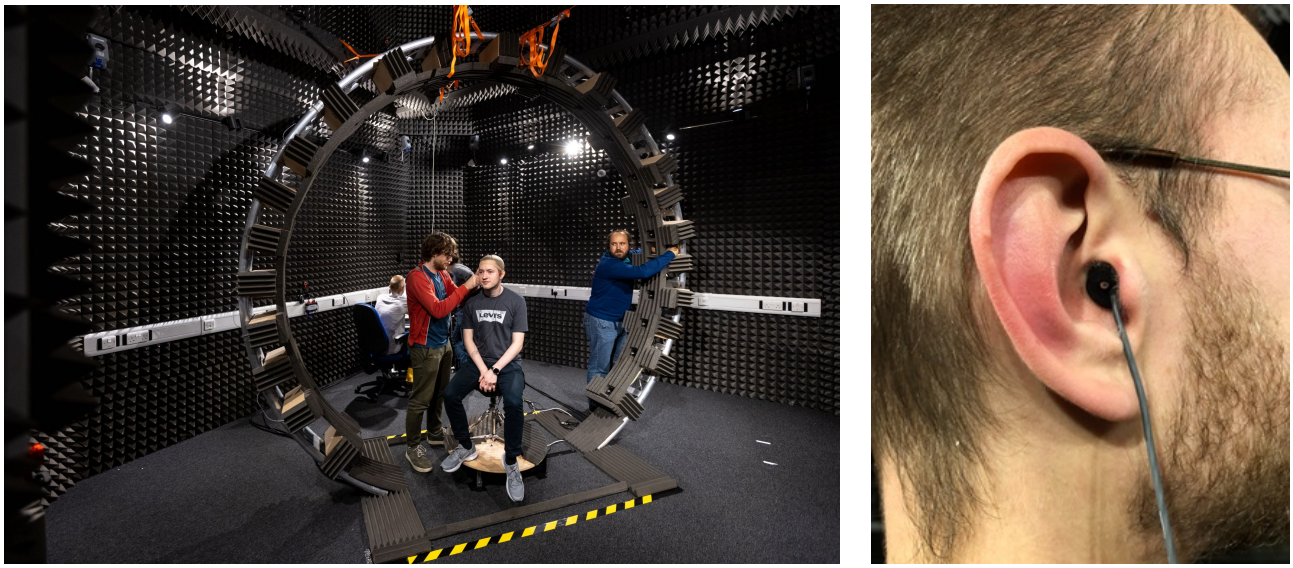
As part of the dataset release, an evaluation of the measurement consistency was undertaken. A set of metrics were chosen to understand the impact of measurement uncertainty on the measured HRTF quality in both numeric and perceptual way. This section briefly presents the dataset and, specifically, the work that was conducted to evaluate the consistency and repeatability of the HRTF measurements.

4.2 HRTF Measurements

HRTFs were measured in an acoustically treated room, shown in Figure 4.1a. Its reverberation time was 60 ms at 1 kHz and A-weighted background noise level was 23 dB . The sounds were played via 23 custom built full-range single-driver loudspeakers that were mounted on a vertical arch. The sounds were recorded using Knowles electret microphones that were placed at the entrance of the blocked subject's ear canals using memory foam earbuds (see Figure 4.1b).

The subjects were sat on a turntable in the centre of the arch, which was rotated in 5° intervals to cover all azimuth directions around the listener (the subjects were rotated by 175° in total because the loudspeakers were mounted symmetrically on both sides of the arch).

[2]Currently available via `https://transfer.ic.ac.uk:9090/#/2022_SONICOM-HRTF-DATASET/` and within the SONICOM ecosystem (`https://ecosystem.sonicom.eu/databases`).



(a) Measurement facility. **(b)** Microphone in the subjects' ear.

Figure 4.1: HRTF measurement setup. Reproduced from Engel et al. 2023.

Multiple exponential sweep method using 500 ms sweeps was used to measure the broadband impulse responses (Majdak 2007). A chin rest was used to stabilise the subject's head during the measurement. The subject's initial position was aligned using mounted plane alignment lasers. During the measurement, the subject's position was continuously tracked using the Optitrack infrared tracking system and a set of reflective markers on a headband placed on the subject's head. The measurements during which the subject's head moved more than the set tolerance (±2.5° for pitch and yaw, ±5° for roll, ±10 cm for displacement) were repeated. The measurement was conducted using a custom-built application based on ExpSuite (Acoustics Research Institute of the Austrian Academy of Sciences, Austria)[3].

The recorded impulse responses were firstly windowed to 5 ms to remove room reflections. Then, they were compensated for the measurement system response using an inverse of the reference impulse response measured in the middle of the arch without the subject present. Two versions of the HRTFs are produced: one with the inverse of the original recorded reference measurement and one with its minimum phase equivalent. These versions of the HRTFs as well as the intermediate (unprocessed) impulse responses are available from an online repository[4]. For the purposes of this thesis, the HRTFs, compensated using a minimum phase filter were used (available as 'FreeFieldCompMinPhase' from the repository).

Beyond the acoustic HRTF measurements, the repository contains individual headphone equalisation filters for Sennheiser HD 650 headphones. They were obtained from a regularised

[3] https://www.oeaw.ac.at/isf/expsuite
[4] https://transfer.ic.ac.uk:9090/#/2022_SONICOM-HRTF-DATASET/

inversion of an average impulse response over five measurements. These filters were used in the listening tests presented in this thesis. The repository also includes high quality 3D head scans using EinScan Pro 2X 2020 3D scanner and associated simulated HRTFs using the boundary element method via Mesh2HRTF software (Brinkmann et al. 2023). Finally, 360° photos of the subject's head at 5° intervals were taken with a smartphone camera. They are available from the research team upon request and can be used for the development of image-based HRTF personalisation algorithms. The use and evaluation of simulated HRTFs and ML-based personalisation approaches using the images is beyond the scope of this thesis.

4.3 Measurement Validation

To evaluate the repeatability of the HRTF measurements, one subject (P0004) was measured five times, restarting the setup procedure for each measurement (re-inserting the microphones and realigning the subject). The five measurements were compared against each other as well as against five measurements of other subjects (P0001, P0002, P0003, P0006, and P0007). Numerical metrics were used to assess the differences between the measurements.

However, since HRTFs are used in spatial audio rendering, it is important to establish the perceptual impact of measurement uncertainties. Although subjective listening tests are the most truthful way to evaluate the perceptual qualities, designing and running them is demanding for researchers and test subjects. They were therefore deemed to be outside the scope of the HRTF measurement study. Instead, computational auditory models were chosen as a rapid alternative to assess the perceptual differences between the measurements.

In summary, the measurement consistency was analysed using numerical spectral difference metrics, speech intelligibility predictions, and estimated differences in sound localisation performance. Numerical spectral analysis and speech intelligibility predictions, conducted by a colleague, are summarised below, while sound localisation predictions, conducted by the thesis author, are discussed in full detail.

4.3.1 Numerical Assessment

Firstly, the average HRTF magnitudes and their standard deviations (SDs) were computed across the front position of 5 versions of the P0004 HRTF and 5 different HRTFs. It revealed that, generally, the SD across the different HRTFs is higher (above 4 dB) than across the five versions of the same HRTF (up to 3 dB). The SDs are generally higher at higher frequencies for both

multiple measurements of the same subject and for different subjects, especially around 8 kHz (Engel et al. 2023, Figure 8, top panels). This frequency range might be particularly sensitive to the microphone positioning in the ear canal, as well as to small changes in the subject alignment. However, smaller deviations across the different measurements of the same subject than across five different subjects indicate that the anatomical differences between the subjects have a greater influence on the measurements than the measurement setup.

Similar trend is revealed in average magnitude differences between five versions of P0004 HRTFs and 5 different HRTFs across all the measured positions (Engel et al. 2023, Figure 8, bottom panel): above 3 kHz, the average difference between the P0004 HRTFs is up to 2.5 dB, while the difference across five different HRTFs is around 6 dB. Generally, the numerical deviations across multiple versions of the same HRTF were found to be similar to the ones reported by Brinkmann et al. (2019b), indicating that the measurement consistency is similar to previously measured publicly available datasets.

4.3.2 Speech Intelligibility Predictions

Perceptual validity of binaural cues was evaluated using *jelfs2011* speech intelligibility model (Jelfs et al. 2011). The model predicts the binaural benefit in decibels (SRM) of understanding spatially separated speech in noise, based on extracted ITD and ILD values in frequency bands, weighted by their relevance to understanding speech. The predicted value then corresponds to a relative increase in noise level that would result in the same speech intelligibility as in the condition when the noise and the speech are colocated.

As reported in Engel et al. 2023, for positions along the horizontal plane, the model predicted similar SRM values across the repeated versions of P0004 HRTF (around 7.0–7.3 dB) and for different HRTF (around 7.1–7.5 dB). While it is difficult to validate the similarity between the repeated HRTF measurements using the estimated binaural speech advantage, these values are in line with previous literature (Jelfs et al. 2011), suggesting the validity and consistency of the binaural cues in the measured HRTFs.

4.3.3 Sound Localisation Predictions

The *barumerli2023* spherical sound localisation model, summarised in § 2.4.1, can be used to estimate how well a human listener would perform in a sound localisation task when presented with binaural stimuli processed with a target HRTF that differs from their own true HRTF (template). The inputs of the model are two HRTFs (target and template) in Spatially Oriented Format

for Acoustics (SOFA), and the experimental conditions, such as the evaluated sound source directions and the number of repetitions. The model outputs are the predicted sound direction estimations that would be made by a listener. Generally, if the target and template HRTFs are similar, the estimations will be close to the actual sound source directions. On the other hand, if they are different, the estimations will contain larger localisation errors and may contain a large number of front-back and up-down confusions. For an accurate estimate of absolute localisation errors, a set of free model parameters, which control non-HRTF-related aspects of sound localisation performance (e.g. individual pointing accuracy), have to be calibrated using individual sound localisation test data. However, personal sound localisation data was unavailable, and only relative differences in localisation errors were of interest to this study. Therefore, median parameter values, calculated using data from previous sound localisation studies (Majdak et al. 2010, 2013), were used in this study (the calibration procedure is discussed in more detail in § 6.3).

For the evaluation, the five measurements of subject P0004 were used as template HRTFs, representing five hypothetical listeners. For each of them, sound localisation tasks were simulated for the same ten target HRTFs that were analysed in the previous section: the five measured ones from subject P0004 plus the ones from the five randomly selected subjects. This is analogous to asking each listener to evaluate their own HRTF, four alternative versions of individually measured HRTFs, and five non-individual HRTFs. In each task, the listener would have to localise one of 1706 directions, interpolated from the supplied HRTF, for each of the 10 target HRTFs, with 300 repetitions (to account for stochasticity), resulting in 5 118 000 localisation estimations per listener.

Figure 4.2 shows the results of the evaluation. The top row displays the predicted sound localisation errors when using the five P0004 HRTFs as targets, whereas the bottom row shows the errors when using the other subjects' HRTFs as targets. The errors are divided into root-mean-square (RMS) local polar error (first column), RMS lateral error (middle column) and quadrant errors (third column), as defined in Middlebrooks 1999b and summarised in § 2.3.1. While the absolute values of the estimated localisation errors are difficult to compare to real human performance without individual model parameter calibration, they are within the range of previously reported values for human listeners (Majdak et al. 2010; Middlebrooks 1999b; Poirier-Quinot et al. 2022). The results are plotted in a matrix form to show the predicted errors of each target-template pair. Therefore, in the top row, diagonal positions represent modelled errors using the same HRTF as both the target and the template.

Overall, estimated localisation errors when using different versions of the same subject's

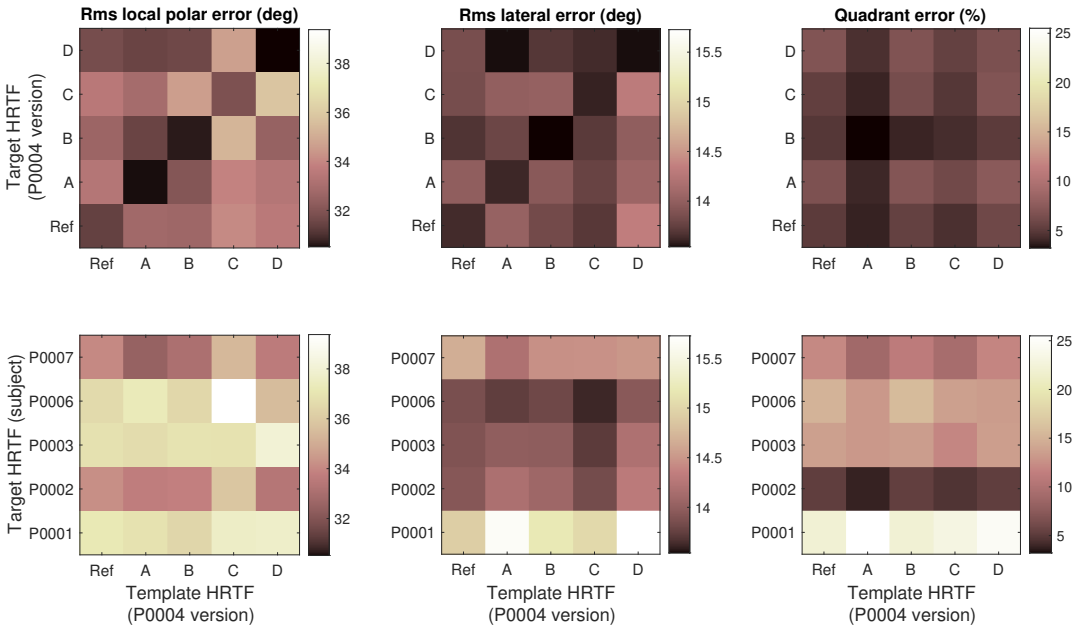


Figure 4.2: Modelled sound localisation errors for five versions of subject P0004 HRTFs when using different measurements of the individual HRTF (top row) and non-individual HRTFs from other subjects (bottom row). Reproduced from Engel et al. 2023.

HRTF are smaller than when using HRTFs from different subjects (i.e. non-individual). On average, estimated RMS local polar error across all the versions of the same-subject HRTFs (top-left subfigure) is around 32.8° (standard deviation (SD): 1.5°), while for different-subject HRTFs (bottom-left subfigure) it is around 35.8° (SD: 1.8°). This is a significant increase in error, according to a two-sample t-test ($t(48) = -6.66$, $p < 0.001$), where the significance level is set to 0.05. Similarly, RMS lateral error is around 13.8° (SD: 0.2°) for the same-subject HRTFs (top-middle subfigure) and around 14.3° (SD: 0.6°) for different-subject HRTFs (bottom-middle subfigure). While lateral error tends to be lower than the polar error and less affected when non-individual HRTFs are used in real listening experiments due to robustness of the interaural cues (Romigh and Simpson 2014), this estimated increase in lateral error is significant ($t(48) = -3.87$, $p < 0.001$). Finally, the average quadrant error is around 5.4 % (SD: 1.2 %) for the same-subject HRTFs (top-right subfigure) and around 13.3 % (SD: 6.2 %) for different-subject HRTFs (bottom-right subfigure), which is also a significant increase ($t(48) = -6.2$, $p < 0.001$). Increased localisation errors with non-individual HRTFs compared to different versions of individual HRTFs suggest the HRTF measurement inconsistencies have a significantly smaller perceptual impact than the anatomical differences between subjects.

These findings are further supported by the similarity of the estimated errors for each of the non-individual HRTFs across different versions of the template HRTFs (bottom row). To assess this similarity of the estimated errors across the template HRTF versions, an intra-class

correlation (ICC) analysis was conducted for each of the three error metrics. The degree of absolute agreement of estimated errors across the five versions of template HRTF was evaluated using a two-way mixed-effects model for averaged measurements, ICC(A,5) (McGraw and Wong 1996). For the estimated RMS local polar error across different non-individual HRTFs (bottom-left subfigure), $\mathrm{ICC(A,5)} = 0.94$ ($p < 0.001$), indicating excellent consistency of the estimated error across the templates (Koo and Li 2016). In comparison, for the same-subject HRTFs (top-left subfigure), $\mathrm{ICC(A,5)} = 0.1$ ($p = 0.38$), suggesting that there is no similarity between the predicted errors for each version of P0004 as a target. Simlarly, for the estimated RMS lateral error across different non-individual HRTFs (bottom-middle subfigure), $\mathrm{ICC(A,5)} = 0.97$ ($p < 0.001$) and for the estimated quadrant error (bottom-right subfigure), $\mathrm{ICC(A,5)} = 0.99$ ($p < 0.001$), indicating excellent consistency of predicted errors across different versions of the template HRTF. High absolute agreement of the errors across different template HRTF versions suggests that the measurement variability of the same subject had a minimal effect for the model outputs compared to the differences between the non-individual HRTFs used as targets.

The analysis based on the simulated sound localisation performance indicates that the measurement inconsistencies have a minor impact on the perceptual qualities of the measured HRTFs and they are significantly smaller than the HRTF differences between different subjects due to anatomical variability. These findings align with the numerical analysis results presented earlier and confirm the high quality and consistency of the measured HRTFs within the dataset.

4.4 Chapter Summary and Conclusions

This chapter presented the SONICOM HRTF dataset, which currently contains over 300 acoustic HRTF measurements alongside high-resolution 3D scans, photographs, individual headphone equalisation filters, and synthetic HRTFs derived from the scans. The measurement consistency analysis demonstrates that measurement uncertainties are minor, validating that the measurements capture the individual characteristics of each subject's HRTF and can serve as the ground truth for perceptual studies. Consequently, this dataset provides a valuable resource for the remainder of this thesis, supporting both the data-driven approaches for addressing HRTF analysis and personalisation challenges presented in Chapter 5, and the perceptual evaluation experiments detailed in Chapters 7 and 8.

Chapter 5

Challenges in HRTF Personalisation Research

This chapter presents my contribution to the development of the inaugural Listener Acoustic Personalisation (LAP) Challenge, and the analysis of its outcomes. The challenge, which addressed several issues in HRTF personalisation research, was organised in 2024, with the award ceremony held during the European Signal Processing Conference (EUSIPCO) 2024. It was supported by the IEEE Signal Processing Society and the SONICOM project[1]. § 5.2 presents Task 1 of the challenge, based on the journal publication **VII** (Daugintis et al. 2025a), where I served as the lead author. § 5.3 provides a shorter summary of Task 2, reflecting a more limited role in its development. It is based on the journal publication **VI** (Hogg et al. 2025).

5.1 Introduction

The field of binaural audio rendering has developed significantly from the earlier experimentation in specialised facilities (Møller 1992; Morimoto and Ando 1980) to widespread consumer applications (Rafaely et al. 2022). While the advancements in hardware and its processing power have made a real-time three or even six-degree-of-freedom[2] binaural rendering possible on wearable consumer devices, more work is necessary before the virtual audio scene can be perceived as indistinguishable from the physical one for every listener. Researchers are still working to understand and improve different aspects of the binaural audio reproduction pipeline (Rafaely et al. 2022), from improving the quality of spatial audio capture from wearable devices (Berebi

[1] https://www.sonicom.eu/lap-challenge/

[2] Three degrees of freedom here refers to the free rotation of the sound scene around the listener along the three rotational axes (pitch, yaw, and roll), while six degrees of freedom also include translational movements.

et al. 2025), to optimising real-time virtual acoustics rendering (Engel et al. 2021; Gari et al. 2024; Martin et al. 2025a), to finding the most appropriate assessment strategies for evaluating the quality of the rendered sound (Meyer-Kahlen et al. 2024).

An important part of this pipeline remains the question of HRTF personalisation, discussed in § 3.4.1. Recent research on improving HRTF-based binaural rendering increasingly relies on data-driven approaches for personalising HRTFs and enhancing rendering quality. However, several challenges remain before the full potential of ML-based methods can be realised. While more high-quality HRTFs are being measured to support the personalisation research, including the SONICOM HRTF datasest, presented in Chapter 4, data availability is still limited for training ML algorithms, restricting generalisation of the models (Fantini et al. 2025). Furthermore, as discussed in Chapter 3, while HRTFs can be thought of as numerical filters, their function can only be meaningfully described through the perceptual outcomes. Therefore, robust perceptually-motivated metrics, based on spatial hearing research, are needed to improve the development, training, and evaluation of the personalisation algorithms (Ananthabhotla et al. 2021; Daugintis et al. 2023b; Doma et al. 2023a; Geronazzo 2025; Yao et al. 2024).

A useful way of addressing these gaps within the research domain and advancing the state of the art is to engage with the wider scientific community by organising open challenges. It is important to highlight that this advancement of knowledge happens in three main stages during the organisation of a challenge. First, the challenge organisers bring their existing expertise together to identify the key research questions that need to be addressed and, importantly, work on designing a robust evaluation framework to benchmark different proposed solutions. The latter part can be challenging in itself because the absence of good solutions can coincide with the lack of a clear evaluation metric that describes the problem in a succinct and quantifiable way. Second, the challenge participants contribute their ideas by proposing solutions to the defined problems. Finally, the analysis of the submissions leads to further insights into the effectiveness of both the newly proposed methods and the challenge evaluation framework in addressing the identified research gap.

While open challenges and the development of the common task frameworks (CTFs) has been an important driver for AI-related research (Kutz et al. 2025), few examples exist on the collection of standardised datasets, development of evaluation frameworks, and challenge organisation within the domain of spatial and immersive audio. Inspired by similar approaches taken by the hearing aid research (Cox et al. 2023; Dabike et al. 2024), the first Listener Acoustic Personalisation (LAP) challenge was organised to identify, describe, and tackle some of the challenges hindering the success of data-driven HRTF personalisation methods (Geronazzo

et al. 2024). The first LAP challenge comprised two tasks. Task 1 focused on HRTF dataset harmonisation to facilitate data aggregation and large-scale analysis. Task 2 asked participants to upsample HRTFs from a spatially sparse set of measurements to a dense grid.

The following sections summarise the setup and outcomes of both challenge tasks. § 5.2 presents the results of Task 1, with most of the content adapted from Daugintis et al. 2025a. This section highlights the thesis author's major contributions to the development of the challenge evaluation framework, analysis of the submissions, and the discussion of the challenge outcomes, which provide new insights into the use of perceptually-motivated metrics to evaluate HRTF quality. § 5.3 provides a more concise summary of Task 2, based on Hogg et al. 2025, reflecting a more limited contribution by the thesis author to this publication.

To the organisers' knowledge, the LAP challenge was the first initiative in the domain of immersive audio, aiming to incorporate both perceptual and numerical aspects of binaural spatial audio. An important aspect of the LAP challenge organisation was the development of perceptually-motivated HRTF evaluation metrics. The LAP challenge presented an opportunity to test the use of the *barumerli2023* auditory model, presented in § 2.4.1, as a perceptually-informed method to validate the processed HRTFs. This exploration reflected the broader thesis goal of establishing an auditory model-based HRTF analysis framework. The use of the model within the context of the challenge stress-tested its assumptions, practically validating its use in perceptually quantifying HRTF differences, while also revealing some of its limitations that have to be accounted for in the future research.

5.2 LAP24 Challenge Task 1: HRTF Dataset Harmonisation

Section adapted from publication **VII** (Daugintis et al. 2025a).

Task 1 of the LAP 2024 challenge was motivated by the limited availability of high-quality HRTF measurements, restricting the generalisation of ML algorithms for HRTF personalization. Various attempts have been made to use ML methods to, for example, synthesise HRTFs from anthropometric pinna measurements (Yao et al. 2022; Zhi et al. 2022) or upsample a small set of measurements to a dense HRTF grid (Hogg et al. 2024). However, the quality of ML solutions can depend on the quantity of training data (Arbel et al. 2024; Hogg et al. 2023). Due to the laborious and costly HRTF acquisition process, typical measured HRTF datasets contain data from fewer than 100 subjects, but few ML methods attempted to combine data from different datasets (Fantini et al. 2025). Reluctance to merge multiple datasets for ML training can be

attributed to measurable dissimilarities between HRTFs from different databases (Andreopoulou et al. 2015).

As discussed in § 3.2, HRTF acquisition methods contain inherent uncertainties that need to be compensated for. However, existing compensation methods are limited in their ability to remove all setup-specific artefacts due to factors such as unaccounted systemic measurement biases, individual measurement variability, and mathematical limitations of traditional signal processing algorithms. This results in numerically significant differences in spectral magnitudes and variations in estimated interaural time differences between measurements of the same mannequin head at different facilities (Andreopoulou et al. 2015). The heterogeneity in HRTF databases becomes even more obvious for ML algorithms, tasked with classifying HRTFs based on their signals in the time or magnitude domain: Analysis of classification performance using gradient boosted trees reveals that low (below 5 kHz) and high (above 15 kHz) frequency regions are particularly informative to the classifiers (Pauwels and Picinali 2023).

Nonetheless, identifying and extracting dataset-specific artefacts remains challenging due to the complex multidimensional nature of HRTFs across direction, frequency, and listener morphology. Previous proposals on mitigating some of the numerical discrepancies between HRTF datasets include normalising the original HRTFs by the global average and standard deviation (Andreopoulou and Roginska 2011) or by directional averages (Wen et al. 2023) of each HRTF database, as well as using Bayesian fusion (Luo et al. 2013) or extreme frequency smoothing (Lu et al. 2025). However, the perceptual impact of these methods on HRTF quality has not been investigated. Therefore, more extensive perceptual evaluations are needed to understand how the numerical differences between HRTFs relate to perceived differences by human listeners.

To address the HRTF dataset incompatibility issue, Task 1 was dedicated to the matter of HRTF dataset harmonisation. An ideal harmonised HRTF was defined as an HRTF that contains all the perceptually relevant components of its original counterpart but with dataset-specific artefacts removed or mitigated so that its origin could not be detected. This section of the thesis presents the development of the evaluation metrics for the challenge, which aim to combine perceptual and numerical HRTF analysis approaches, summarises the analysis of the proposed solutions, and discusses the limitations of the evaluation framework that became clear after running the challenge.

5.2.1 Metric Selection: Motivation and Goals

To assess the HRTF harmonisation task, a reasonable HRTF evaluation criterion must be found to ensure that harmonised HRTFs retain their relevant perceptual qualities while diminishing the idiosyncrasies introduced by specific acquisition methods. From various perceptual attributes that could be used to describe the quality (and validity) of a harmonised HRTF (Simon et al. 2016), the sound source localisability was chosen as the most important aspect of HRTF-based binaural audio rendering. It is expected that a measured individual HRTF should contain all the necessary acoustic cues for the listener to successfully associate the rendered sound with the correct direction of the sound source, despite the minor levels of noise introduced by the measurement system. Therefore, listener performance should be comparable when localising binaural stimuli spatialised with the original and the harmonised HRTF. It must be acknowledged that ensuring comparable sound source localisation performance might not prevent degradation of other qualitative aspects of binaural rendering, such as the perception of sound colouration or externalisation. However, the objective formulation of sound source localisability via localisation error metrics and its strong connection to other subjective qualities (Jenny and Reuter 2020) provide a reliable framework for evaluating perceptual HRTF validity.

Although the *ground truth* for every perceptual assessment are human responses, designing and performing a robust listening test is time-consuming and labour-intensive. It becomes especially difficult when a large number of qualitative comparisons must be made, leading to problems with the reliability of the results (Andreopoulou and Katz 2016; Kim et al. 2020). To partly address these issues, computational auditory models can be used to estimate human responses and reduce the need for physical tests (Geronazzo et al. 2018). Auditory models that aim to predict human sound localisation performance can provide a useful metric to assess whether the harmonised HRTF has retained its essential perceptual components. As discussed in § 2.4, many of these models work on the principle of template comparison (e.g. Barumerli et al. 2023; Baumgartner et al. 2014): They compare auditory cues (ITD, ILD, and/or spectral cues) contained in the template (e.g. original) HRTF and the target (e.g. harmonised) HRTF. Their output is an estimate of the sound source direction that is on par with human subject responses when performing a sound localisation test with sounds rendered using a target HRTF. The average localisation error across all simulated positions for each HRTF pair provides an interpretable metric framework (Barumerli et al. 2018b) that can be compared to human localization performance (Majdak et al. 2014).

While retaining the perceptual components of the original HRTF that provide accurate

spatial cues to the listener, the harmonised version should ideally not contain any identifiable *signature*, specific to its acquisition system, in its signal. Differences between datasets can be perceptual (that is, identifiable by listeners) and/or numerical (measurable by means of numerical analysis). Although the elimination of perceptual differences is usually the main goal of any HRTF processing method, defining its success is cumbersome due to the aforementioned challenges with performing listening experiments. Compared to human hearing, numerical techniques may exhibit greater sensitivity to HRTF variations and hence provide an efficient and reproducible assessment of harmonisation performance. This, together with the desire to allow participants to check their own solutions, served as motivation to use numerical ML methods tasked with identifying HRTF datasets (Pauwels and Picinali 2023) as evaluation criteria for HRTF harmonisation. Although numerical evaluation methods should ideally align with perceptual models to avoid minimising negligible HRTF differences, the development of such methods requires further research into task-specific perceptual sensitivity.

The two constraints—HRTF perceptual validity and its identifiability—can act in opposition to each other: removing database-specific HRTF artefacts often risks degrading HRTF's perceptual quality. This polarity poses a challenge when defining an optimal evaluation metric that takes both sides into account. Similar considerations could be found in previously organised challenges in the audio processing domain, such as the ICASSP 2023 Clarity Challenge (Cox et al. 2023) and the second Cadenza Challenge (Dabike et al. 2024). In the LAP challenge, a two-stage evaluation process was designed to first assess whether the submitted HRTFs were still perceptually valid and then assess how distinguishable each collection of the harmonised HRTFs is. This procedure is summarised in Figure 5.1.

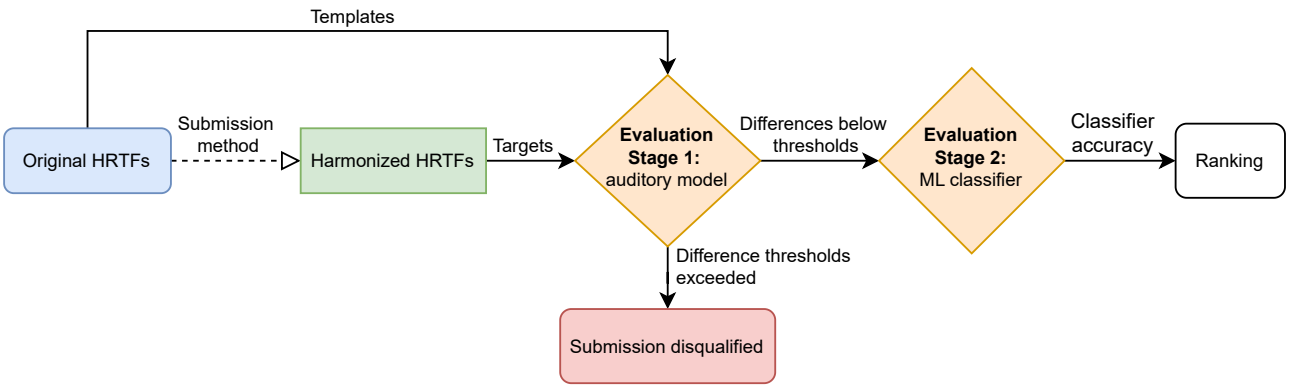


Figure 5.1: The challenge evaluation structure. Reproduced from Daugintis et al. 2025a.

5.2.2 Task Overview

Participants were asked to harmonise HRTFs from eight datasets from different research facilities. Datasets included six collections of acoustically measured HRTFs: 3D3A (Sridhar et al. 2017), HUTUBS (Brinkmann et al. 2019b), RIEC (Watanabe et al. 2014), SADIE II (Armstrong et al. 2018), SCUT (Yu et al. 2018), and SONICOM (Engel et al. 2023); one collection of full head and torso transfer function simulations—CHEDAR (Ghorbal et al. 2020); and one set of pinnae-only transfer function simulations—WiDESPREaD (Guezenoc and Séguier 2020). These datasets differed in sampling rate, HRIR duration, number of directions, and source distance (full details provided in Daugintis et al. 2025a, Table 1). The challenge participants were provided with a collection of 80 original HRTFs (ten from each dataset) and asked to submit their harmonised versions.

Stage 1 evaluation was performed using an auditory sound source localisation model developed by Barumerli et al. (2023) and calibrated according to Daugintis et al. 2023b. The model is presented in more detail in § 2.4.1. It was used to calculate the differences in predicted localisation errors between the original and the harmonised version of the HRTF. The metrics included lateral and polar biases and errors, according to Middlebrooks (1999b) and a polar gain (the slope of the regression line for response vs. target polar angle for positions within 30° lateral angle from the median plane (Macpherson and Middlebrooks 2000)).

Stage 1 evaluation, which aimed to determine if the harmonised HRTF is perceptually valid (i.e. it contains the main individual features of the original HRTF) faced a challenge of determining a validity criteria due to the absence of an error-free reference HRTF dataset to be compared against. Instead, a reasonable maximum allowable deviation of the localisation errors from the original HRTFs had to be established as a validity threshold for the harmonised HRTFs. To this end, the *Club Fritz* HRTF collection, comprising ten HRTFs, each measured using the same mannequin head (Neumann KU100) but in different research facilities (Andreopoulou et al. 2015), was investigated. Differences between the HRTF measurements within this collection provided a useful reference for establishing the thresholds, as it represented an expected variability between high-quality HRTF measurements.

A round-robin modelling exercise was performed using *Barumerli2023* model and different *Club Fritz* HRTFs as templates and targets (Barumerli et al. 2018b). A set of 72 common HRTF grid positions were found across the *Club Fritz* HRTFs (Figure 5.2A) and used in the modelling exercise. The maximum observed differences in the estimated localisation errors across all the HRTFs were then set as the thresholds for maximum allowable deviations between the original

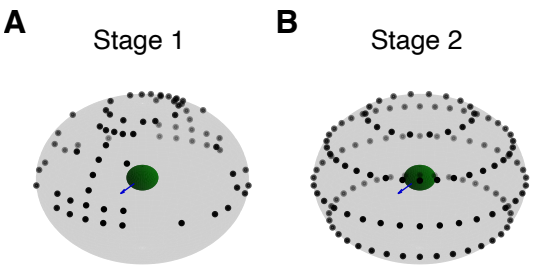


Figure 5.2: HRTF directions used for Stage 1 (A) and Stage 2 (B) evaluation. The central sphere with an arrow represents the position of the head, pointing forwards. Reproduced from Daugintis et al. 2025a.

and harmonised HRTFs. These thresholds are presented in Daugintis et al. 2025a, Table 2. To minimise metric variability due to different grids of the original *Club Fritz* collection and the LAP challenge task dataset, the error metrics were calculated from the model estimates for the 72 positions, closest to the *Club Fritz* common grid positions within each HRTF grid (Figure 5.2A). Participants had to guarantee that at least 64 (80 %) of their 80 harmonised HRTFs had localisation metric differences below the thresholds for their submission to be passed to the second evaluation stage. Otherwise, the submission would be disqualified from the challenge.

The second stage evaluated the effectiveness of the harmonisation based on the principle outlined in Pauwels and Picinali 2023: If HRTFs are perfectly harmonised, an ML classifier should not be able to find any dataset-specific differences between them. To assign a score to the harmonisation, the classifiers were trained by trying to maximise the separability between classes (dataset names in this case). However, the interpretation of the results was reversed: a failing classifier meant a successful harmonisation. Therefore, the chance level of 0.125 (due to the eight different classes) represented perfect harmonisation, while a perfect classification accuracy of 1 represented no harmonisation at all. Minimal harmonisation was applied to the input HRTFs as a preprocessing step to ensure meaningful classification results. This included resampling to the minimum rate (44.1 kHz) and truncating to the minimum HRIR duration (5.333 ms), as in Pauwels and Picinali 2023. A combination of classifiers (support vector machine classifiers with linear and radial basis function kernels, logistic regression, and a decision tree classifier) and input types (time-domain HRIRs, linear HRTF magnitudes, and logarithmic HRTF magnitudes) was employed to ensure the robustness of evaluation. A common grid of source directions was established between all datasets such that the corresponding HRIRs could be directly compared without interpolation (Figure 5.2B). This ensured that the classifier did not rely on the differences in HRTF directions to distinguish between datasets. It is important to emphasise that different grids were used for Stage 1 and Stage 2 evaluations, each selected to ensure

the result consistency within each stage. However, as discussed in the following sections, this choice introduced some limitations to the challenge design.

5.2.3 Summary of the Challenge Outcomes

Three submissions were received: CoWiDeq, which relied on diffuse-field equalisation and dynamic compression, EqPCA-UoA, which was based on principal component analysis, and IOA3D, which employed a neural network architecture (Zhao et al. 2025). All three submissions passed the Stage 1 validation criteria. Two submissions (CoWiDeq and EqPCA-UoA) managed to slightly reduce the the Stage 2 classifier accuracy (0.95 and 0.92, compared to a perfect score of 1 when original HRTFs were used). The third submission (IOA3D) managed to generally deceive the classifier, resulting in a mean classifier accuracy score of 0.27. This submission was declared the winner. More information on the three methods and their evaluation using the challenge evaluation metric are provided in Daugintis et al. 2025a.

Despite successful Stage 1 perceptual validation for all three HRTF harmonisation methods submitted, informal listening tests revealed a significant degradation in binaural rendering quality for some harmonised HRTFs. This effect of the proposed methods on HRTFs was further numerically analysed to understand the limitations and shortcomings of the current challenge setup.

Figures 5.3 A and B present example magnitude spectra along the median and horizontal planes of the original left ear HRTFs and the corresponding harmonised HRTFs from the three submissions. The examples include two acoustically measured HRTFs (SONICOM and HUTUBS) and two simulated ones (CHEDAR and WiDESPREaD). The figure shows a heterogeneous effect that each of the three proposed methods has on the spectral content of the HRTFs. A clear degradation of HRTF features is visible in two of the three submissions (EqPCA-UoA and IOA3D). The EqPCA-UoA method appears to blur sharp frequency-dependent notches in the HRTF magnitude spectra, while the magnitude spectra of the IOA3D solutions contain noise at the positions used by the Stage 2 evaluation (Fig. 5.2 B). This is shown more clearly in Figure 5.3 C, which presents the magnitude spectra of the example IOA3D HRTF (corresponding to the original SONICOM HRTF) at the Stage 2 grid positions only. The figure shows the logarithmic magnitude response input to the classifiers, where each panel represents one of the four horizontal planes at different elevations, comprising the grid shown in Fig. 5.2 B. The figure reveals that when the positions not used by the classifier are removed, the harmonised HRTF spectra lose all the typical HRTF characteristics, such as the frequency-dependent notches, visible in the original

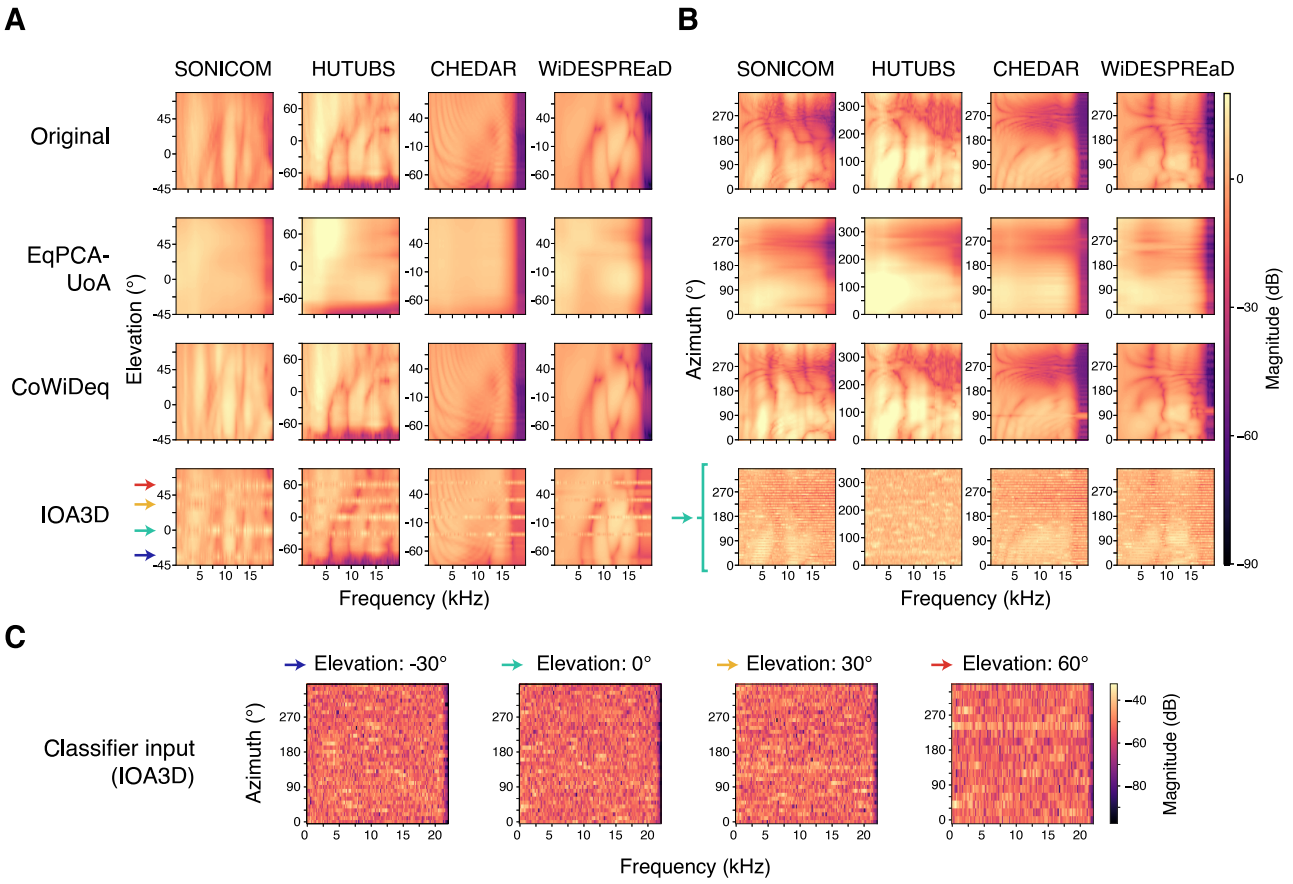


Figure 5.3: A-B: Example original and harmonised left ear HRTF magnitude spectra for positions on the median (A) and horizontal (B) plane from four of the challenge HRTFs. C: Example logarithmic magnitude response input to the Stage 2 classifier of a harmonized SONICOM HRTF using IOA3D method. Four panels contain HRTF positions across four horizontal planes comprising the full Stage 2 grid (Fig. 5.2 B), indicated by the arrows in the full SONICOM HRTF spectra (A, bottom left). The 'Elevation: 0°' panel in C is also a subset of the full horizontal plane spectra (B, bottom left). Reproduced from Daugintis et al. 2025a.

HRTF spectra (Fig. 5.3 B). Further examination of binaural cues revealed minimal changes in ITDs but substantial anomalies in ILDs, particularly in the winning submission, which erased ILD features at the Stage 2 evaluation positions (Daugintis et al. 2025a, Figure 6).

The submission analysis revealed that the winning solution exploited a flaw in the challenge setup that arose from the use of two evaluation grids. To confirm this hypothesis and better understand the effect of the chosen grid on Stage 1 evaluation, the localization metrics were also estimated using the Stage 2 grid, as well as a grid of 25 positions, common between the two grids. In these scenarios, the CoWiDeq method passed the validation thresholds, regardless of the grid used: 71 of 80 CoWiDeq HRTFs fell below the thresholds when using the Stage 2 grid and 64 were below when the common grid was used. However, the other two solutions did not meet the Stage 1 criteria when a different grid was used for the evaluation. For the EqPCA-UoA solution, only 19 HRTFs passed the criteria with the Stage 2 grid and 48 with the common grid,

while 0 IOA3D HRTFs passed the thresholds using the Stage 2 grid and 1 with the common grid.

This analysis of the submitted HRTFs demonstrated some limitations of the harmonisation task design. It exposed important challenges in developing a robust evaluation framework for a task, which lacks the ground truth data to validate the results against. These findings led to an in-depth deliberation on the task definition, metric selection and the challenge outcomes, which provide a useful resource for future research into HRTF dataset harmonisation and perceptual HRTF evaluation. This discussion, presented in the following section, contributes to the thesis goal of determining the strengths and limitations of the use of auditory models as perceptually-motivated HRTF quality metrics.

5.2.4 Discussion

The challenge evaluation integrated the numerical and perceptual aspects of HRTF research, with the aim of a robust assessment of the submissions that reflected the multifaceted definition of the HRTF harmonization procedure. Nonetheless, the plotted magnitude responses and additional result analysis demonstrate limitations of the current challenge structure. This work highlights the complex nature of the HRTF harmonization process that can be summarised by the following polarities:

- acoustic vs. perceptual HRTF definitions;
- interpretable vs. closed-box evaluation methods;
- global (averaged) vs. local (direction-dependent) HRTF evaluation strategies.

These dualities must be addressed in order to improve the LAP challenge framework. They thus underlie further discussion on the challenge evaluation outcomes.

Stage 1 Evaluation

The use of auditory models marked an initial step toward automating the evaluation of harmonised HRTFs, but the challenge results demonstrated that precisely engineered solutions can exploit model limitations. The magnitude response figures suggest that the Stage 1 evaluation is not sufficiently sensitive to the degradation of HRTF magnitude spectra. The blurring of the sharp frequency-dependent notch pattern observed in two of the submissions, although potentially imperceptible in some contexts of horizontal localisation (Kulkarni and Colburn 1998), is likely to impair the perception of sound source elevation, as it removes the main notches associated with vertical sound localisation (Langendijk and Bronkhorst 2002). Yet, the averaging of the error metrics over the full sphere and large thresholds made it impossible for the model-

based evaluation to detect such frequency blurring. Although the thresholds were based on the maximum estimated variability across real measurements, their values for some of the metric differences appear to be very high (e.g. quadrant error difference threshold of 34.56 percentage points (pp) and lateral RMS error difference threshold of 20.71° (Daugintis et al. 2025a, Table 2)). For comparison, Middlebrooks (1999b) reports an increase of less than 30 pp in quadrant errors and less than 10° increase in lateral RMS error for most subjects, localising with non-individual HRTFs compared to individual ones. The loose Stage 1 thresholds suggest that a numerically engineered but perceptually invalid solution (e.g. replacing all HRTFs with one non-individual HRTF) could still pass the Stage 1 thresholds. Furthermore, considering that an average quadrant error in a real localisation experiment with individual HRTFs can be around 20 % (Majdak et al. 2010), the quadrant error difference threshold above 30 pp allows for a harmonised HRTF with an estimated quadrant error above the chance rate (50 %) to pass the criterion for this metric.

Initially, a more strict set of thresholds was imposed for the Stage 1 assessment, corresponding to the third quartiles of the estimated metric differences within the *Club Fritz* collection. However, the thresholds were increased from the third quartile of the metrics differences obtained from the round-robin comparison to the maximum metrics difference after participant feedback, suggesting that perceptually insignificant processing (such as low-passing the HRTFs at 18 kHz) would make harmonised HRTFs fail the Stage 1 assessment. This evidence highlights the limitations of using a model as a surrogate for perceptual tests when applied outside of its intended conditions. Designed for broadband noise sources and realistic individual HRTFs, the model accurately reproduces listener localisation patterns by applying equal weighting to all spectral cues between 700 Hz and 18 kHz (Barumerli et al. 2023). Deviations from these conditions expose the limitations of the model, which could be mitigated by incorporating spectral weighting functions to better reflect vertical localisation patterns (Lladó et al. 2025; Zonooz et al. 2019). While the model provides a useful tool for automating HRTF harmonisation validation, the results reveal its limited view of auditory perception processes, which means that it can be easily misled by engineered modifications.

A potential solution to improve Stage 1 threshold sensitivity involves rethinking the two-step evaluation process. Rather than relying on a binary role of the Stage 1 validation, the perceptual quality of the HRTF could contribute to a combined final metric. This metric would synthesise the results from both stages. For each stage, a normalised representative metric could be derived and then aggregated, either by applying equal weights, akin to the ICASSP 2023 Clarity Challenge, that balanced the importance of speech intelligibility and audio quality criteria (Cox et al. 2023), or using a more intricate weighting strategy that aligns with the priorities of the

challenge. This could result in methods being penalised due to predicted increases in localisation errors despite their good performance according to Stage 2 evaluation.

Stage 2 Evaluation

While the Stage 1 evaluation was designed to offer an interpretable metric, Stage 2 evaluation offered a straightforward numerical evaluation that served the purposes of the challenge, yet lacked interpretability of its results. The traditional ML techniques used in the evaluation process are fairly simple, which allows for some degree of insight when examining the fitted models, particularly for decision tree classifiers (Pauwels and Picinali 2023). Nonetheless, transforming this analysis into a format relevant to the perceptual context and extracting meaningful dataset-specific features that classifiers use to distinguish between HRTF datasets remains a challenge. A future research avenue would be to find a more interpretable model for evaluating HRTF harmonization. Besides providing a sanity check for the classifier performance, its results would aid in understanding the perceptual significance of the disparities among the datasets. Additionally, it could function as a reverse engineering tool, supporting the development of dataset-specific adjustments based on the model's outputs.

In addition to improving interpretability, future enhancements to the classifier should ensure its resilience against deliberately corruptive HRTF alterations. The authors of one of the solutions observed that the classifier's current accuracy could be compromised by introducing pure tone signatures to random HRTFs. They recommended exploring metrics based on perceptual evaluation. Similarly, the winning solution was more tailored towards deceiving the classifier than aligning HRTFs in a perceptual way. Nevertheless, identifying a suitable perceptual metric or input features for ML classifiers requires further investigation, as any proposed data processing might mask undesirable artifacts and could be exploited by proposed harmonization solutions. Generally, the Stage 2 metric should be considered alongside the Stage 1 evaluation, which specifically targets perceptual HRTF quality. A comprehensive set of evaluation metrics would thus be more effective in identifying intentional perceptual HRTF degradation for the purpose of reducing its detectability.

Evaluation Grids

A crucial aspect in designing Task 1 was determining which spatial grid to use for evaluating the submitted solutions. Stage 1 relied on a grid aligned with the *Club Fritz* collection, targeting perceptually relevant angular sectors (e.g. near the median plane) where human behavioural

responses are especially sensitive to spatial cues (Middlebrooks 1999b). In contrast, Stage 2 employed a uniform sampling of the sphere to promote geometric consistency and avoid spatial interpolation across datasets for the classifier. As a result, each approach prioritised different factors (i.e. perceptual relevance versus uniform acoustic coverage) leading to an inconsistency in how HRTFs were evaluated across the two stages.

With the evaluation of the submissions, the grid discrepancy exposed two main issues. First, there was no unified requirement to maintain spatial consistency in HRTF harmonisation across the entire sphere. Second, evaluating harmonised HRTFs over the sphere proved inadequate for capturing local anomalies (e.g. if the distribution of errors is highly variable or skewed, a simple mean can mask substantial inaccuracies in specific regions). Even with a unified grid, behavioural localisation metrics often cover a limited range of positions, typically within 30° of the median plane for polar-based metrics, leaving many directions unexamined (Middlebrooks 1999b). This gap allowed ML models to exploit unused portions of the grid, as demonstrated by the winning solution: By injecting random noise specifically at the Stage 2 positions and retaining perceptually relevant features where Stage 1 metrics were measured, it passed Stage 1 despite producing larger errors in other directions. Meanwhile, globally consistent approaches (two other submissions) appeared to be more robust, but still performed inconsistently when Stage 1 evaluation was conducted using different grids. Consequently, a supposedly *global* solution does not necessarily guarantee local coherence and can be exploited by methods optimised for particular subsets of directions.

In the future, caution is needed to balance perceptual and acoustic considerations on spatial grids in the harmonisation process, reflecting the chosen HRTF definition. If the goal is to preserve perceptual relevance, metrics might focus on directions where human responses are most accurate or behaviourally critical, while still accounting for broader coverage. For instance, a more detailed analysis of quadrant confusion types (Poirier-Quinot et al. 2022), compensation for polar angle compression (Carlile et al. 2014), or the probabilistic representation of behavioural responses (Barumerli and Majdak 2025) could improve the robustness of the evaluation. In parallel, ML-based evaluation might consider weighting errors, depending on spatial cue variability in the region (Middlebrooks 1992) and adopt methods to ensure both local and global consistency. By unifying these perspectives, perceptual and acoustic, future versions of the challenge could effectively address the discrepancies that allowed submitted methods to selectively optimise for one stage at the expense of the other.

5.2.5 Future Avenues for HRTF Harmonisation

Task 1 of the LAP challenge 2024 served as an important step in systematising the HRTF harmonisation problem and merging perceptually informed HRTF evaluation approaches with ML methods. The work generated valuable discussions and highlighted key limitations that will help refine future approaches and advance the field. The difficulties associated with defining a good evaluation metric for this challenge reflect a broader challenge with the reliability of the metrics used as targets (Rodamar 2018). In this context, solutions tend to be optimised for higher metric scores, rather than addressing the challenge as a whole. This is particularly true for ML models, which, driven by numeric loss functions, can exploit evaluation loopholes. Nevertheless, organising multiple iterations of the challenge with refined evaluation would ideally lead to an improved HRTF harmonisation method, stress-testing its metrics during each run. A cyclic procedure, similar to the series of 'Clarity enhancement and prediction challenges' (Graetzer et al. 2020), could be considered, in which a challenge to find the best metric to assess harmonisation would be organised in turn with a challenge to find the best harmonisation procedure.

While listening tests should ultimately be included in challenge evaluation, their design requires further consideration. A fully crossed study, where each participant undergoes HRTF measurements at multiple facilities and evaluates all their harmonised versions, would be financially unfeasible, especially considering a high inter- and intra-subject result variability, requiring a large sample size. Such evaluation would also lack the repeatability needed for rapid, iterative solution testing by participating teams. As an alternative, an initial screening phase using listening tests could be added to the current evaluation setup to ensure a minimum qualitative standard without requiring individual HRTF measurements. However, defining the minimum perceptual validity criteria based on such data remains a challenge.

Generally, several concepts should be clarified before proceeding with the next challenge. Firstly, a clearer task-oriented HRTF evaluation criteria must be established before defining the improved set of metrics. The current iteration focused on localisation accuracy; however, more qualitative aspects, such as externalisation or timbral colouration, could also be accounted for, potentially by using already existing models (Baumgartner and Majdak 2021; McKenzie and Brinkmann 2025). In addition, the choice of the task dataset needs to be examined. The current selection included both measured and modelled HRTF datasets, featuring WiDESPREaD (Guezenoc and Séguier 2020), which only simulated pinnae transfer functions. The challenge aimed to incorporate a diverse range of HRTFs to explore the possibility of consolidating them into a unified HRTF dataset. However, pinnae-only responses fundamentally differ from full

HRIRs, which is evident from considerably smaller ITDs and ILDs, shown in Daugintis et al. 2025a, Figure 6. Therefore, aligning them with other HRTFs while maintaining relative perceptual similarity to their original versions might be an ill-posed task and fall outside the scope of what would typically be considered harmonisation in this context. Overall, given the scarcity of research in the HRTF harmonisation domain to date, considerations like this present an opportunity to establish HRTF harmonisation baselines that would be useful for a wider scientific community.

5.3 LAP24 Challenge Task 2: HRTF Upsampling

A summary of publication **VI** (Hogg et al. 2025).

5.3.1 Background

While Task 1 of the LAP 2024 challenge was concerned with global HRTF differences across datasets, Task 2 concentrated on the individual analysis of each HRTF for the purpose of increasing their spatial resolution. It was motivated by a desire to create more efficient and scalable HRTF measurement pipelines. Traditional high-quality measured HRTFs contain measurements from more than 200 directions, requiring a multichannel loudspeaker rig for an efficient measurement procedure (Engel et al. 2023). However, such dense measurements might be redundant, if direction and subject-dependent trends in HRTFs can be recognised and used in reconstructing HRTF filters between the measured positions, reducing requirements for measurement equipment. If a reliable method is found to upsample HRTFs from a small set of measurements, it could be used to reduce the time and cost of acquiring individual HRTFs, making personalised binaural audio more accessible.

Generally, existing upsampling approaches could be divided into two categories: algorithmic and learning-based (Hogg et al. 2025). They differ in the information used to reconstruct the HRTF at an unknown direction: algorithmic methods interpolate the HRTF based on a superposition of the available measurements only while learning-based approaches are able to condense the knowledge about the patterns in HRTFs based on their training datasets. Algorithmic interpolation methods are commonly used to ensure smooth auditory scene during head rotations in dynamic binaural rendering. The most popular methods include barycentric interpolation, based on a weighted sum of the three (or four, if distance is also considered (Gamper 2013)) nearest measurements (Cuevas-Rodríguez et al. 2019), and spherical harmonic interpolation, based on

HRTF spatial decomposition using spherical basis functions (Arend et al. 2021a). However, these traditional methods generally fail at very low sparsity levels. Data-driven methods can be more powerful, but they require care in designing the models and the training datasets to minimise overfitting and ensure the ability to reconstruct an arbitrary HRTF at an arbitrary position.

While various HRTF upsampling methods have been proposed (Fantini et al. 2025), absence of common task frameworks that include standard test datasets and evaluation frameworks has made it difficult to compare different approaches and identify the most effective upsampling strategies. Therefore, LAP 2024 challenge provided a perfect opportunity to bring together the research community and establish a benchmark for HRTF upsampling methods.

5.3.2 Task Overview

Task 2 of the LAP challenge 2024 asked participants to spatially upsample HRTFs from a small set of measurements to their original dense grid of directions. To ensure that proposed upsampling methods are universal enough to deal with different types of upsampling tasks, the challenge dataset was constructed from an unpublished set of 12 SONICOM HRTFs covering four levels of sparsity: three HRTFs contained 100 uniformly distributed directions, three were composed of 19 uniformly distributed directions, three consisted of five directions concentrated in front of the listener, and the final three had three directions containing front, left, and top measurements only. The original versions of HRTFs were made up of 793 directions and participants were asked to reconstruct HRTFs in their original grid. The publicly-available part of the SONICOM HRTF dataset, consisting of 200 HRTFs at the time the challenge was organised, was available for challenge participants as a training dataset. Full measurement details of the SONICOM HRTF dataset are provided in Chapter 4.

Three common numerical metrics, discussed in § 3.5.1, were selected to evaluate the submissions, based on their widespread use in HRTF upsampling research. They included LSD, the difference in ILD, and the difference in ITD, all averaged over all available directions of the original HRTFs. While perceptually motivated metrics, including *barumerli2023* auditory model, used in Task 1, were also considered for Task 2 evaluation, a well-established set of numerical metrics was chosen instead to ensure the clarity of the evaluation procedure to the participants during the inaugural version of the challenge. The main factor, which facilitated the Task 2 evaluation process, compared to the Task 1, was the availability of the ground-truth data (the original dense HRTFs) that the submissions could be compared against. However, the auditory model was used in further analysis of the challenge outcomes and provided additional insights into the

perceptual quality of the upsampled HRTFs, as discussed in the next section.

In summary, a two-stage evaluation was developed. For Stage 1, maximum error thresholds were set to all three metrics, based on the performance of two baseline algorithmic interpolation methods: barycentric and spherical harmonic. The difference between the original and the upsampled HRTF had to be lower than the thresholds for at least 10 out of 12 HRTFs. This way, the consistency of the method across the 12 HRTFs was evaluated. For those submissions that passed the first stage, their differences in all three metrics were normalised and z-scores used to rank them. The method with the lowest total z-score, summed up over the four sparsity conditions, were deemed the winner of the challenge.

Below, the key findings from the analysis of the submissions are summarised to provide the a more complete picture of the LAP 2024 challenge outcomes. However, full details and the discussion are omitted from the thesis due to a limited involvement of the thesis author to this analysis and its minor relevance to the rest of the work presented here. However, full details on the challenge outcomes are provided in Hogg et al. 2025.

5.3.3 Summary of the Challenge Outcomes

Eight submissions from seven teams were received during the challenge. Seven of them used data-driven approaches and one relied on an algorithmic interpolation method. One submission failed the Stage 1 evaluation and the rest were ranked according to the Stage 2 metrics. The top solutions were learning based and managed to achieve an average LSD below 5 dB, an average ILD difference of around 1 dB, and an average ITD difference of around 20 µs for the lowest sparsity level (i.e. upsampling from 3 measurements) (Hogg et al. 2025, Figure 2).

Further result analysis was performed to understand the effect of challenge evaluation metric choice on the challenge outcomes. Firstly, a spatial variation of the metrics was recognised. While the current evaluation averaged the errors over all directions, some positions, which might be less perceptually relevant in a spatial audio rendering context (e.g. positions at very low elevations), were found to have higher LSD. This analysis suggests that a spatial metric weighting might be beneficial for optimising the evaluation metrics. Furthermore, the effect of using different weights when summing the three metrics on the overall ranking was investigated. While the winning solution was found to be quite robust across all three metrics, some solutions showed a varied performance across them. Therefore, their ranking was strongly affected by the specific evaluation scheme used.

Then, the upsampled HRTFs were evaluated using the *barumerli2023* auditory sound local-

isation model. The evaluation revealed a good performance of the winning solution; however, some solutions underperformed the baseline methods in terms of predicted localisation errors. This analysis shows that while the numerical metrics used in the challenge were sufficient in determining the top submissions, it also highlights the importance of including perceptually motivated metrics in further refinements of the challenge.

Finally, the performance of the Task 2 submissions was compared against the performance of an average SONICOM HRTF. The comparison revealed that at low sparsity levels, the upsampled HRTFs tend to approach the average HRTF. While the winning solution generally outperformed the average SONICOM HRTF in the evaluation metrics, the latter produced lower errors than some of the other submissions, especially at higher upsampling (lower sparsity) levels. This analysis highlights the challenge in encoding the individual nature of HRTFs from a small set of measurements.

In general, the evaluation of LAP challenge Task 2 was more effective than that of Task 1 in benchmarking the submissions due to the availability of the ground truth—the original HRTFs. Still, the result analysis highlighted the need for further refinement of the evaluation procedure to ensure its robustness and perceptual relevance. This finding further reinforces the main aim of this thesis of developing a robust perceptually motivated HRTF analysis metric. Future iterations of the challenge should consider including the metrics for spatial smoothness of the upsampled HRTFs as well as a perceptual assessment based on listening tests. Nonetheless, the challenge outcomes demonstrated the potential of data-driven methods in HRTF upsampling and provided a benchmark for future research in this area.

5.4 Chapter Summary and Conclusions

This chapter overviewed recent efforts to engage with the broader scientific community and address existing challenges in data-driven HRTF personalisation research. The work was conducted within the remit of the 2024 Listener Acoustics Personalisation (LAP) challenge, which addressed two priority tasks: harmonisation of HRTFs from different datasets and upsampling of HRTFs from a small set of measurements. The development of the challenge evaluation framework required consolidating knowledge from spatial hearing, auditory modelling, and numerical HRTF analysis, presented in Chapters 2 and 3, in order to design robust perceptually motivated metrics and assess the perceptual quality of the processed HRTFs.

The challenge presented an opportunity to explore the use of the *barumerli2023* auditory

model as a perceptually motivated HRTF evaluation metric. The model was especially useful for the HRTF dataset harmonisation task, which was a more challenging scenario compared to the HRTF upsampling task due to the inherent lack of ground truth data. While the auditory model proved to be a useful tool for automating the evaluation of harmonised HRTFs, the challenge outcomes also revealed some of its limitations. It highlighted the need for further refinements of the evaluation framework to robustly account for the directional and spectral HRTFs variations. This exploration provides valuable context for the further development of the non-individual HRTF evaluation framework based on the *barumerli2023* auditory model, presented in Chapter 6 and evaluated in Chapters 7 and 8.

Chapter 6

Auditory Model-Based HRTF Similarity Metric

This chapter presents the development and calibration of the auditory model-based HRTF similarity metric. § 6.1–§ 6.5 are adapted from the conference publication **II**, presented at the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (Daugintis et al. 2023b). § 6.6 shows further metric analysis, comparing it with previously reported perceptual HRTF rating method. It is based on a conference publication **V**, presented at the 2025 Forum Acusticum Euronoise (Daugintis et al. 2025b).

6.1 Introduction

§ 3.4 discussed the importance of using individual HRTFs for the most accurate sound localisation and realism of binaurally rendered spatial audio experiences. However, high-quality HRTF measurements require specialised equipment and controlled environments, such as the setup presented in Chapter 4, making them impractical for widespread use. Instead, various methods have been proposed to personalise HRTFs when individual measurements are not available, as overviewed in § 3.4.1, some of which are based on selecting a non-individual HRTF. The latter methods rely on a chosen HRTF similarity metric, which can be qualitative (e.g. based on subjective human ratings (Katz and Parseihian 2012)) or quantitative (e.g. spectral distance (Middlebrooks 1999a), geometric pinnae matching based on notch frequency analysis (Geronazzo et al. 2014, 2020), or computer vision techniques (Zhi et al. 2022)).

While the subjective tests are known to produce inconsistent results, especially for naïve listeners (Andreopoulou and Katz 2016; Kim et al. 2020), the quantitative numerical similarity

metrics, summarised in § 3.5.1, lack the perceptual foundations to describe the quality of the HRTF fit in a meaningful way (they may be well suited for machine audition aiming for the 'best localisation' but not necessarily a human one). More complex numerical metrics, including ML approaches, have been used to develop a more perceptually informed HRTF distance metric based on human sound localisation data (Ananthabhotla et al. 2021). Other HRTF personalisation studies used machine learning techniques with psychoacoustic input in stages of, e.g. feature generation (Qi and Wang 2019) or dimensionality reduction (Grijalva et al. 2016). However, as exemplified in Chapter 5, the HRTF selection problem is constrained by limited datasets and perceptual complexities of the auditory system, making it difficult to develop robust and generalisable ML models. Therefore, domain knowledge can help develop more efficient and precise automated HRTF fitting procedures.

A few quantitative non-individual HRTF selection studies have used computational auditory models to evaluate the quality of spatial processing without extensive listening tests (Geronazzo et al. 2018; Warnecke et al. 2022). The human sound localisation models used in these studies estimate the localisation error, focusing on the sound direction along a sagittal plane, where listener uncertainty is the most impactful (Baumgartner et al. 2013, 2014). As summarised in § 2.4, these auditory models rely on the assumption that the auditory system has learned the mapping between spatial features and sound direction (Middlebrooks 1992). This procedure, known as template matching, enables the subject to estimate the source direction from the observed spatial features. To account for the variation across humans in distinguishing the sound localisation cues, the models have a set of parameters, which can be tuned based on individual sound localisation data to accurately predict individual localisation performance. Nevertheless, since collecting the individual localisation data is time-consuming, previous model-aided HRTF selection studies relied on average non-individual model parameters (Barumerli et al. 2018a; Geronazzo et al. 2018). However, the impact of additional modelling uncertainty due to such model simplification on an HRTF classification task has not yet been addressed.

Emphasising the synergy between human (perceptual) and machine (numerical) approaches (Heller et al. 2023), a set of automated non-individual HRTF classification criteria, which acts as a perceptually motivated HRTF similarity metric, is proposed based on the sound localisation performances simulated with an individually-calibrated computational auditory model, developed by Barumerli et al. (2023). The general *barumerli2023* model structure was already summarised in the § 2.4.1. Furthermore, its use in perceptually validating HRTF measurement consistency was presented in § 4.3.3, while its application in evaluating perceptual HRTF differences was discussed in Chapter 5. However, Chapter 5 also revealed some important considerations when

using the model for perceptual HRTF analysis, including the directional variability of modelled localisation errors and the importance of validating the model against perceptual data. Building on these insights, this chapter provides more in-depth investigation into the model, its parameters, and the use of its predictions for developing the HRTF similarity metric.

Firstly, the model calibration using previous sound localisation data is presented in § 6.3. Then the proposed HRTF classification procedure is described in § 6.4. Due to the complexity of sound localisation error distributions over different directions around the listener, a novel methodology that accounts for the error distributions within a range of directions in front of the listener was implemented, rather than aggregated global error values, to classify non-individual HRTFs as *good* or *bad* and select one *best*/*worst*, representing the extremes of the proposed metric. § 6.5 discusses the validity of using the proposed selection methodology with individual and non-individual (averaged) model parameters. Finally, the selection based on the proposed method is compared to a previously conducted perceptual HRTF rating in § 6.6.

6.2 Data and Performance Metrics

Data from the AMT collected as part of two previous studies was used (6 subjects from Majdak et al. (2010) and 11 from Majdak et al. (2013)). The datasets contain subjects' individual HRTFs and their responses from a sound localisation test using short broadband noise bursts in a virtual reality environment. In each simulation of the localisation test, the *barumerli2023* model (Figure 2.2) was supplied with the individual HRTF as the template and one of the HRTFs as a set of target inputs.

To separate the directions, dominated by the binaural vs monaural cues, an interaural-polar coordinate system was used for localisation data analysis, shown in Figure 2.1. Three error metrics were used to analyse the modelled responses, based on Middlebrooks 1999b: PE, QE, and LE. Their mathematical formulation is provided in Equations 2.2–2.4 of § 2.3.1.

6.3 Model Calibration

The free model parameters (σ_{itd}, σ_{ild}, σ_{mon}, σ_{prior}, and σ_{motor}, described in § 2.4.1) were tuned based on the individual data from the real localisation tests. The model parameters for the first five subjects had already been reported in Barumerli et al. 2023. Following that calibration procedure, the model was fitted to additional eleven subjects using the greedy algorithm,

Algorithm 6.1 Greedy *barumerli2023* calibration algorithm. © 2023 IEEE.

$\sigma_{ild} = 1, \sigma_{mon} = 5, \sigma_{motor} = 10$ ▷ Initialise parameters
Calculate $\epsilon_{LE}, \epsilon_{PE}, \epsilon_{QE}(RAU)$ ▷ Quick model setup
while $\epsilon_{QE}(RAU) \geq 0.2$ **do**:
  σ_{mon} −= sign $\epsilon_{QE}(RAU)$
  Calculate $\epsilon_{LE}, \epsilon_{PE}, \epsilon_{QE}(RAU)$ ▷ Quick setup
  if $\sigma_{mon} \leq 2$ **or** $\sigma_{mon} \geq 8$ **then break**
while $\epsilon_{PE} \geq 0.15$ **do**:
  σ_{motor} −= sign ϵ_{PE}
  Calculate $\epsilon_{LE}, \epsilon_{PE}, \epsilon_{QE}(RAU)$ ▷ Quick setup
  if $\sigma_{motor} \leq 5$ **or** $\sigma_{motor} \geq 25$ **then break**
if $\epsilon_{LE} \geq 0.1$ **then**:
  Check if ϵ_{LE} reduces with $\sigma_{ild} = 0.5$ ▷ Quick setup
  while $\epsilon_{LE} \geq 0.1$ **do**:
    σ_{motor} −= sign ϵ_{LE}
    Calculate $\epsilon_{LE}, \epsilon_{PE}, \epsilon_{QE}(RAU)$ ▷ Quick setup
    if $\sigma_{motor} \leq 5$ **or** $\sigma_{motor} \geq 25$ **then break**
Calculate $\epsilon_{LE}, \epsilon_{PE}, \epsilon_{QE}(RAU)$ to validate ▷ Full setup

formalised in Algorithm 6.1, which was empirically developed to produce the most stable results (Barumerli et al. 2023). For each subject, the model was supplied with individual HRTFs as both the template and the target, and relative differences between aggregated real and simulated errors, ϵ_{PE}, ϵ_{LE}, and $\epsilon_{QE}(RAU)$, were computed. For the calibration procedure only, QE percentage values were transformed to a scale comparable to the other two angle-based metrics using a rationalised arcsine transform (RAU) (Studebaker 1985).

The values of $\sigma_{itd} = 0.569$ and $\sigma_{prior} = 11.5°$ were set based on derivations from previous psycho-acoustics experiments (Ege et al. 2018; Reijniers et al. 2014). σ_{mon}, σ_{motor} and σ_{ild} were tuned in the specified order to reduce ϵ's below previously reported thresholds (Barumerli et al. 2023). A quick model setup with five repetitions was used during the adjustment process to reduce computational time, while a full setup (300 repetitions) was run at the end to validate the calibration. Termination conditions were added to keep the parameters within stable bounds, but none of the subjects exceeded them. Once the automated calibration was finished, the parameters were fine-tuned to one decimal place to reduce the ϵ's further by running the model manually (a more precise automated calibration procedure is subject to future research).

6.4 HRTF Classification Procedure

Localisation errors were modelled for every subject using each of the 16 HRTFs (1 individual and 15 non-individuals). The calibration procedure, presented in § 6.3, was performed using

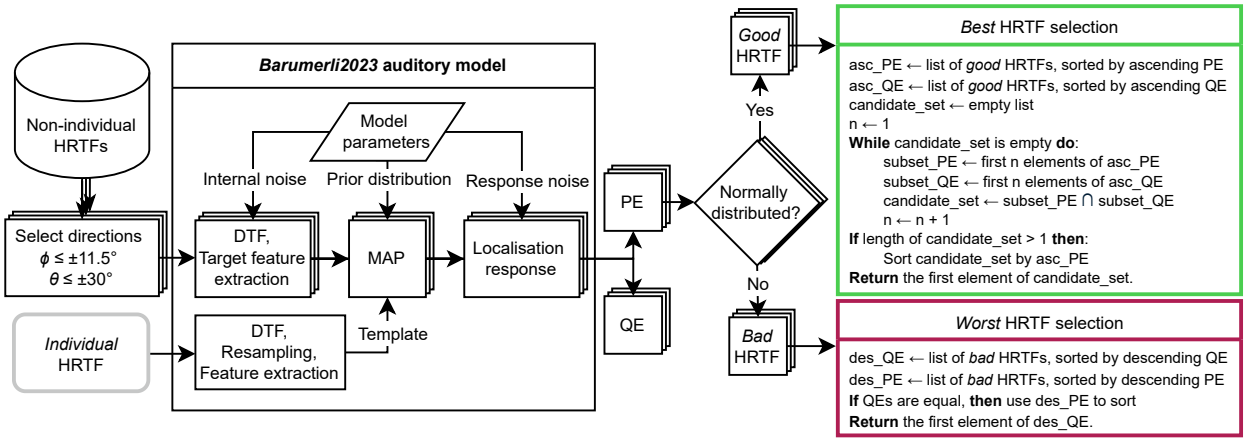


Figure 6.1: Non-individual HRTF classification procedure. *barumerli2023* auditory model is supplied with the individual HRTF as a template, which it compares against different non-individual HRTFs within a limited range of directions using Bayesian inference (MAP) and estimates localisation responses for each. The responses are analysed and compared using PE and QE, classifying them into *good* and *bad* and selecting the *best* and the *worst* from the groups. ∩ represents intersection. *Barumerli2023* auditory model structure is adapted from Figure 1 in Barumerli et al. 2023.

global errors averaged over all the simulated directions and modelling iterations. For the HRTF classification procedure it was interesting to investigate the spread of localisation errors for different directions around the listener, so that the robustness of the predicted global errors across target HRTFs could be evaluated. Therefore, for the classification procedure, the errors were aggregated for each of the direction over all the model iterations (300 in total), resulting in 16 distributions per subject per error. The resulting full sphere PE distributions appeared to be multimodal for different areas around the listener, so the directions of interest were limited to a section in front of the listener within $\pm 11.5°$ polar angle, equivalent to σ_{prior}. In this range, a more direction-independent localisation performance is expected, according to Ege et al. (2018). A hypothesis was made that a *good* non-individual HRTF would correspond to a normally distributed PE within this frontal area. At the same time, skewed or multimodal distribution would indicate that a modelled listener cannot effectively use the localisation cues embedded in the HRTF and might rely too much on prior beliefs. Since LE shows less variability among non-individual HRTFs (Romigh and Simpson 2014), the method focused on PE and QE. The classification was performed in the following steps **for each subject**, separately (summarised in Figure 6.1):

1. divide target HRTFs into *good* ($p > 0.05$) and *bad* ($p < 0.05$) based on the Shapiro-Wilk test on the PE distributions of the 15 simulated localisation tests;
2. ***good* HRTFs**: order PE distributions of *good* HRTFs by rms of the distribution and their QE distributions by the 3rd quartile, respectively;

3. ***bad* HRTFs**: order QE distributions by the median.

The *best* and the *worst* non-individual HRTFs for each subject were selected from the 15 available using the following criteria:

1. **The *best* HRTF**: Look for a common HRTF in top $n = 1$ *good* PE and QE distributions. Increase n until the intersection is found.
2. **The *worst* HRTF**: select the highest median QE; If multiple HRTFs have the same highest median QE then base the selection on the 3rd quartile, or rms PE value if QE distributions are identical.

Finally, the classification procedure was repeated for each subject using non-individual (median) model parameters to investigate how the results are affected when individual model parameters are not used.

6.5 Classification Results and Discussion

Table 6.1: Model parameters for 11 individual subjects, and median parameter values across 16 subjects. © 2023 IEEE.

Subject	Parameters		
	σ_{ild} (dB)	σ_{mon} (deg)	σ_{motor} (deg)
NH12, 15-18	*available from Barumerli et al. 2023*		
NH39	1	5.4	11.8
NH43	1	2.2	20.5
NH46	1	3.2	13.8
NH53	1	2.4	11.1
NH55	0.5	5.1	17.5
NH58	0.5	3.1	9.2
NH62	1	7.2	13.5
NH64	0.5	4.7	13.2
NH68	0.5	7.5	13.4
NH71	0.5	4.5	13.6
NH72	1	6.9	11.1
Median	**0.75**	**4.3**	**13.45**

Table 6.1 reports individual model parameters for 11 of the subjects used in the study, obtained from the calibration procedure detailed in § 6.3. Median parameter values across all 16 subjects (including 5 from the previous study), used to test the selection methodology with non-individual model parameters, are also presented.

Figure 6.2 shows an example of modelled PE and QE distributions across different target HRTFs for subject NH39, using individual model parameters. The figure presents error distri-

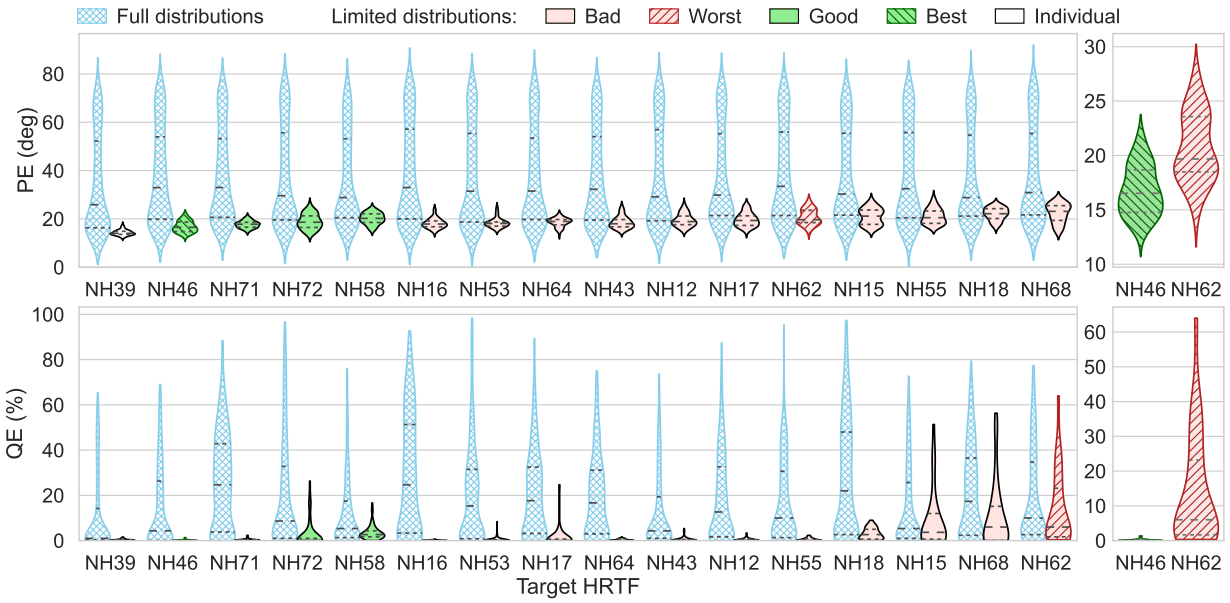


Figure 6.2: Violin plots of modelled error distributions for subject NH39 with different target HRTF. Dashed lines indicate quartiles. Full distributions across the listener sphere are plotted next to the distributions within a limited range of $|\phi| \leq 11.5°$ in the front. The non-individual HRTF are categorised and arranged from left to right based on PE and QE values, respectively, according to the proposed procedure (§ 6.4). The limited distributions for the *best* and *worst* HRTF are plotted separately on the right. © 2023 IEEE.

butions for all directions (within $|\theta| \leq 30°$ as per error definitions) and the limited ones in the front, colour-coded based on the PE normality criterion. It also includes distributions from the individual HRTF which, as expected, are generally smaller than from the non-individual HRTFs.

Generally, medians of direction-limited PE and QE distributions for *good* non-individual HRTFs are one of the smallest ones across the simulated HRTFs. The imposed normality criteria may sometimes place HRTFs with median PE, similar to the *good* HRTFs, in the *bad* category (e.g. NH16, NH53, or NH64). However, when the full-sphere error distributions are considered, the *bad* HRTFs appear to be associated with higher overall QE rates. This could suggest that the strong evidence of bimodality of the direction-limited PE distribution may act as a good indicator for a higher QE in a full localisation test with the particular target HRTF.

The error distributions of the *best* and the *worst* target HRTFs for subject NH39 are also plotted separately in Figure 6.2. The bimodal nature of the PE distribution for NH62 is clearly visible in the violin plot, while the evidence for the bimodality of the NH46 is not strong enough for the normality hypothesis to be rejected. NH46 is also associated with negligibly small QE, compared to NH62.

Figure 6.3 summarises the classification results for all 16 subjects using both individual and non-individual (median) model parameters. The number of HRTFs classified as *good* varies considerably across subjects. For instance, when considering the individualised predictions, 11

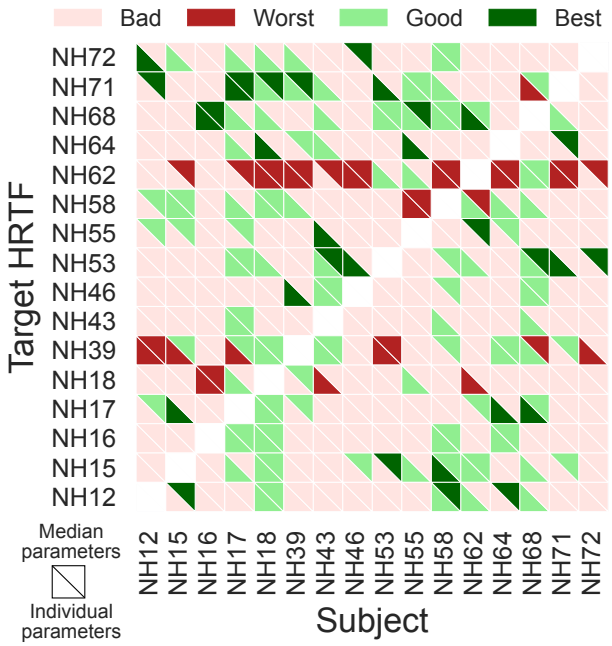


Figure 6.3: Non-individual HRTF selection using individual (bottom left half of the cells) and median (top right half) model parameters. © 2023 IEEE.

out of 15 non-individual HRTFs are classified as *good* for subject NH18, while none passed the normality criterion for subject NH72. Although the results might be affected by the limited HRTF selection pool, some subjects appear to be responsive to a wider range of sound localisation cues than others. However, the phenomenon is not reciprocal, i.e. HRTF of a responsive subject is not generally categorised as *good* for other subjects. In fact, NH18 is marked as the *worst* for 3 of the subjects. The addition of noise across various stages of the auditory model makes the classification non-commutative. This asymmetry is in line with the outcomes from previous perceptual studies (Cuevas-Rodriguez et al. 2021; Katz and Parseihian 2012). On the other hand, some HRTFs are often selected as the *worst* (e.g. NH39 and NH62). The repeated classification suggests that those HRTFs may have less distinguishable spectral cues compared to the rest of the dataset and could be discarded in a process of database optimisation for non-individual HRTF selection (similarly to Geronazzo et al. 2018).

When the non-individual model parameters are used, the resulting HRTF selection is affected, as indicated by the colour mismatch within each cell of Figure 6.3. Generally, the selection methodology for the *worst* HRTF appears to be more robust to model parameter changes than the one for the *best*, indicating a relationship between the parameters and the shape of the PE distributions. The HRTF categorisation is also more robust to parameter changes for some subjects. For example, the selection for NH16 does not depend on the model parameters, while 8 HRTFs are misclassified for NH68, including the *worst* HRTF, which is considered *good* if

individual model parameters are used. A more systematic analysis of the effect of the parameter change on the HRTF selection is required in the future. Nevertheless, the results demonstrate that one must take care when interpreting results from sound localisation models without individually calibrated parameters.

6.6 Comparison with Perceptual HRTF Rating

In this section, results are compared between the HRTF selection based on the discussed classification procedure and a perceptual rating (Katz and Parseihian 2012). The ability to explain subjective rating results with the use of an auditory model would provide a validation of the latter to be employed for numerical HRTF selection.

6.6.1 Subjective Listening Test

The procedure for the subjective HRTF rating, based on a listening test, is detailed in Katz and Parseihian 2012. In summary, 45 subjects participated in a test in which they rated 46 HRTFs[1]. The participants listened to a series of repeated noise bursts, rendered in horizontal (every 30°) and vertical (every 15°) trajectories with a corresponding HRTF. They had to rate each HRTF according to how well the rendering corresponded to the trajectories, assigning them one of three available ratings: 'bad', 'ok', or 'excellent'. For the purposes of this study, the results of 44 out of 45 subjects and 45 out of 46 HRTFs are used because one of the HRTFs was unavailable.

6.6.2 Model-Based HRTF Selection

For the purposes of this study, a selection of the *best* and the *worst* non-individual HRTF was made by comparing the predicted PE and QE, across different non-individual HRTFs, as described in § 6.4. Initially, the original classification methodology, which only considers errors for sound source directions in front of the listener, was used (referred to as the *original* method). Since the subjective test was conducted using the specific directions along the two trajectories, a second version of the model selection was made, where errors were calculated using the specific directions used in the listening test (the *trajectory* method). In this case, error distributions were not analysed, so only the *best* and the *worst* HRTF were selected: The *best* HRTF was selected from the lowest PE and QE (as described in § 6.4) and the *worst* was defined as the HRTF which

[1]The set included individual HRTFs, which were indicated to the participants. However, individual HRTF ratings are not considered in this study.

resulted in the highest quadrant error rate estimate. In both cases, only non-individual HRTFs were used in the selection procedure, because including individual HRTFs would have resulted in them being selected as the *best* HRTFs.

6.6.3 Results

Figure 6.4 reproduces the subjective rating results, presented in Katz and Parseihian 2012, Figure 1, for non-individual HRTFs, considered in this study. In total, 17.9 % of the time the HRTFs were rated as 'excellent', 42.0 % as 'ok' (making 59.9 % of subjective ratings at least 'ok'), and 40.1 % as 'bad'. These percentages indicate the chance level for selecting an HRTF to be in one of the categories.

Figure 6.5a shows model-based selection, following the *original* method. To understand the level of agreement between the two results, Figure 6.6 presents the *best* and the *worst* HRTFs according to the original model selection and their subjective rankings from the listening test. Comparing the model-based selection with the perceptual ratings, the *best* HRTFs according to the model were rated as 'excellent' 15.9 % of the time while 59.1 % of the time they were rated to be at least 'ok'. These levels are similar to chance, indicating poor agreement between the model and the subjective test. On the other hand, the *worst* HRTFs, according to the model-based selection, were rated as 'bad' in the perceptual test at 52.3 %, indicating an above chance level of correspondence between the two methods.

Figure 6.5b shows the results of the selection when the directions along the trajectories used in the subjective test were considered. Figure 6.7 then shows the corresponding subjective ratings for the *best* and the *worst* modelled HRTFs. In this case, 29.5 % of the *best* HRTFs were rated as 'excellent' in the listening test and 77.3 % were rated to be at least 'ok'. These values are above the estimated chance level, indicating some correspondence between the modelled and the subjective results. However, 45.5 % of the *worst* HRTFs were rated as 'bad' in the subjective test, which is only slightly above chance level and lower than with the *original* method.

The method for selecting one *best* HRTF intentionally omitted individual HRTFs, because they would be selected for most participants. However, in the case of directions along the trajectory, the selection was also made using a modification were two *best* and *worst* HRTFs were selected, including the individual one (shown in Figure 6.8). Figure 6.9 reveals the subjective rankings of the model-selected HRTFs. In the case of top two HRTFs, the matching rate increases due to the individual HRTFs. However, the accuracy of the 'bad' selection decreases.

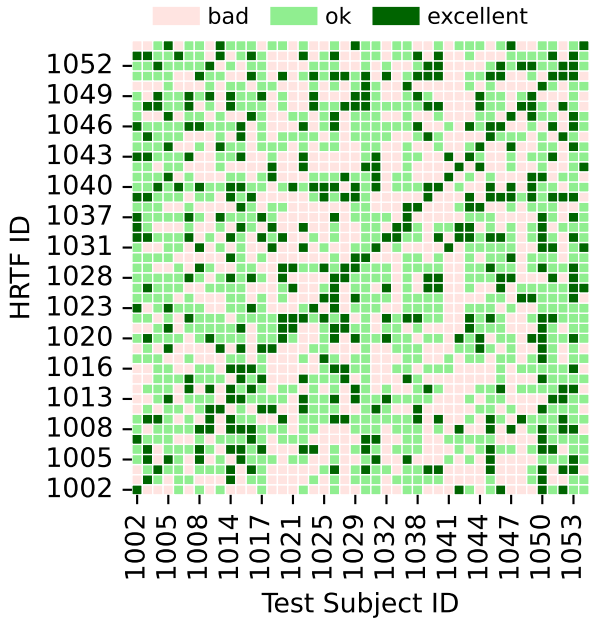
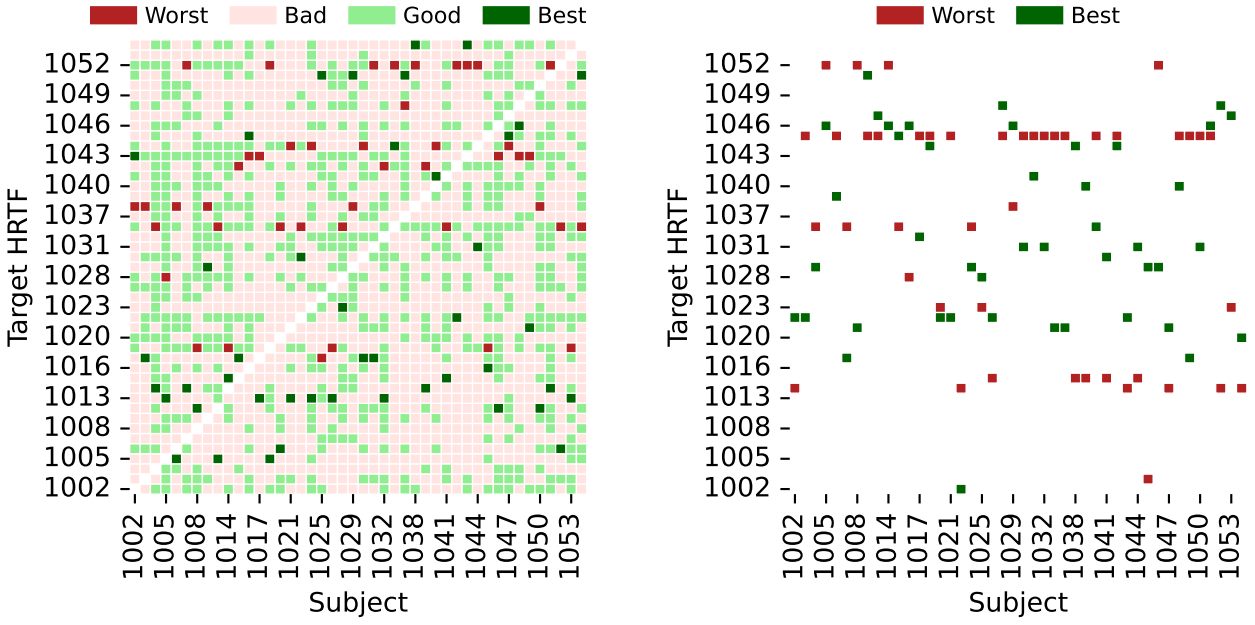


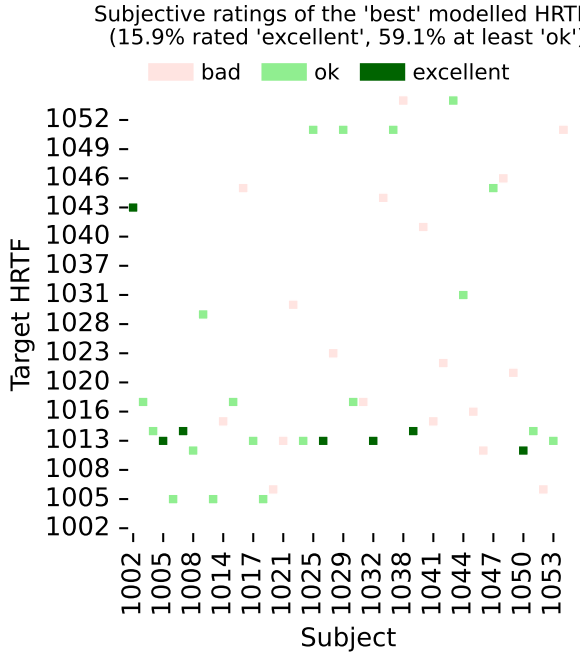
Figure 6.4: Subjective non-individual HRTF ratings. Figure reproduced from Katz and Parseihian 2012.



(a) The *original* method.

(b) The *trajectory* method.

Figure 6.5: Auditory model-based HRTF selection.



(a) Selected *best* HRTFs.

(b) Selected *worst* HRTFs.

Figure 6.6: Selected HRTFs using the *original* method and their associated subjective ratings (i.e. colours indicate the ratings, presented in Figure 6.4).

(a) Selected *best* HRTFs.

(b) Selected *worst* HRTFs.

Figure 6.7: Selected HRTFs using the *trajectory* method and their associated subjective ratings (i.e. colours indicate the ratings, presented in Figure 6.4).

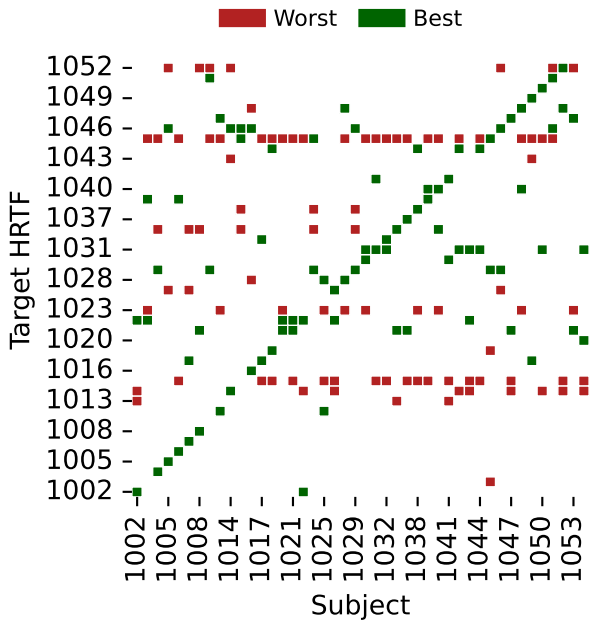
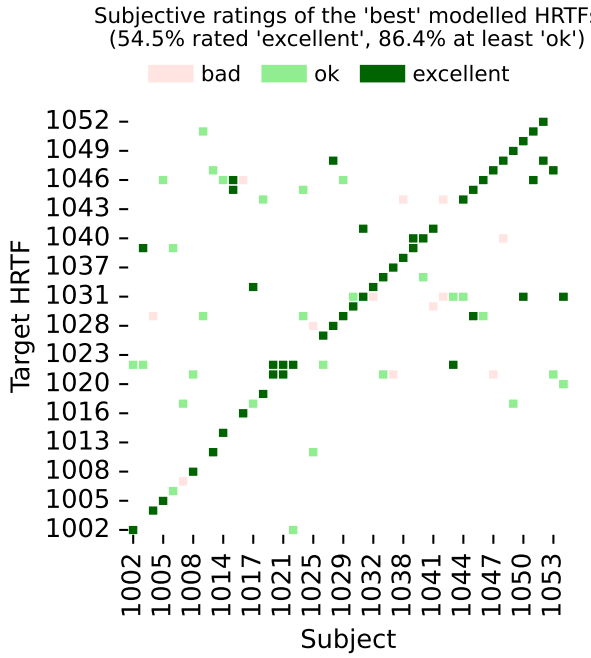
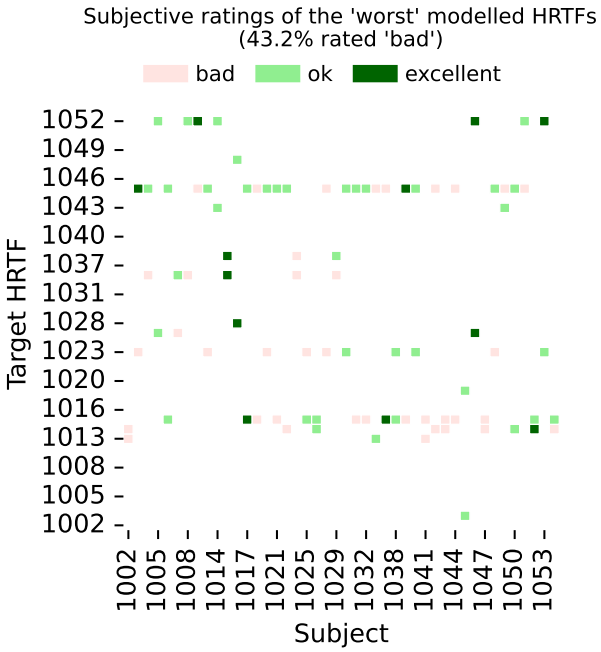


Figure 6.8: Model-based selection of two *best*/*worst* HRTFs, including the individual HRTFs, using the *trajectory* method.



(a) Selected *best* HRTFs.



(b) Selected *worst* HRTFs.

Figure 6.9: Selection of two *best*/*worst* HRTFs for each participant, using the *trajectory* method, and their associated subjective ratings (i.e. colours indicate the ratings, presented in Figure 6.4).

6.6.4 Discussion

The limited success of the modelling results can be attributed both to the challenges of consistency and repeatability of the subjective listening tests (the task of rating 46 HRTFs one after another in a consistent fashion is challenging, especially for less experienced participants (Andreopoulou and Katz 2016)) and to the limitations of the model-based selection procedure. The auditory model was specifically designed to replicate a static sound localisation task. However, the subjective test used predefined trajectories to present the HRTFs, so the participants were performing a comparison between the expected sound location and the binaural rendering instead of a static blind localisation task. Studies comparing similar HRTF assessment methods and metrics indicated that the correlation between different methods depended on the specific metrics used (Zagala et al. 2020) and the listeners who performed the tests (Kim et al. 2020).

The metrics used for the model-based selection might have to be reconsidered to better correspond to the perceptual task. For example, quadrant ambiguity might play a smaller role when listening to a sound source from a known direction, which would explain the poor agreement between the *worst* HRTFs, selected based on QE, and their subjective ratings. Instead, a polar error might be a better metric for the selection of the *worst* HRTF. Furthermore, the model, which uses multiple noise parameters to predict individual localisation responses based on real sound localisation performance, was not calibrated to each subject due to the unavailability of the required localisation data. Therefore, individual-level predictions might be inaccurate. Finally, the interaural time differences (ITDs) of the HRTFs used for the subjective test were altered to match each subject's ITD, whereas the HRTFs used in the model selection had unaltered ITDs. Model-based selection only considered errors along the confusion cones, mainly dependent on spectral cue mismatch (Romigh and Simpson 2014), while the perceptual judgement of horizontal trajectory rendering could have been dominated by binaural cue matching, not taken into account by the chosen estimated metrics.

In summary, the model-based HRTF selection, presented in § 6.4, was not able to predict perceptual HRTF quality, according to the subjective ratings from Katz and Parseihian 2012. However, the modified model-based selection, which accounted for the specific directions used in the subjective test showed an improved agreement with the perceptual ratings for the *best* HRTFs. While the improvement suggests that the predicted localisation error along the specific directions used in the subjective test could be a partial indicator of perceived HRTF rendering quality, limited results can be attributed to differences between the model's intended scenario and the exact setup of the subjective test and repeatability challenges of a subjective listening

test. This comparison shows the importance of considering the specific perceptual task when designing numerical HRTF selection procedures to ensure that the most relevant aspects of HRTF quality are acounted for in that scenario.

6.7 Chapter Summary and Conclusions

Following the targeted applications of the *Barumerli2023* Bayesian sound localisation model presented earlier, including assessment of HRTF measurement consistency in Chapter 4 and perceptual quality of harmonised HRTFs in Chapter 5, this chapter offered a deeper analysis of the model itself and its use in non-individual HRTF evaluation. It first introduced the model calibration procedure, based on previously collected localisation data from 16 normal-hearing listeners. The chapter then detailed a classification framework for establishing perceptual similarity of non-individual HRTFs. The framework classifies non-individual HRTFs into *good* and *bad* categories based on predicted polar and quadrant error distributions within a limited frontal area and selects one *best* and one *worst* non-individual HRTF for each subject. The robustness of this procedure to variations in model parameters was examined, and the selection was compared against subjective ratings gathered in an earlier listening experiment.

The comparison between model-derived HRTF selections and listener ratings showed that the predictive power of the model depends strongly on the perceptual task in question, suggesting that different aspects of HRTF quality could be important in different listening scenarios. As a result, a comprehensive evaluation of the metric is required to determine its validity as well as its limitations. The remainder of the thesis therefore focuses on this perceptual assessment. Chapter 7 presents a direct validation of the model-based selection using a static localisation experiment and investigates how localisation performance relates to perceived rendering quality. Chapter 8 extends the evaluation to an indirect task, based on intelligibility of binaurally rendered reverberant speech, to examine the role of HRTF selection in more complex conditions. Together, these chapters provide an extensive perceptual evaluation of the auditory model-based metric and assess the importance of HRTF personalisation across diverse scenarios.

Chapter 7

Perceptual Evaluation of the HRTF Similarity Metric

This chapter presents the perceptual evaluation of the metric, discussed in Ch. 6. The text is adapted from the manuscript **VIII** that is currently undergoing a peer review process with the Journal of the Acoustical Society of America.

7.1 Introduction

Chapter 6 presented a novel auditory model-based similarity metric for selecting the *best* and the *worst* non-individual HRTFs based on their predicted sound localisation performance compared to an individual HRTF. However, comparing the metric's predictions against prior perceptual data revealed challenges arising from variability in experimental conditions and methodologies across different studies. Therefore, a dedicated perceptual evaluation is needed to establish the validity of the proposed metric.

This chapter aims to evaluate the proposed HRTF similarity metric, introduced in § 6.4, through listening experiments. It presents a multitask perceptual evaluation comprising a static sound localisation test and a dynamic spatial audio quality assessment. A cross-experiment analysis establishes a statistical framework for assessing HRTFs by focusing on pairs with extreme perceptual dissimilarity, as predicted by the metric. This framework enables the identification of correlations between acoustic features, localisation performance, and qualitative perceptual attributes, helping to reveal patterns in auditory perception. The following hypotheses are tested in this study:

H1: the selected *worst* non-individual HRTF provides a significantly worse sound localisation

performance than the *individual* and the *best* non-individual HRTFs;

H2: the selected *best* non-individual HRTF provides similar sound localisation performance to the *individual* HRTF; and

H3: the perceived *overall differences* in spatial audio quality, *tone colour*, *externalisation*, and *naturalness* between the *individual* HRTF and the *worst* one are larger than between the *individual* and the *best* HRTFs.

The work is presented as follows: § 7.2 describes the listening test methodology used to evaluate the procedure and § 7.3 overviews its results. The results and study limitations are discussed in § 7.4.

7.2 Evaluation Methodology

Two listening tests were conducted to assess the validity of the proposed model-based selection of *best* and *worst* non-individual HRTFs. In addition to the static localisation test (Experiment 1, § 7.2.4), which the auditory model aims to predict directly, this HRTF selection was also evaluated using a spatial audio quality assessment (Experiment 2, § 7.2.5) to evaluate the extent to which the differences in localisation performance correspond to the perception of rendering quality. A pilot study design was presented in Daugintis et al. 2023a, improvements to which are presented here.

7.2.1 Test Setup

Both tests were administered to participants in a single session with the localisation test presented first, followed by the spatial audio quality assessment. The participants were allowed to take breaks between the two experiments and between individual blocks of each experiment. The total duration of the session was approximately 1.5 hours.

The listening tests were conducted using a pair of Sennheiser HD 650 headphones and a Meta Quest 3 VR headset. A standalone VR test application was developed in Unity using the Unity wrapper of the 3D Tune-In (3DTI) Toolkit (Cuevas-Rodríguez et al. 2019, Unity package available from 3DTune-In 2023) to render the binaural sounds via the desired HRTFs. For both tests, the VR environment contained a minimal visual environment with a starry night skybox and three coordinate axes to help with orientation (see Figs. 7.2 and 7.3). The subjects used hand-held controllers (which had a virtual laser extension in the VR scene) to navigate the test environments, indicate the perceived direction of a sound source (in case of the localisation

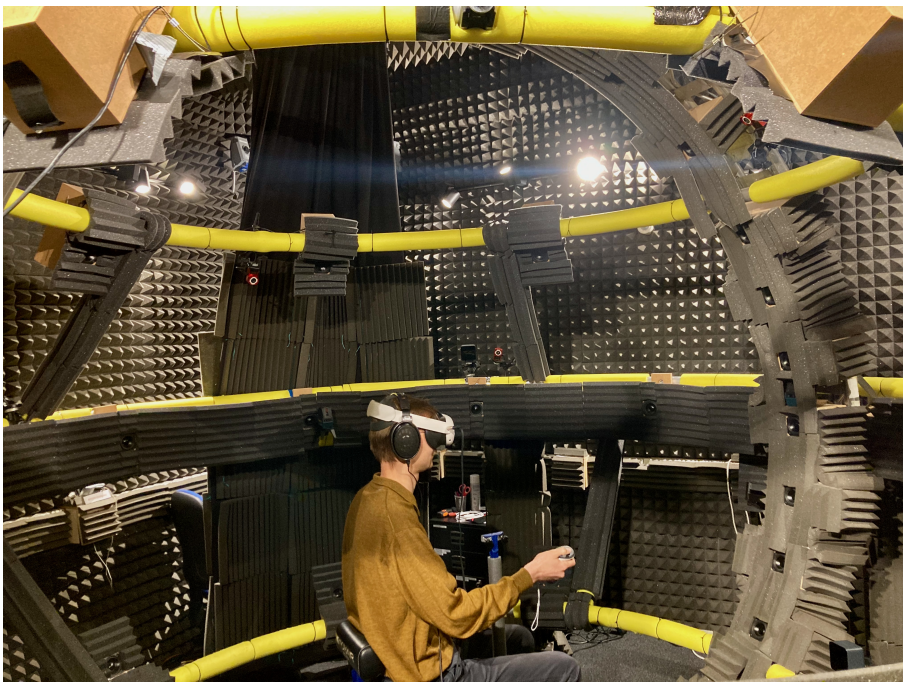


Figure 7.1: Listening test setup. The headphones were removed and the sound stimuli played through the loudspeakers during the Experiment 1 familiarisation session.

test) and rate the differences in the audio quality assessment. The tests were performed in a sound-insulated room that was also used for the HRTF measurements (details are available in Engel et al. 2023), which can be seen in Figure 7.1. The sound level setting in the VR headset was fixed for all participants. The sound pressure level (measured using miniDSP EARS) varied between 65 dBA and 75 dBA, depending on the stimulus and its position.

Instead of headphones, the Experiment 1 familiarisation session (procedure described in § 7.2.4) employed a set of 29 custom-built one-way loudspeakers (built by Wilmslow Audio Ltd, UK, using full-range 3-inch Peerless 830987 drivers) mounted on a dome around the listener (see Figure 7.1). The speakers were powered by Triad TS-PAMP8-100 class-D amplifiers. The audio was played using a Max 8 patcher on a Windows 10 PC, connected to the MOTU audio interfaces using the AVB protocol. The playback was controlled by Open Sound Control (OSC) messages sent from the VR headset to the Max 8 patcher via a local Wi-Fi network.

7.2.2 Participants

Seventeen participants performed listening tests. Their age was 18 to 43 years (median: 26). Four of them identified as female and thirteen as male. One participant was left-handed and one ambidextrous (the participants were allowed to use the VR controller of their preferred hand). Participants also provided information on their familiarity with immersive audio, years of musical training, playing habits of video games, and level of English via a written questionnaire. All but

one participant had at least some experience with immersive audio, and 13 of them had some previous musical training. No participant reported any hearing impairments.

All participants have given their informed consent according to the approval issued by the Imperial College London Science, Engineering, and Technology Research Ethics Committee (SETREC number: 7046527).

7.2.3 HRTFs

Participants who performed the two listening tests had their HRTFs measured as part of the SONICOM dataset (Engel et al. 2023) prior to performing the tests. For each participant, a model-based HRTF similarity metric was used to classify the first 204 SONICOM HRTFs that were available at the time the study was conducted (P0001 to P0204) and select one *best* and one *worst* non-individual HRTF. Three HRTFs (including *individual*) were then used in both listening tests. The free-field compensated versions of the HRTFs (using the minimum-phase inverse free-field response filter) were used. Additionally, all sounds were equalised using individualised headphone equalisation filters, measured as part of the SONICOM dataset (the same headphones used for the measurement were used during the test).

7.2.4 Experiment 1: Static Sound Localisation Test

Static sound localisation test paradigm was chosen for the first experiment to reflect the *barumerli2023* auditory model assumptions of static listening conditions and to directly evaluate the model-based HRTF selection procedure. The design of the sound localisation test was inspired by that conducted by Majdak et al. 2010, and later by Geronazzo et al. 2019 and Geronazzo et al. 2020.

Stimulus

Macé et al. (2012) showed improved localisation performance with repeated noise bursts. Therefore, a short sequence of three Gaussian noise bursts, each 100 ms long and windowed using Hann windows, was chosen as a stimulus for the static localisation test.

Procedure

The test started with the familiarisation session, which provided some minimal procedural training of sound localisation within the same VR environment but without exposing the listener

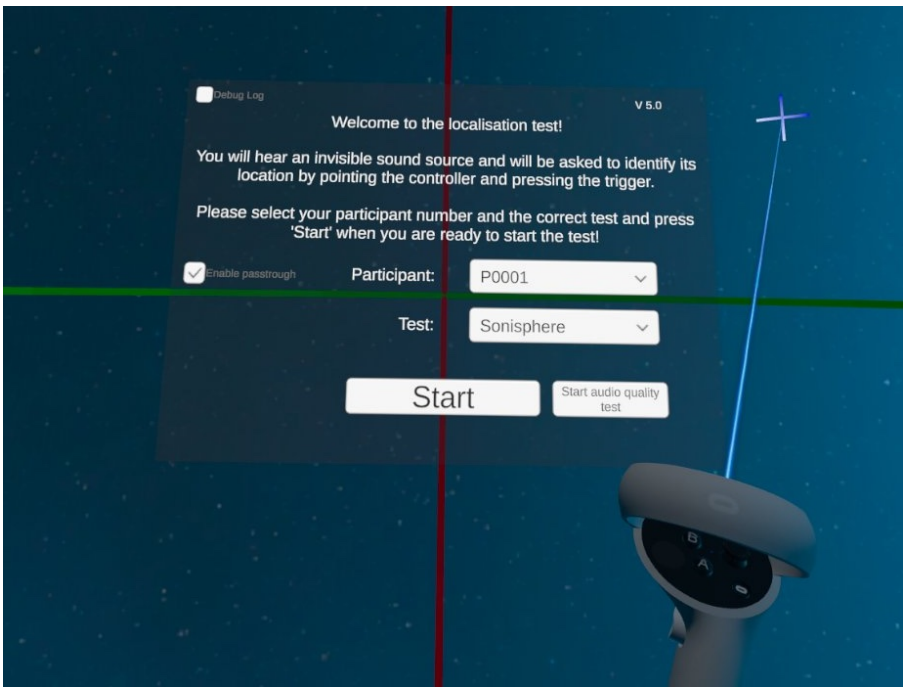


Figure 7.2: The start screen of the VR localization Test.

to the HRTFs tested (the protocol is described below). After the familiarisation, three blocks of the main test were presented. Each block contained 99 trials, which corresponded to a randomly interleaved set of 33 distinct source positions evenly spaced around the listener (down to −30° elevation), rendered using each of the three HRTFs. Note that although the HRTF selection was based on predicted localisation performance within a small region in front of the listener, the test evaluated performance across the entire virtual sphere to assess how well the selection generalises globally. In total, the complete listening test contained three repetitions of each direction and HRTF condition pairings. Participants were allowed to take a break after the familiarisation session and between the blocks.

Before each trial, subjects were aided by a virtual reticle and a target to position themselves in the middle of the virtual sphere and orient their gaze towards the front ($\theta = 0°$ and $\phi = 0°$). They were then asked to remain still while the sound, rendered in a specific direction using one of the HRTFs tested, was played (if the participant rotated their head by more than 2.5° away from the front position during that time, the sound would stop and the trial would restart). They were then asked to indicate the perceived direction of the sound source using the virtual laser, attached to the hand-held controller (visible in Figure 7.2).

Familiarisation Session

Previous studies acknowledged the need to familiarise subjects with the test environment without providing perceptual training on the test conditions, which would hinder experimental results (Majdak et al. 2010). However, a simple familiarisation routine, implemented in the pilot study Daugintis et al. 2023a, proved to be insufficient in increasing the reliability of subjects' localisation performance. Instead, a more extensive training routine was devised, aimed at providing some training on general sound localisation skills with participants' own ears and familiarisation with the test protocol without exposing them to the specific binaural rendering conditions.

The familiarisation session was administered using the same VR environment, but the headphones were removed and the sound stimuli (the same 3×100 ms Gaussian noise train) were played through the surrounding speakers (shown in Figure 7.1). Subjects received visual feedback after each response: If the response was within 5° of the target, a green sphere would appear at that position, and the response was registered as correct. However, for responses that were further away, a different colour sphere appeared in the response position (yellow, orange, and finally red, if the response was further away than 35° from the target). The sphere was also accompanied by an arrow pointing towards the direction of the target; the participant was then given a maximum of two more attempts to point to the correct position of the target before moving to the next position. The training consisted of 87 trials, each loudspeaker position being repeated three times in a randomised order. The position of the sound source was visible from the first attempt for the first 27 trials. For the rest of the session, the sound source was made visible only for the third attempt (i.e. if the participant's responses were wrong twice).

Data Analysis

The data of the static sound localisation test were analysed based on the metrics described in § 2.3.1, namely, mean GCE, confusion rate classification, PAE, and LE. For each metric, the normality of the error distributions associated with each HRTF condition (*individual*, *best*, and *worst*) across the participants was assessed using the Shapiro-Wilk test. If the test did not reject the normality hypothesis for all three distributions ($p > 0.05$), repeated measures analysis of variance (RM ANOVA) was used to assess the main effect of the HRTF condition on the localisation performance metric. Greenhouse-Geisser correction was used on p-values if Mauchly test for sphericity indicated a violation ($p < 0.05$). If RM ANOVA returned a significant main effect ($p < 0.05$), post hoc paired t-tests were used to investigate differences between conditions.

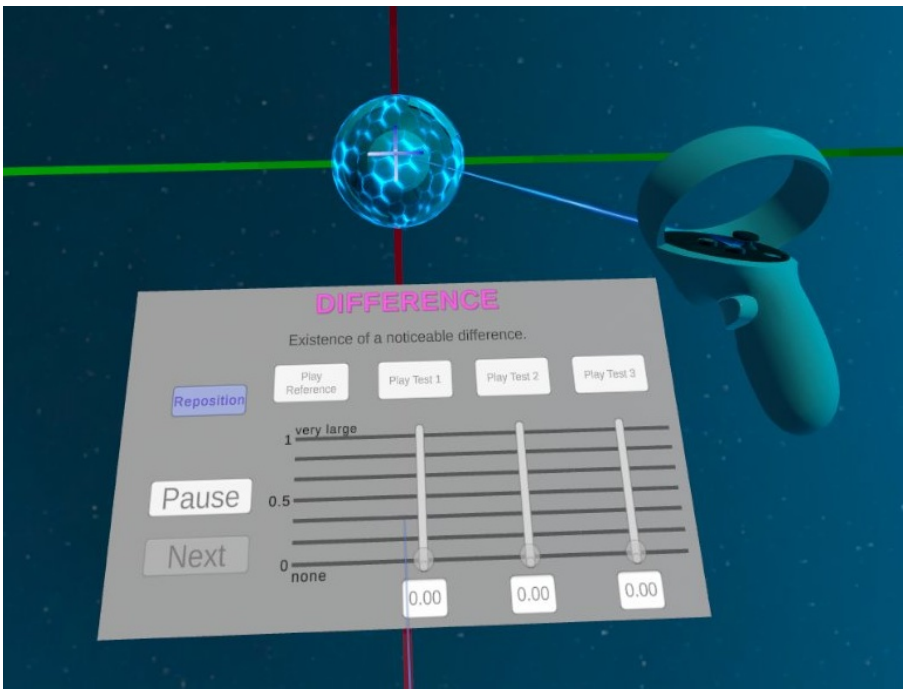


Figure 7.3: The VR environment for spatial audio quality assessment. The sphere represents a visible sound source that can be moved using the VR controller.

For scenarios in which at least one distribution was deemed not normal, the Friedman test was used to assess the main effect of the HRTF condition, followed by the Wilcoxon signed rank post hoc test. Bonferroni correction was applied to the p-values of both types of post hoc tests to control the family-wise error rate. The effect of the HRTF condition was deemed significant if the corrected p-value was below 0.05. Statistical analysis was performed using a *Pingouin* package (Vallat 2018) in Python 3.12.7.

7.2.5 Experiment 2: Spatial Audio Quality Test

A spatial audio quality test was inspired by the studies presented by Brinkmann et al. (2019a,b) and used the design principles and attributes of the Spatial Audio Quality Inventory (SAQI) (Lindau et al. 2014).

Stimuli

The experiment was based on dynamic listening conditions in which the sound source was visible and the listeners could interact with it by either moving it around (without changing the distance) or rotating themselves. Four sound clips were used: a series of 1 s pink noise bursts, followed by 1 s silence (similar to the one used by Brinkmann et al. 2019a), an anechoic English male speech (from Bang & Olufsen: Music for Archimedes, 1992), an anechoic castanet rhythm,

and an anechoic excerpt of a guitar solo (Capriccio Arabe by Francisco Tárrega, performed by Ingolf Olsen from Bang & Olufsen: Music for Archimedes, 1992).

Procedure

Each trial consisted of a comparison between a reference sound and three test sounds. The reference sound was always rendered via *individual* HRTFs, while the three tests used *individual* (a control condition), *best* or *worst* non-individual HRTF in a randomised order. The participants were presented with a graphical user interface (see Figure 7.3) where they could play and pause the sound and switch between the four conditions. They were then asked to use the sliders provided to rate the difference between each test and the reference sound. Following the SAQI protocol (Lindau et al. 2014), they were initially asked about the *overall difference* between the test sound sources and the reference and asked to rate it between 0 (none) and 1 (very large). Then more detailed questions on *tone colour bright-dark* (accompanied by a bipolar scale of -1 (darker) to 1 (brighter)), *externalisation* (from -1 (more internalised) to 1 (more externalised)), *naturalness* (from -1 (lower) to 1 (higher)) were asked in a randomised order. The description of each attribute, as detailed by Lindau et al. 2014, was also presented on the screen. Finally, they were asked to rate any other noticeable difference that participants could also describe verbally to the examiner. Participants were encouraged to explore the sound scene dynamically before rating the differences. There were a total of four sessions (five questions each), corresponding to the different audio samples. Their order was randomly selected for each participant. The test was preceded by a set of instructions to familiarise the subjects with the setup.

Familiarisation Session

The qualitative test followed the localisation test that was conducted in the same visual VR environment, so the subjects were already familiar with the general test setup. Before starting the test, the examiner introduced the procedure verbally. Then a set of instruction panels was presented within the VR environment highlighting the main controls of the test (visible in Figure 7.3) to subjects: the buttons to play and pause each of the conditions, the rating slider, and the ability to move the audio source by selecting it with the controller. Subjects were allowed to try each control without registering the selection during the instructions. They could also ask further questions to clarify the meaning of each attribute during the test. Importantly, they were not informed that the test included a control condition identical to the reference.

Data Analysis

The hidden control condition (*individual* HRTF) in the qualitative test was used to post-screen the responses and ensure the reliability of the results. Trials in which participants rated the *individual* HRTF test conditions as more different (over 0.1 point tolerance) from the reference than the *best* or *worst* HRTF conditions were flagged as outliers and not considered in further result analysis (for *tone colour*, *naturalness*, and *externalisation* ratings, absolute rating values were compared).

Further analysis was performed on the ratings of the *best* and *worst* HRTF conditions using linear mixed effect models via *lme4* package (Bates et al. 2015) in R 4.4.3 (R Core Team 2025). Initially, the most general linear model was fitted for each attribute to account for the maximum viable set of effects: `Rating ~ HRTF + Stimulus + (HRTF + Stimulus | Subject)`, where the formula is of a form `Response ~ Fixed Effects + (Random Effects | Grouping Factor)` and `HRTF` and `Stimulus` are ordered categorical variables for the HRTF condition (*best* and *worst* only, since the *individual* HRTF condition was only used for the verification of the trial) and the type of stimulus.

The significance of each effect was assessed using a step-down approach, implemented in *lmerTest* package (Kuznetsova et al. 2017). Due to the small sample size of the test, the Kenward-Roger method (Kenward and Roger 1997) via *pbkrtest* package (Halekoh and Højsgaard 2014) was used to compare the models and estimate the significance of fixed effects. The method starts from the full model of a form `Rating ~ HRTF + Stimulus + (HRTF + Stimulus | Subject)` each time (`abs(Rating)` in the case of *tone colour*). The method first assesses each random effect by removing it from the model and comparing the full and the reduced models using the likelihood ratio test. The effects are retained if the p-value of the likelihood ratio statistic (which is asymptotically distributed as χ^2) is below 0.1 (`(1 | Subject)` random intercepts were kept to retain the mixed-model structure, even if their contribution was not significant). It then assesses each fixed effect by calculating its F statistics (using the Kenward-Roger approximation) and removes the fixed effects that result in a non-significant change ($p > 0.05$) when the effect is removed from the model (Kuznetsova et al. 2017). For each attrbute, the procedure is summarised in tables, presented in the Result section. Each row of the table presents the statistics of the model with the corresponding parameter removed. For the random effects, N_{par} represents the number of model parameters, LL—log-likelihood, AIC—Akaike information criterion, LRT—likelihood ratio test statistic, df—degrees of freedom, and p $(> \chi^2)$—its associated p-value. For the fixed effects, SS is the sum of squares, MS—mean square, df_{num}—numerator

degrees of freedom, df_{den}—denominator degrees of freedom, F—the value of the F-statistic, and p (> F)—its associated p-value. p-values below elimination thresholds are in bold.

The results of the final model are reported, following the guidance of Meteyard and Davies 2020. Reported information on the fixed effects includes the estimate of the rating for each condition (in a 0 – 1 or -1 – 1 SAQI rating scale, depending on the attribute), standard error (SE) and a bootstrapped 95 % confidence interval (CI) of the estimate, degrees of freedom (df), t-statistic, and the associated p-value. The variance and standard deviation (SD) of the random effects (and residuals) and the correlation between them are reported as well. Finally, marginal and conditional r-squared values (according to Nakagawa et al. 2017), and intraclass correlation (ICC) are reported to assess the variance components of the model.

7.3 Results

7.3.1 Experiment 1: Static Sound Localisation

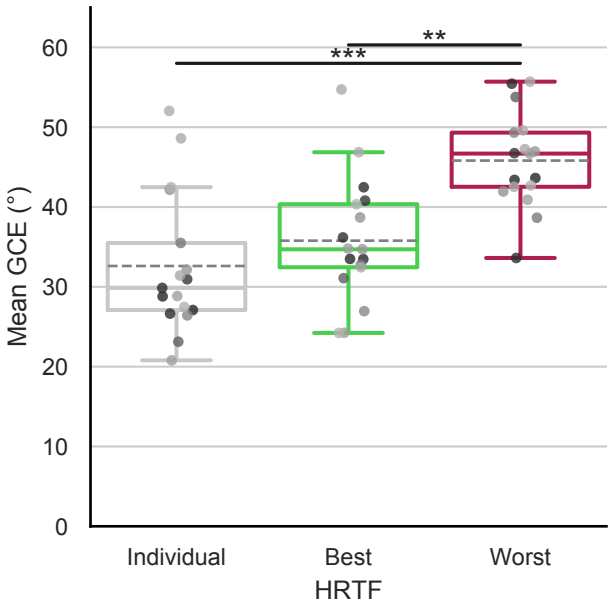


Figure 7.4: Distributions of mean great-circle error (GCE) across all trials per subject for three HRTF conditions. Shades of grey represent individual subject data. The boxes represent the interquartile range (IQR) with a line indicating the median, while the whiskers extend to the most extreme data points within 1.5×IQR. Dashed lines show the distribution means. Bars and asterisks indicate statistically significant differences (**: $p < 0.01$, ***: $p < 0.001$).

Figure 7.4 shows the distribution of mean GCE across all trials for each subject, grouped by HRTF condition. The distribution for *individual* HRTFs is non-normal ($p = 0.045$). The Friedman test revealed a significant effect of the HRTF condition on mean GCE ($\chi^2(2) = 16.35$, $p < 0.001$).

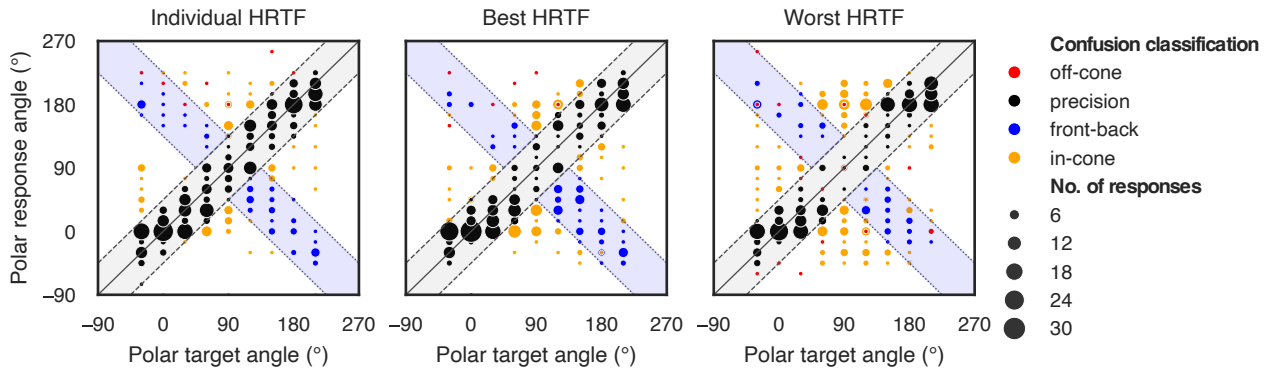


Figure 7.5: Polar angle responses for target positions along the median plane ($\theta = 0°$) across all participants, grouped into 15° intervals. The responses are categorised, according to Algorithm 2.1. Precision and front-back confusion areas are shaded in grey and blue, accordingly (some responses within these areas might be classified differently due to large lateral angle errors).

Post hoc tests showed significant differences between the *individual* and the *worst* HRTFs ($W = 5$, $p < 0.001$), and between the *best* and the *worst* HRTFs ($W = 14$, $p = 0.005$), but not between the *individual* and the *best* conditions ($W = 32$, $p = 0.1$). Based on the distribution medians, the *worst* HRTFs increased the mean GCE by approximately 17° compared to the *individual* HRTFs, while the error with the *best* HRTFs was only about 5° higher.

Focusing on the polar axis, Figure 7.5 shows responses for target positions on the median plane only, aggregated across participants. It indicates an increasing spread of responses away from the precision region with non-individual HRTFs. The confusion rates (according to Algorithm 2.1) across all target positions around the sphere for each participant are plotted in Figure 7.6. The effect of the HRTF condition on each type of confusion was analysed using

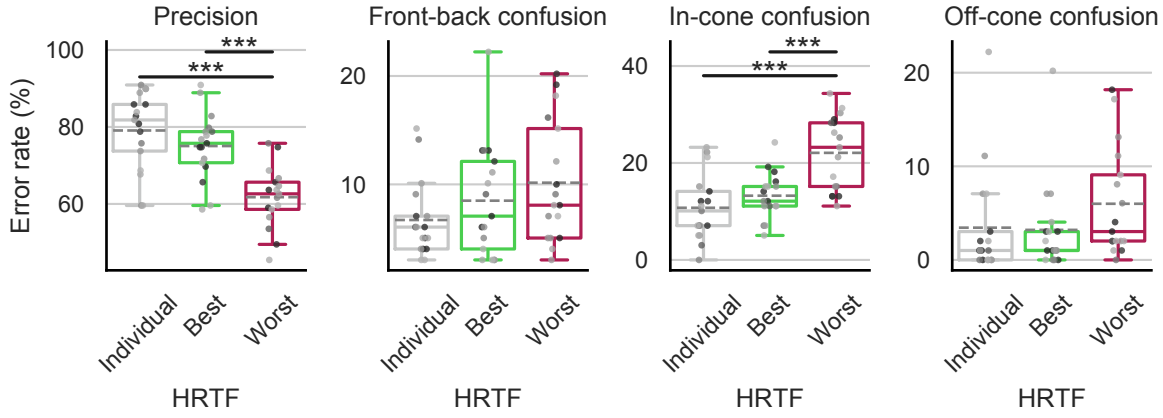


Figure 7.6: Distributions of confusion rates per participant across all the target positions. Shades of grey represent individual subject data. The boxes represent the IQR with a line indicating the median, while the whiskers extend to the most extreme data points within 1.5×IQR. Dashed lines show the distribution means. Bars and asterisks indicate statistically significant differences (***: $p < 0.001$).

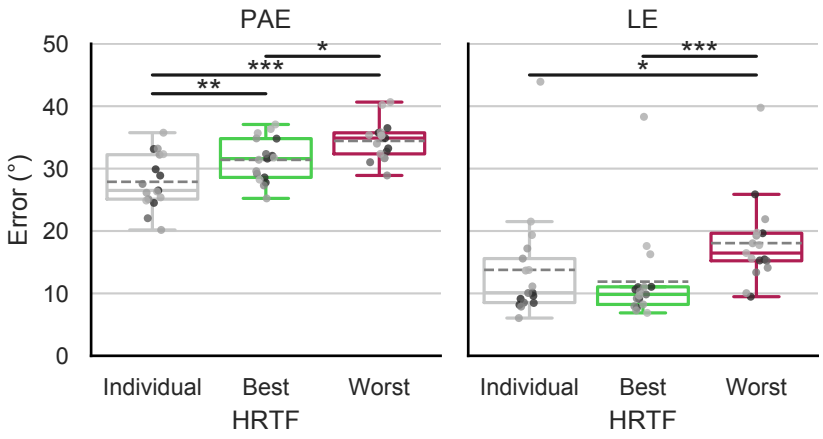


Figure 7.7: Distributions of polar angle error (PAE) and lateral angle error (LE) across subjects for each HRTF condition. Shades of grey represent individual subject data. The boxes represent the IQR with a line indicating the median, while the whiskers extend to the most extreme data points within 1.5×IQR. Dashed lines show the distribution means. Bars and asterisks indicate statistically significant differences (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

univariate statistics, but the p-values were corrected using the Bonferroni method to account for the four tests. RM ANOVA showed a main effect of the HRTF condition on the precision rate ($F(2, 32) = 25.33$, $p < 0.001$). Post hoc tests revealed significant differences between the *individual* and the *worst* HRTF condition ($t(16) = 5.84$, $p < 0.001$) as well as the *best* and the *worst* HRTF condition ($t(16) = 4.90$, $p < 0.001$). However, the difference between the *individual* and the *best* HRTF conditions was not significant ($t(16) = 2.23$, $p = 0.12$). A main effect of the HRTF condition was also found on the in-cone confusion rate ($F(2, 32) = 23.10$, $p < 0.001$). Again, post hoc tests revealed significant differences between the *individual* vs. the *worst* HRTF condition ($t(16) = -5.68$, $p < 0.001$) and the *best* vs. the *worst* HRTF condition ($t(16) = -4.77$, $p < 0.001$) but not between the *individual* vs. the *best* HRTF condition ($t(16) = -1.86$, $p = 0.25$). No significant differences between the HRTF conditions were found in the front-back confusion rates (Friedman test results: $\chi^2(2) = 2.76$, $p = 1$) and the off-cone confusion rates ($\chi^2(2) = 7.18$, $p = 0.11$). All in all, the *worst* HRTFs reduced precision by an average of 17 percentage points (pp) and increased in-cone confusion rates by about 11 pp while the *best* HRTFs retained precision (within 4 pp) and in-cone confusion (within 2 pp) rates comparable to the *individual* HRTFs.

A more detailed analysis on the local errors was performed to investigate the differences in localisation error when the response was in the same quadrant as the target. The left panel of Figure 7.7 presents the PAE across the participants for each HRTF condition. A main effect of HRTF condition on PAE was found ($F(2, 32) = 26.19$, $p < 0.001$), and post hoc tests showed significant differences between all pairs: *individual* vs. *worst* ($t(16) = -6.38$, $p < 0.001$), *individual* vs. *best* ($t(16) = -4.45$, $p = 0.001$), and *best* vs. *worst* ($t(16) = -3.43$, $p = 0.010$). On average,

the *worst* HRTF increased the PAE by about 7° compared to the *individual* HRTF, while the *best* HRTF increased it by about 4°.

The right panel of Figure 7.7 shows the LE, which was generally lower than the PAE across all conditions. A Friedman test indicated a main effect of the HRTF condition on LE ($\chi^2(2) = 13.76$, $p = 0.001$). Post hoc tests revealed significant differences between the *best* and the *worst* ($W = 3$, $p < 0.001$) as well as the *individual* and the *worst* ($W = 18$, $p = 0.012$) but not between the *individual* and the *best* ($W = 35$, $p = 0.15$) HRTF conditions. While median LE across the participants was approximately 10° with both the *individual* and the *best* HRTFs, it increased to about 16° with the *worst* HRTF. Overall, the *worst* HRTF yielded the poorest localisation precision on both coordinate axes, while the *best* HRTF showed better performance, but still resulted in higher errors along the polar axis compared to the *individual* HRTF.

7.3.2 Experiment 2: Spatial Audio Quality

Figure 7.8 presents results from the spatial audio quality assessment, which aimed to determine whether predicted differences in localisation performance are also perceived qualitatively. It shows difference ratings for the two non-individual HRTFs, each compared to the *individual* HRTF, across the selected SAQI attributes. Ratings for the hidden reference (*individual* HRTF) condition are also included and were used to identify outliers in the data.

The number of outliers identified varies across the audio sample and the test attribute conditions. The ratings for the difference in *externalisation* were the least reliable, with 48 ratings flagged as outliers. Furthermore, ratings for the guitar sample contain the highest number of outliers: 78 in total across all attributes. For example, six participants ranked *individual* HRTFs as more different than non-individual ones in the case of *overall difference* and seven in the *tone colour* ratings, which constitutes more than a third of all participants. Due to the high number of outliers, guitar ratings were not considered in the linear mixed effects regression analysis to avoid highly unbalanced observation sets.

The ratings for each attribute were analysed using mixed-effects models, as discussed in § 7.2.5. For the *overall difference* attribute (Figure 7.8A-C), a mixed-effects model `Rating ~ HRTF + (HRTF | Subject)` was found to best fit the data (details on backward effect elimination are provided in Table 7.1), suggesting a minimal effect of the stimulus type on the subjective rating. Table 7.2 presents the results of the final reduced model. While the *best* HRTFs were rated as significantly different from the *individual* HRTF (intercept $\beta = 0.38$, $p < 0.001$), the ratings for the *worst* HRTFs were significantly larger ($\beta = 0.23$, $p = 0.002$). Model fit parameters

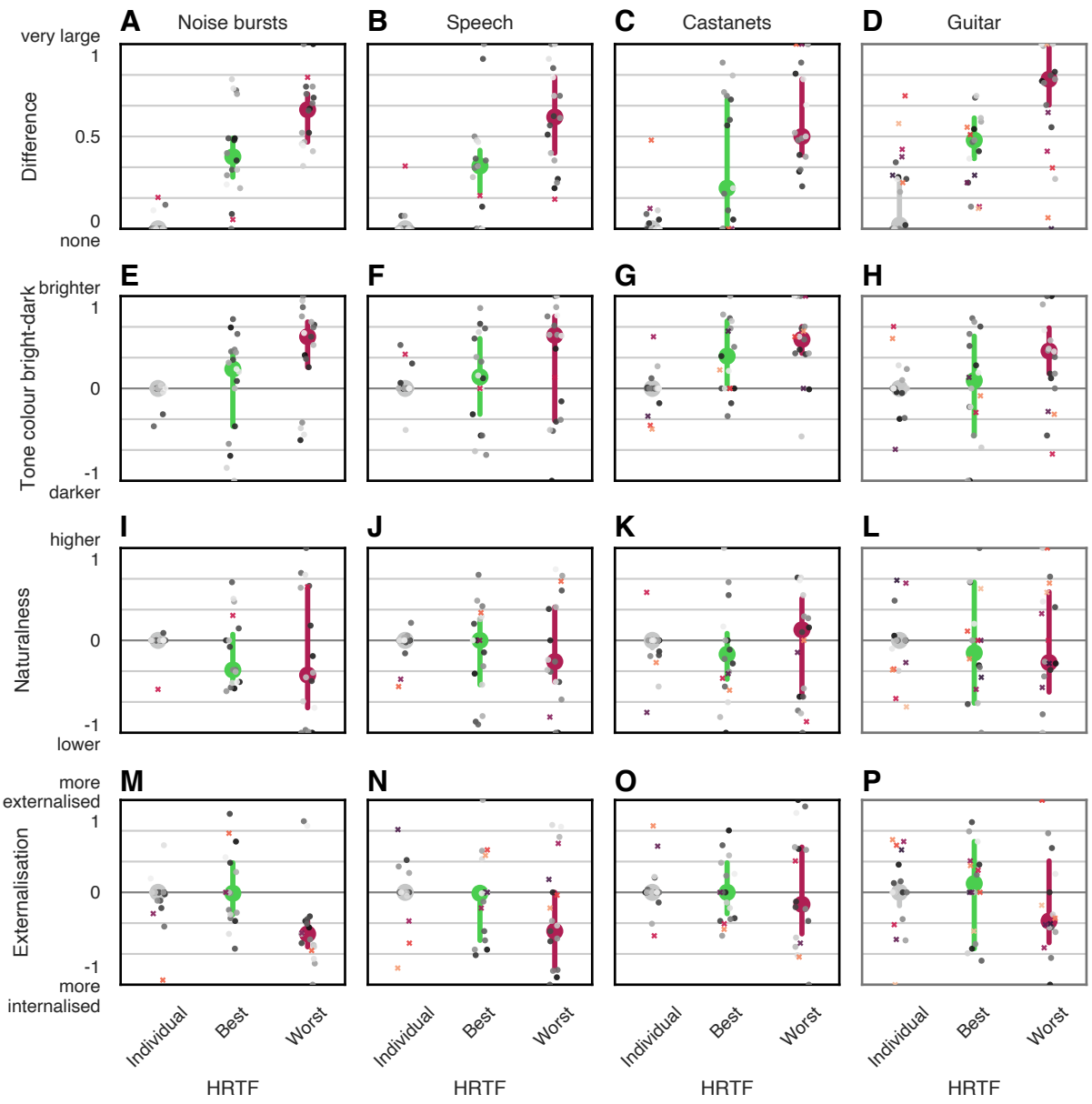


Figure 7.8: Spatial audio quality ratings, in comparison to the reference *individual* HRTF. Individual responses are represented by shades of grey. x markers in shades of red show outliers excluded from the analysis because the hidden *individual* HRTF condition (left), identical to reference, was rated as more different than other conditions. Markers with bars indicate medians and bootstrapped 95 % confidence intervals across the responses (excluding the outliers). Ratings for the guitar sample were excluded from data analysis due to high number of outliers.

indicate some subject-level variability. This is evident for the castanets (Figure 7.8C), where ratings for the *best* HRTF are bimodal: some participants perceived little to no difference, while others reported large deviations from the *individual* HRTF. Overall, the analysis confirms that the *worst* HRTF was associated with greater perceptual deviation from the *individual* HRTF than the *best* HRTF.

The rating patterns for other SAQI attributes were also analysed to better understand the nature of perceived differences between HRTF conditions. The distributions of *tone colour*

Table 7.1: Backward elimination of random and fixed effects for *overall difference* ratings.

Random Effect	Eliminated	N_{par}	LL	AIC	df	LRT	$p(>\chi^2)$
[full model]		15	-11.32	52.64			
Stimulus (slope)	yes	8	-12.90	41.81	7	3.17	0.868
HRTF (slope)	no	6	-15.34	42.68	2	4.88	**0.087**
Fixed effect	**Eliminated**	**SS**	**MS**	df_{num}	df_{den}	**F**	**p (> F)**
Stimulus	yes	0.05	0.02	2	61.00	0.41	0.662
HRTF	no	0.73	0.73	1	15.84	13.34	**0.002**

Final model: `Rating ~ HRTF + (HRTF | Subject)`

Table 7.2: Rating estimates and statistics of the final model for *overall difference*.

Fixed Effects						
	Estimate	SE	95 % CI	df	t	p
Intercept (*best* HRTF)	0.38	0.05	0.27 – 0.49	15.91	6.96	<0.001
worst HRTF	0.23	0.06	0.11 – 0.37	15.84	3.65	0.002
Random Effects						
Group	Parameter		Variance	SD	Correlation	
Subject	Intercept (*best* HRTF)		0.03	0.17		
Subject	*worst* HRTF		0.03	0.18	-0.94	
Residual			0.05	0.23		
Model Fit						
Marginal R^2	Conditional R^2		ICC			
0.16	0.36		0.36			

ratings (Figure 7.8E–H) were bimodal, reflecting that some participants perceived non-individual HRTFs as darker, others as brighter. To account for this, the absolute rating values were used, and the model `abs(Rating) ~ HRTF + (1 | Subject)` was fit using a backward elimination procedure (Table 7.3). The final model, presented in Table 7.4, showed no significant effect of stimulus type and little variability across subjects. Although the random intercept was retained for consistency with other SAQI models, its contribution was negligible. While overall model fit was poorer compared to the model for *overall difference*, it revealed a similar pattern: the *best* HRTFs were rated as significantly different in *tone colour* from the *individual* HRTFs (intercept $\beta = 0.44$, $p < 0.001$), with the *worst* HRTFs rated as even more different ($\beta = 0.16$, $p = 0.004$).

Difference ratings for the *naturalness* and *externalisation* attributes were generally centred around zero, with the exception of lower *externalisation* ratings for noise bursts with the *worst* HRTF (Figure 7.8M). Mixed-effects models for both attributes showed no significant effects of stimulus type or HRTF condition. For *naturalness*, only the random intercept was significant (Table 7.5), indicating that subjects applied individual but consistent rating offsets. The intraclass

Table 7.3: Backward elimination of random and fixed effects for absolute *tone colour* difference ratings.

Random Effect	Eliminated	N_{par}	LL	AIC	df	LRT	$p(>\chi^2)$
[full model]		15	-13.32	56.65			
Stimulus (slope)	yes	8	-16.93	49.86	7	7.21	0.408
HRTF (slope)	yes	6	-17.31	46.61	2	0.75	0.686
Intercept	no (fixed)	5	-18.01	46.01	1	1.40	0.237
Fixed effect	**Eliminated**	**SS**	**MS**	df_{num}	df_{den}	**F**	**p (> F)**
Stimulus	yes	0.02	0.01	2	76.47	0.12	0.883
HRTF	no	0.60	0.60	1	74.19	8.86	**0.004**

Final model: `abs(Rating) ~ HRTF + (1 | Subject)`

Table 7.4: Absolute rating estimates and statistics of the final model for *tone colour*.

Fixed Effects						
	Estimate	SE	95 % CI	df	t	p
Intercept (*best* HRTF)	0.44	0.04	0.35 – 0.52	38.40	10.01	<0.001
worst HRTF	0.16	0.05	0.06 – 0.27	74.19	2.98	0.004
Random Effects						
Group	Parameter		Variance	SD		
Subject	Intercept (*best* HRTF)		0.01	0.09		
Residual			0.07	0.27		
Model Fit						
Marginal R^2	Conditional R^2		ICC			
0.08	0.17		0.10			

correlation coefficient (ICC) from a simplified model (`Rating ~ (1 | Subject)`) was 0.35, suggesting moderate intra-subject rating consistency. In contrast, no fixed or random effects were significant for *externalisation* (Table 7.6), and the model `Rating ~ (1 | Subject)` yielded a low ICC of 0.08, indicating little to no subject-specific pattern. Overall, these results suggest that HRTF fit had minimal impact on perceived *naturalness* and *externalisation* in this listening scenario.

7.3.3 Cross-Experiment Analysis

To understand the relationship between the two experiments, the data were compared using statistical correlation methods. Mean GCE (Figure 7.4) was chosen as the most general metric to represent each individual's overall sound localisation performance, because unlike the local error metrics, it incorporates all tested directions. The intra-subject differences in mean GCEs (ΔGCE) between the *best* and *individual* as well as the *worst* versus *individual* HRTF conditions

Table 7.5: Backward elimination of random and fixed effects for *naturalness* difference ratings.

Random Effect	Eliminated	N_{par}	LL	AIC	df	LRT	$p(>\chi^2)$
[full model]		15	-67.16	164.33			
Stimulus (slope)	yes	8	-69.92	155.84	7	5.51	0.598
HRTF (slope)	yes	6	-72.07	156.15	2	4.31	0.116
Intercept	no	5	-78.57	167.15	1	13.00	**0.001**
Fixed effect	**Eliminated**	**SS**	**MS**	df_{num}	df_{den}	**F**	**p (> F)**
HRTF	yes	0.00	0.00	1	70.14	0.00	0.949
Stimulus	yes	0.04	0.02	2	74.22	0.09	0.913

Final model: `Rating ~ (1 | Subject)`

Table 7.6: Backward elimination of random and fixed effects for *externalisation* difference ratings.

Random Effect	Eliminated	N_{par}	LL	AIC	df	LRT	$p(>\chi^2)$
[full model]		15	-59.72	149.44			
HRTF (slope)	yes	11	-61.11	144.21	4	2.77	0.597
Stimulus (slope)	yes	6	-65.04	142.08	5	7.87	0.164
Intercept	no (fixed)	5	-65.50	140.99	1	0.91	0.340
Fixed effect	**Eliminated**	**SS**	**MS**	df_{num}	df_{den}	**F**	**p (> F)**
Stimulus	yes	0.89	0.45	2	72.25	1.89	0.158
HRTF	yes	0.71	0.71	1	66.51	2.97	**0.09**

Final model: `Rating ~ (1 | Subject)`

were calculated to produce a metric comparable to the SAQI difference ratings. ΔGCE was compared to the medians of individual ratings across three stimuli (noise bursts, speech, and castanets) for the *best* and the *worst* HRTF conditions (ratings associated with the guitar sample were not considered due to unreliability of the data, as discussed in the previous section).

The median scores for *overall difference* (Figure 7.8A-C) and *tone colour* difference (Figure 7.8E-G) were plotted against ΔGCE in Figs. 7.9 and 7.10, respectively. A repeated measures correlation was used to assess the relationship, accounting for paired observations within subjects by allowing varying intercepts but a shared slope across individuals (Bakdash and Marusich 2017). The correlation coefficient (r_{rm}) and bootstrapped 95 % confidence intervals reflect the consistency of this relationship. Each figure includes a regression line showing the predicted marginal effect of ΔGCE on the median rating, based on the estimated fixed-effect slope and model intercept.

The r_{rm} indicates a strong, significant relationship between differences in sound localisation performance and the qualitative *overall difference* rating ($r_{rm} = 0.63$, 95 % CI $[0.21, 0.90]$, $p = 0.005$). The correlation with median *tone colour* difference ratings was moderate but still

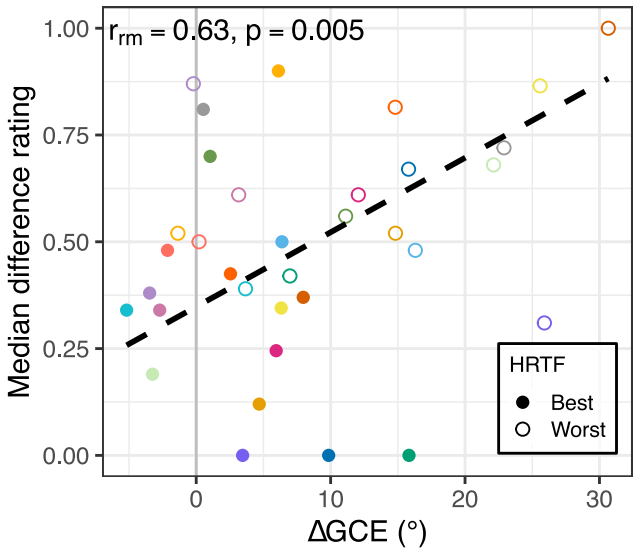


Figure 7.9: Median *overall difference* rating across different stimuli against the difference in mean great circle error compared to *individual* HRTF conditions. Colours represent individual data. Black regression line shows predicted marginal effect, obtained from repeated measures correlation analysis.

significant ($r_{rm} = 0.60$, 95 % CI $[-0.02, 0.87]$, $p = 0.009$), though the wider confidence interval suggests greater variability across subjects. Because the *tone colour* rating used here is signed, the correlation implies that HRTFs with localisation performance closer to the *individual* HRTF were more often perceived as darker, while those with greater localisation errors tended to be rated as brighter.

To investigate trends in *tone colour* ratings with the *best* and the *worst* HRTFs further, the ratings were compared against the difference in the spectral centroids of the associated HRTFs. Spectral centroids, which are commonly associated with perceived brightness (Grey and Gordon 1978), were calculated according to the following formula:

$$f_{centroid} = \frac{1}{2} \sum_{i \in \{L,R\}} \frac{\sum_{k=0}^{K-1} f[k]\,|H_i[0,0,k]|}{\sum_{k=0}^{K-1} |H_i[0,0,k]|}. \tag{7.1}$$

Here, k represents a frequency bin index, K is the total number of bins up to the Nyquist frequency, $f[k]$ is the frequency at bin k, and $|H_i[0,0,k]|$ is the magnitude of the left ($i = L$) or the right ($i = R$) ear HRTF, for the front position ($\theta = 0°, \phi = 0°$), corresponding to the frequency $f[k]$. The front position was chosen because it was the starting position of the sound source for each trial in the experiment. The average spectral centroid between the left and right ear HRTF was used to account for any asymmetries between the two ears.

The top plot of Figure 7.11 shows the median *tone colour* difference ratings plotted against the

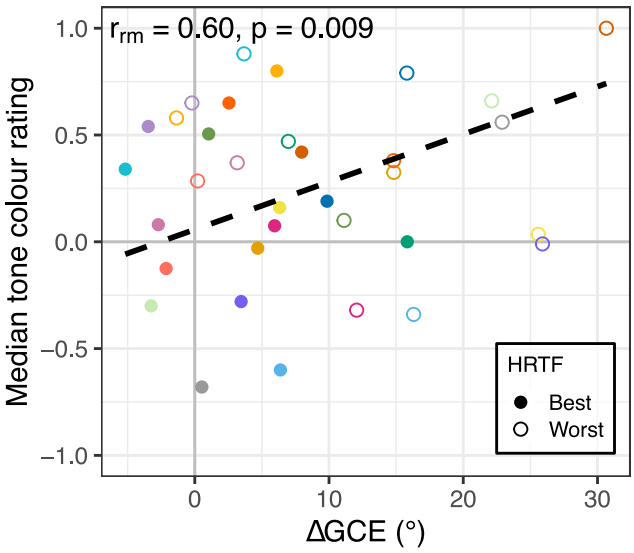


Figure 7.10: Median *tone colour* rating across different stimuli against the difference in mean great circle error compared to *individual* HRTF conditions. Colours represent individual data. Black regression line shows predicted marginal effect, obtained from repeated measures correlation analysis.

difference in spectral centroids between the test and the reference (*individual*) HRTF conditions. Repeated measures correlation revealed a moderate but significant relationship between the two metrics ($r_{rm} = 0.53, 95\,\%\,\mathrm{CI}\,[-0.07, 0.89], p = 0.023$); however, the spread of CI suggests large variability of this association between subjects. The bottom of the figure also shows the distributions of the spectral centroid differences for *best* and *worst* HRTF conditions. The paired t-test (using the reference HRTF, associated with each subject, as a pairing variable) revealed a significant difference between the two groups ($t(16) = -3.37, p = 0.004$), suggesting that the *best* HRTFs, picked by the selection procedure, tended to have lower spectral centroids compared to the *worst* HRTFs for each participant.

Comparison of localisation performance with the other two qualitative ratings—*naturalness* and *externalisation*—revealed no relationship between the data (*naturalness* vs. ΔGCE: $r_{rm} = -0.21, 95\,\%\,\mathrm{CI}\,[-0.64, 0.52], p = 0.41$, *externalisation* vs. ΔGCE: $r_{rm} = -0.26, 95\,\%\,\mathrm{CI}\,[-0.76, 0.30], p = 0.29$). Similarly to the data analysis in § 7.3.2, these results suggest little effect of the HRTF choice on the higher-level perceptual qualities of the audio rendering in the controlled test environment, used for these experiments.

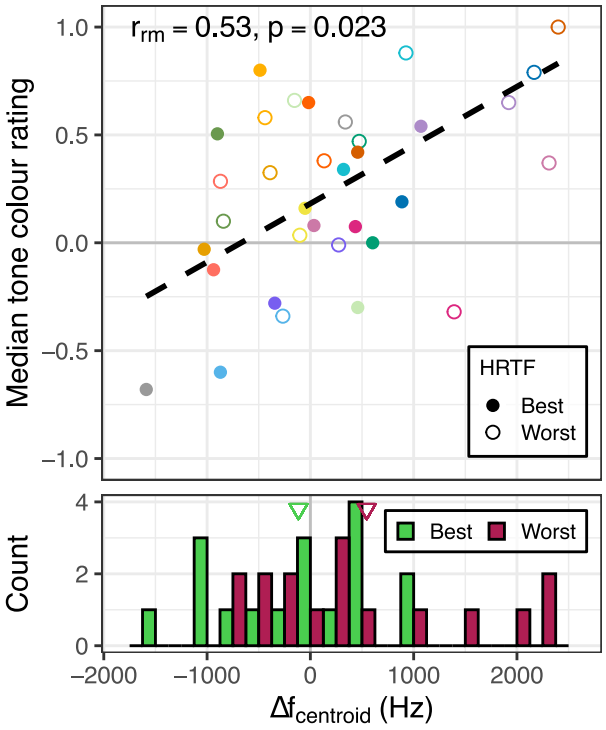


Figure 7.11: Top: Median *tone colour* rating across different stimuli against the difference in spectral centroid (in comparison to the *individual* HRTF) at frontal position. Colours represent individual data. Black regression line shows predicted marginal effect, obtained from repeated measures correlation analysis. Bottom: distribution of spectral centroid differences per HRTF condition. Triangles represent distribution means.

7.4 Discussion

The static sound localisation test demonstrated clear differences in localisation performance between the *individual* and the two selected non-individual HRTFs. This finding was also reflected in participants' subjective ratings during the dynamic spatial audio quality assessment.

7.4.1 Sound Localisation: Best Close to Individual HRTF

The significant increase in mean GCE, PAE, in-cone confusion rate, and the decreased precision rate when localising sound sources with the *worst* HRTFs compared to the *individual* supports the hypothesis **H1**. On the other hand, the *best* HRTFs provide confusion rates and mean GCE similar to the *individual* HRTFs, but are associated with slightly, yet significantly, higher PAE. Therefore, the results partially support **H2**: while the *best* HRTFs provide comparable global localisation performance, they still reduce localisation precision on a local scale. This suggests that although *individual* HRTFs may be necessary for applications requiring high localisation precision, a well-matched non-individual HRTF may be adequate for most casual listening scenarios.

It is noteworthy that no significant differences were detected among the HRTF conditions regarding front-back and off-cone confusion rates. These rates were generally lower than the in-cone confusion rates, and the distributions across participants were predominantly non-normal, exhibiting a skew towards zero, consequently diminishing the statistical power of the analysis. According to Figure 7.5, participants experienced a comparable degree of front-back confusion on the median plane with both the *best* and *individual* HRTFs, albeit with increased confusion at rear target positions with the *best* HRTF. Conversely, responses with the *worst* HRTF displayed greater scattering across polar angles, with fewer of them classified as strictly front-back confusions. This implies that the interpretation of results is contingent upon the category thresholds defined in Algorithm 2.1. Nonetheless, although non-significant, the trends in front-back and especially off-cone confusion rates align with those of the in-cone confusion rates, indicating that more pronounced differences between the *worst* and *individual*/*best* HRTF conditions might emerge in studies with larger sample sizes.

Although LE was not used as a metric in the HRTF selection, its significant increase with the *worst* HRTF indicates that binaural features have affected the PE and QE estimates within the auditory model. On the other hand, lateral localisation with the *best* HRTF was very similar, with some instances showing improved performance over the *individual* HRTF, indicating a remarkably good binaural cue matching. Overall, the LE with *individual* and non-individual HRTFs was consistent with outcomes reported in earlier research (e.g. Middlebrooks 1999b, Figure 13).

7.4.2 Sound Localisation: Performance with Individual HRTF Comparable to Previous Studies

Poirier-Quinot et al. (2022) surveyed a range of studies that trained participants in localising sound sources with various HRTFs. Their Figs. 6-12 summarise the reported mean GCEs and confusion rates throughout the training sessions. The localisation errors with *individual* HRTFs in this study are similar to or lower than the reported localisation performances of the untrained subjects with their own HRTFs in these studies (session IDs 1 and 2 in the figures). For example, mean GCEs (Figure 1 in the reference) with *individual* HRTFs from Majdak et al. (2010) and Stitt et al. (2019) tests were approximately 34° to 38°, while the mean GCE with *individual* HRTFs across all subjects in this study is 33°. The mean GCEs and confusion rates with non-individual HRTFs are generally lower in the current study than in the previous works, such as Parseihian and Katz (2012), Poirier-Quinot and Katz (2021) and Stitt et al. (2019) where the *best* and the *worst* HRTFs were selected based on the participants' subjective rankings. However, methodological

differences between the localisation tests and variations in the HRTF datasets hinder meaningful comparisons of the localisation performances across different studies.

7.4.3 Spatial Audio Quality: Variability and Context Dependence

In the qualitative assessment, the significantly smaller difference in *overall* rendering quality with the *best* versus *individual* HRTF compared to the *worst* versus *individual* HRTF partially supports **H3**. Moreover, a similar trend observed with absolute *tone colour* difference ratings suggests that perception of *overall difference* could have been influenced by the perceived differences in *tone colour* of the audio rendering.

However, the lack of a statistically significant effect in *naturalness* and *externalisation* ratings suggests that HRTF individualisation might have a smaller effect on these higher-level qualitative ratings. Brinkmann and Weinzierl (2023) report a tendency of non-individual anechoic and reverberant binaural renderings of pink noise bursts to be rated as less externalised and natural than the *individual* ones. A similar trend can be observed in the *externalisation* ratings for the pink noise bursts with the *worst* HRTFs in this study, while a number of participants rated the noise bursts with the *best* HRTF as more externalised than with the *individual* HRTF. Previous research on externalisation has largely failed to reliably demonstrate any strong impact from HRTF personalisation, indicating that variables such as reverberation and context may be more significant (Best et al. 2020). This is reflected in the increased variability of *externalisation* ratings with more semantically complex audio stimuli in this study. Similarly, *naturalness* is a context-dependent quality strongly influenced by audiovisual material. Participant feedback indicated challenges in assessing this attribute within the artificial anechoic VR environment used for this study, which is evident in highly varied ratings for this attribute. Therefore, these higher-level attributes could be deemed unreliable proxies for assessing HRTF fit without considering a more realistic test environment that includes, for example, reverberation.

Interestingly, despite some minor tendencies that could be observed in Figure 7.8, no statistically significant effect of audio stimuli (excluding the guitar sample) on ratings could be determined. This outcome is in part due to a high inter-subject rating variability and a small sample size, which decreases the statistical power of the analysis. Nevertheless, outlier analysis revealed differences in the reliability of the ratings, depending on the stimuli. It suggests that ratings with more temporally and semantically complex (as well as band-limited) audio content may be less reliable due to contextual factors affecting the results. For example, the unreliability of the solo guitar stimuli could be due to the fact that the audio loop was too long for participants

to be able to compare the same audio content side by side when switching between the HRTF conditions. Similarly, Brinkmann et al. (2019a) found that pink noise bursts resulted in more noticeable differences compared to music across various SAQI attributes when comparing measured and modelled room auralisations. Therefore, although using a more meaningful audio stimulus, such as music or speech, provides a more relatable test scenario, it also poses challenges for the test design.

7.4.4 Emerging Patterns from Cross-Experiment Analysis

The cross-experiment analysis suggests that the proposed similarity metric could serve as a valuable tool for optimising the design of listening tests. Studies involving a large number of conditions often struggle to detect perceptual differences reliably (Kim et al. 2020). By selecting HRTFs that represent perceptually extreme cases, the similarity metric can help reduce test complexity and improve the reliability of results.

In this study, the fit quality of the *best* and *worst* HRTFs varied across participants, as reflected in both localisation performance and subjective ratings. Nevertheless, the clustering observed in Figure 7.9—with *best* HRTFs appearing in the lower-left quadrant (indicating lower localisation error and perceived difference), and *worst* HRTFs in the upper-right—supports the overall validity of the grouping. Furthermore, the moderate-to-strong correlation between localisation error and perceived *overall difference* suggests that localisation performance can act as a proxy for some other perceptual qualities, although this relationship appears weaker for more context-dependent attributes like *externalisation* and *naturalness*.

Spectral centroid analysis revealed a tendency for the similarity metric to select *best* HRTFs with a lower spectral centroid and a darker perceived *tone colour* compared to *worst* HRTFs. This bias might be caused by HRTFs with more high-frequency content yielding larger modelled localisation errors, because high frequencies tend to be noisier in HRTF measurements (Andreopoulou et al. 2015). Besides, a relatively moderate correlation with large confidence intervals between the spectral centroid and the subjective *tone colour* rating suggests that other factors may influence the *tone colour* rating. For example, familiarity with spectral cues might reduce the perceived colouration of the HRTFs. However, the present analysis was limited to comparing the spectral centroids of only the front HRTF direction, whereas the qualitative assessment allowed dynamic movement of the sound sources. Therefore, more research is needed to better understand the relationship between similarity metric results, HRTF spectral content, and subjective perception of colouration.

7.4.5 Practical Limitations of HRTF Research

It is important to underline that the statistical power of the evaluation is affected by the limited participant sample size, as well as the HRTF selection pool. Although the collection of 204 HRTFs employed in this study is one of the largest measured HRTF datasets, in the context of the human population, it corresponds to only a very small sample with limited demographic diversity. One can expect the contrast between the *best* and the *worst* HRTFs to increase with increasing database size, resulting in larger observed differences in listening tests. Extending HRTF datasets is challenging, while combining multiple datasets is also non-trivial due to their inherent differences (Andreopoulou et al. 2015; Daugintis et al. 2025a).

The current study only evaluated the differences between the individual HRTF and two non-individual HRTFs, which represent the extreme ends of the proposed similarity metric. While largely validating their relative ranking (that is, *best* is better than *worst*), it cannot universally confirm that no better or worse HRTFs would be found in the dataset if participants performed the listening tests with all 204 HRTFs. Although a more detailed HRTF metric and evaluation could be devised, it presents a practical challenge of controlling the reliability of listening tests with multiple HRTFs (Andreopoulou and Katz 2016; Kim et al. 2020). Adding a few more HRTFs would increase the response noise but would not confirm that no better or worse HRTFs exist in the whole dataset. Ultimately, the classification of *best* and *worst* HRTFs remains specific to the chosen auditory model and testing context, and should not be interpreted as universally valid.

The perceptual evaluation presented in this chapter was conducted in a controlled anechoic environment, which may not fully represent real-world listening scenarios. While the static localisation test paradigm provides a sensitive framework for assessing HRTF fit, it is important to consider how these findings translate to dynamic localisation scenarios. As discussed in § 2.3, dynamic sound localisation tests present additional challenges in controlling confounding factors associated with free head movements, but experimental paradigms such as those used by Bolia et al. (1999) and Lladó et al. (2024a) could be employed to further investigate the non-individual HRTF fit. Similarly, the anechoic environment used for spatial audio quality assessment proved to be suboptimal for evaluating higher-level perceptual attributes. However, incorporating physically correct binaural reverberation for each HRTF condition is not straightforward in dynamic scenarios, and results are likely to be influenced by the specific reverberation used, making it difficult to isolate the effects of HRTF fit alone. These methodological considerations highlight the inherent trade-off between experimental control and ecological validity in perceptual HRTF evaluation.

7.5 Chapter Summary and Conclusions

The chapter presented a perceptual evaluation of the auditory model-based HRTF similarity metric introduced in Chapter 6. The evaluation comprised two listening experiments: a static sound localisation test and a dynamic spatial audio quality assessment. The experiments compared localisation performance and assessed subjective differences between the *individual* HRTF and two non-individual HRTFs (*best* and *worst*) selected from the SONICOM HRTF database, presented in Chapter 4. While the *individual* HRTF was associated with the lowest localisation errors, the *best* HRTF provided a very similar localisation performance, which could be adequate for many practical applications. In contrast, the *worst* HRTF significantly degraded localisation ability across all tested metrics. The subjective assessment revealed that while the *best* HRTF was perceived as different from the *individual* HRTF in terms of *overall rendering quality* and *tone colour*, the differences were considerably smaller than those with the *worst* HRTF. Furthermore, the cross-experiment analysis demonstrated significant correlations between localisation performance and these subjective ratings, indicating the link between localisation cue mismatch and perceived binaural rendering quality. These results validate the proposed metric's ability to identify perceptually relevant differences in non-individual HRTF fit.

Although the evaluation successfully established the positive effect of HRTF personalisation within the anechoic controlled environment, the assessment also revealed some limitations of this artificial setup in evaluating more complex aspects of rendering quality. This was exemplified by high response variability for *externalisation* and *naturalness* ratings and when listening to musical (solo guitar) stimuli. Whilst more naturalistic listening test scenarios are needed to understand the effect of HRTF fit on these higher-level perceptual attributes, they present additional challenges in controlling for the confounding factors. One such scenario, which aims to assess the affect of the HRTF fit on the perception of binaurally rendered reverberant speech, is presented in the following chapter (Chapter 8).

Chapter 8

Perceptual Evaluation on a Reverberant Speech-on-Speech Masking Task

This chapter presents a study that uses auditory model-based HRTF selection to assess the effect of HRTF personalisation on understanding reverberant masked speech. The text is adapted from the conference publication **IV**, presented at the Audio Engineering Society's 5th International Conference on Audio for Virtual and Augmented Reality (AES AVAR) (Daugintis et al. 2024).

8.1 Introduction

Chapter 7 validated the ability of the auditory-model-based HRTF similarity metric, introduced in Chapter 6, to identify well- and poorly-fitting non-individual HRTFs in terms of static sound localisation performance and perceived spatial audio quality. However, this evaluation was limited to an artificial anechoic environment. In practical applications of spatial audio, sounds may include reverberation to enhance the sense of space (Begault 2001). Furthermore, real-world VR/AR applications may exploit spatial hearing indirectly, for example when rendering speech in a multitalker teleconferencing environment. To better understand how HRTF fit influences performance under such more realistic conditions, this chapter presents a listening test investigating its effect on speech intelligibility in a reverberant speech-on-speech masking scenario.

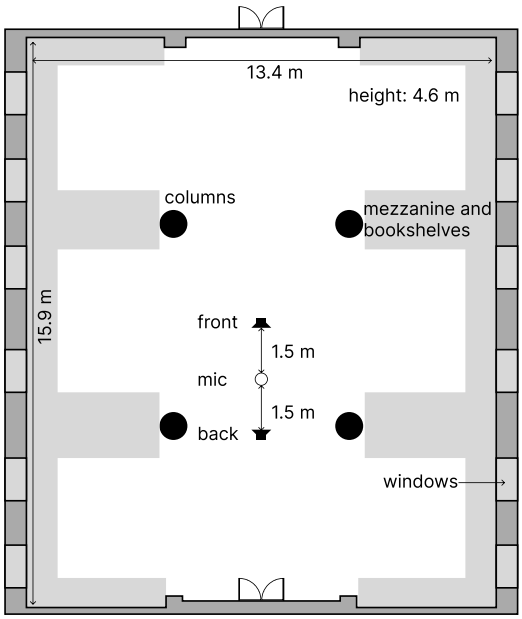
A significant, albeit small, HRTF-dependent effect has already been observed when understanding anechoic masked speech, rendered binaurally (Cuevas-Rodriguez et al. 2021; González-Toledo et al. 2024; Zenke and Rosen 2022). However, when recreating a realistic multi-talker conversation, especially in AR, reverberation mimicking a physical environment should be ap-

plied to the virtual sources. As discussed in § 2.6, perception of attended masked speech is altered by the reverberant sound field both because the interfering speech is less correlated between the two ears—making it harder to reduce its impact via the binaural unmasking mechanism—and because the target speech is blurred (Lavandier and Culling 2008). In a physical space, reverberation is spatially distributed based on the room geometry and its surfaces (Alary et al. 2021; Engel et al. 2021). A hypothesis was formed that, due to binaural unmasking effects, realistic reverb can affect the masked speech perception in a way that depends on the spatial characteristics of the reverb itself. Consequently, the choice of the HRTF may influence the perceived spatial distribution of the reverb around the listener and, in turn, the perception of speech.

To study the effect of HRTF and realistic binaural reverb reproduction on speech-in-noise understanding, a speech-on-speech masking task was designed. By asking the participants to identify binaurally rendered reverberant target speech while ignoring interfering speakers, the listening test examined the differences in speech identification performance across individual and non-individual HRTFs, as well as across different reverberation rendering and interfering speech positioning scenarios.

8.2 Reverberation Measurements

Measured spatial room impulse responses (SRIRs) were used for this study. SRIRs were measured in a large seminar room with a high ceiling and three alcoves on each side of the longer axis, separated by columns and bookcases (see Figure 8.1a). A 32-capsule Eigenmike (mh acoustics) spherical microphone array was used to obtain 4$^{\text{th}}$-order Ambisonic SRIRs. Two positions at 1.5 m distance in front (shown in Figure 8.1b) and back of the microphone array along the longer room axis were measured using Genelec 8030A loudspeaker and exponential sine sweep technique (Farina 2007). The loudspeaker distance was chosen based on the direct-to-reverberant ratio (DRR) to ensure sufficient direct sound energy and balance the test difficulty. For this study, DRR was calculated as the ratio between the summed energy of the first 30 samples from the direct impulse peak and the summed energy of the rest of the IR in decibels. The SRIRs were denoised and encoded in the spherical harmonics domain, by replacing the noisy tail of the measurements with a synthesised decaying noise sequence extending the estimated reverberation slope, as described in Massé et al. 2020. RTs, extrapolated from the 30 dB decay (T_{30}) and DRRs, calculated from the omnidirectional channel of the encoded 4$^{\text{th}}$-order Ambisonic



(a) Seminar room layout, showing the position of the microphone array and the loudspeaker for the front and back SRIR measurements.

(b) SRIR measurement setup for the front position. Reproduced from Daugintis et al. 2024.

Figure 8.1: SRIR measurement setup in the seminar room.

SRIRs of both positions are presented in Table 8.1. Although RT is similar in both positions, the DRR at the back position is slightly higher than in the front. This might have been caused by the microphone stand alignment or asymmetry of absorptive surfaces in the room.

Two reverb conditions were tested:

- realistic **dichotic**, based on binaurally decoded SRIRs,
- **diotic**, based on the convolution with the omnidirectional channel of the SRIRs only and applied equally to both ears.

The latter was employed as a baseline (the direct path was always rendered dichotically by directly convolving the speech signals with an appropriate HRTF in both reverb conditions, as discussed in § 8.4).

Table 8.1: RT (T_{30}) and DRR of the SRIRs, measured at two positions used in the test. Reproduced from Daugintis et al. 2024.

Frequency (Hz)	125	250	500	1000	2000	4000	8000
T_{30} (s), front	1.1	1.1	1.2	1.1	0.9	0.7	0.5
T_{30} (s), back	1.1	1.1	1.1	1.1	0.9	0.8	0.6
Broadband DRR (dB), front				1.9			
Broadband DRR (dB), back				2.8			

8.3 Speech-on-Speech Masking Task

The test was based on a speech identification task, using the coordinate response measure (CRM) speech corpus (Bolia et al. 2000). By superposing competing audible semantic streams that have to be disentangled (i.e. simultaneous talkers), an informational masking scenario was devised. The choice was motivated by previous findings that such tasks benefit more from the perceived spatial separation of sources, as compared to purely energetic masking cases, when masking occurs only due to overlap of spectral energy (e.g. speech in background noise) (Brungart 2001b).

During each trial, three simultaneous phrases of a form 'Ready [*call sign*] go to [*colour*] [*number*] now' were presented by different talkers. Participants had to listen to the target call sign 'Baron' and identify the heard colour-number combination while ignoring the other two interfering speakers. The test followed a procedure similar to the one described by González-Toledo et al. (2024). The original recordings of the CRM corpus, low-pass filtered at 18 kHz (instead of the publicly available dataset, which is filtered at 8 kHz) were used in this study because of the influence of high-frequency content on vertical and front/back discrimination via the spectral pinnae cues (Martin et al. 2012). Although the intelligibility of CRM corpus across different talkers, colours, and numbers is not fully balanced (Brungart 2001a), the easy structure of the test, which allows for quick unsupervised response procedure, meant that multiple repetitions of each condition could be used to smooth the result variability.

8.4 Binaural Rendering

The sounds for this study were rendered binaurally. Two spatial positions were used: in front of the listener and directly opposite at the back. The target speech (corresponding to the call sign 'Baron' that participants had to attend to) was placed in front while both interferers were positioned either in front (co-located condition) or at the back (separated condition). These directions were chosen to limit the effect of better ear listening, and because they are known to be particularly difficult to externalise, so the quality of rendering due to the choice of an HRTF or reverberation condition could have a stronger effect on the task performance.

For the binaural reproduction, the direct sound part and the reverberation were rendered separately and summed up at the end, similarly to Engel et al. 2021. The SRIR peaks plus 30 samples were removed from the reverberation part. Instead, delayed and level-adjusted mono speech samples were rendered directly through the appropriate HRTF using SPARTA

binauraliser[1].

For the reverberant part, the rendering pipeline depended on the reverberation condition. In the dichotic case, the same dry speech samples were convolved with the 4^{th}-order Ambisonic SRIRs using spat5[2] multichannel convolver and then decoded using SPARTA AmbiBIN[3] ambisonic-to-binaural decoder. The magnitude least-squares method was used for the decoding to preserve the relevant ITD cues in the low frequencies (Schörkhuber et al. 2018). For the diotic reverb version, the dry speech samples were convolved with only the first channel of the Ambisonic SRIRs (which corresponds to the 0^{th}-order omnidirectional part of the SRIR) and the same reverberant signal was used for both headphone channels. The direct and the reverberant part of the output signal were summed up at the end and the resulting signal convolved with an individual headphone equalisation filter from the SONICOM HRTF database (Engel et al. 2023).

The direct and the reverberant (in the dichotic reverb scenario) parts of the sound were rendered binaurally using one of three different HRTFs: individual and two non-individual HRTFs from the SONICOM HRTF dataset (Engel et al. 2023). These two were selected for each subject using a recently proposed non-individual HRTF matching procedure (Daugintis et al. 2023b and § 6), based on the *best* and the *worst* predicted static sound localisation performance according to auditory model calculations (Barumerli et al. 2023). For each participant, the two non-individual HRTFs were selected from the remaining 199 HRTFs, available from the SONICOM HRTF database.

8.5 Listening Test Setup and Procedure

The test was administered via a laptop running Max 8 (Cycling '74) in an acoustically treated room (the same room was used to measure HRTFs; its acoustic parameters are reported in Engel et al. 2023). The sound was played via the Motu M2 audio interface and Sennheiser HD 650 headphones. The audio gain was fixed to the same level across all the participants (The sound pressure level at the ear canal was measured to be approximately 60-65 dBA during the playback).

The graphical user interface (GUI) of the test is shown in Figure 8.2. The test was divided into 12 blocks, 30 trials each. The blocks corresponded to the combinations of 3 HRTF and 2

[1] https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#binauraliser
[2] https://forum.ircam.fr/projects/detail/spat/
[3] https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#ambibin

reverberation conditions, each repeated twice. Within each block, the co-located and separated maskers conditions were interleaved. Therefore, there were 30 trials for each HRTF, reverberation, and maskers condition in the test. Participants were unaware of the different conditions but were asked to concentrate on the target speaker, which was always rendered in front of the listener.

Before the test, participants were presented with a familiarisation block. The first 10 trials contained only the target speech, while the next 20 both the target and one masker. During the training, HRTF and reverberation conditions, as well as masker positions, were randomised without replacement. After each trial of the training session, the participants were informed whether their answers were correct.

A total of 13 participants performed the test. All of them were either native English speakers or had resided in an English-speaking country for a sustained period.

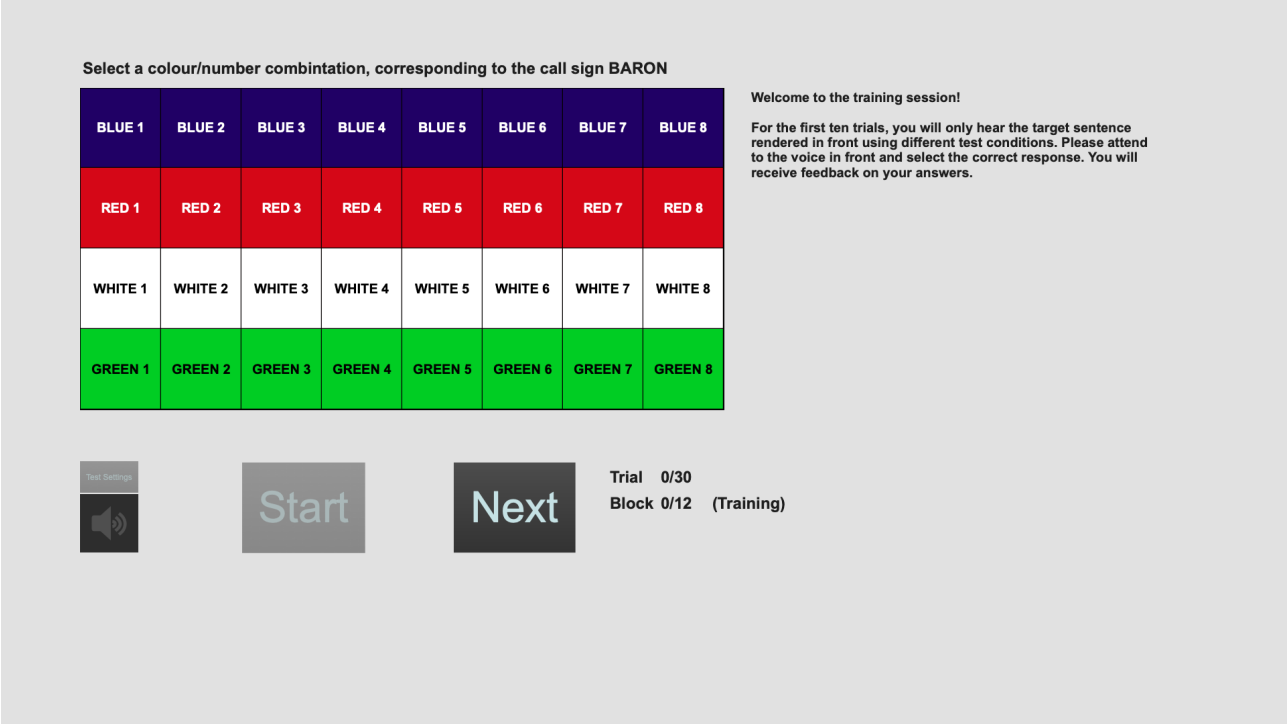


Figure 8.2: Graphical user interface of the listening test. Reproduced from Daugintis et al. 2024.

8.6 Results

The main results from the listening test are presented in Figure 8.3. The boxplot shows the percentage of correct responses (colour-number combinations) across participants when the masker speakers were separated from the target and placed at the back (left panel) and when they were co-located with the target at the front (right panel). Results in each panel are also

subdivided by the reverberation and HRTF conditions.

To understand the effect of maskers, reverb, and HRTF conditions on the percentage of correct responses, a three-way RM ANOVA was performed. The normality of the distributions was confirmed by the Shapiro-Wilk test and the sphericity by Mauchly's test. RM ANOVA revealed a statistically significant effect of the maskers' position ($F(1, 12) = 179.294$, $p < 0.0001$), indicating, as expected, that it was more challenging to understand the target speech when the interfering speech was co-located. There was also a significant interaction between the maskers' position and the reverb type ($F(1, 12) = 27.131$, $p < 0.001$). Additional two-way RM ANOVAs were performed on the two separate groups split by maskers' position. It showed a significant effect of reverberation type in the case of co-located maskers ($F(1, 12) = 28.998$, $p < 0.001$). The figure reveals that when the maskers were co-located the task was on average more difficult to perform in the diotic reverb scenario. No significant effect of reverberation was found in the separated maskers case, as well as no significant differences between HRTF conditions.

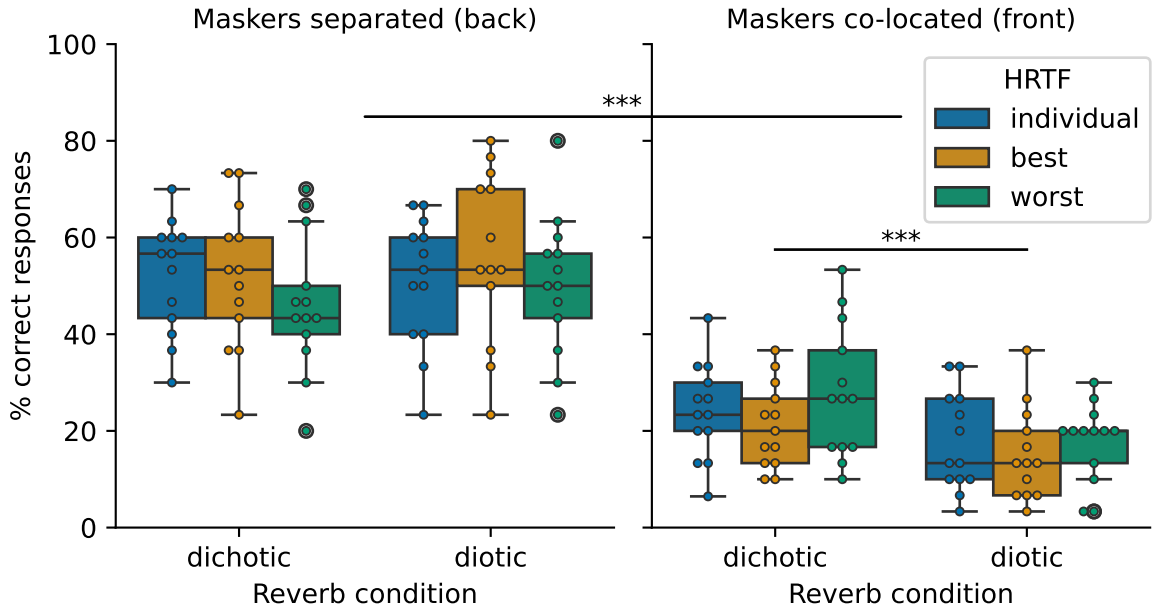


Figure 8.3: Percentage correct responses for different maskers positions and different reverb and HRTF conditions. Points represent individual performances and boxplots show medians and the inter-quartile range of the performance distributions. `***' denotes significant differences ($p < 0.001$) in distributions between maskers' positions and between reverb conditions in the case of the co-located maskers. Reproduced from Daugintis et al. 2024.

8.6.1 Spatial Release from Masking

As discussed in § 2.6, the improvement of speech perception when the target and masking speech sources are spatially separated is often refered to as spatial release from masking (SRM).

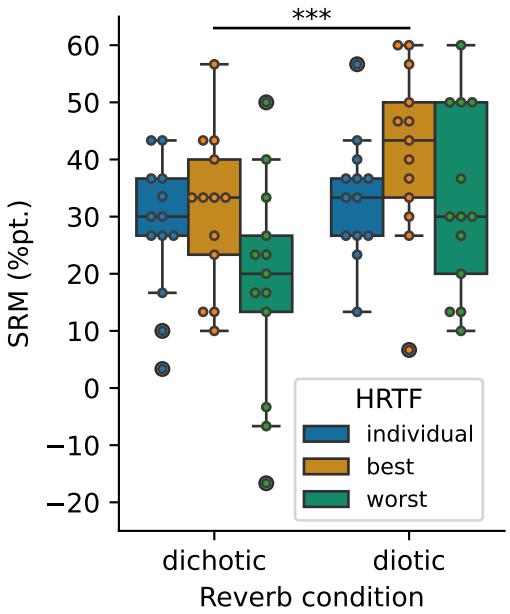


Figure 8.4: Spatial release from masking as percentage point differences between test performance with separated vs co-located maskers for different reverb and HRTF conditions. Points represent individual results and boxplots show medians and the inter-quartile range of the distributions. `***' shows significant differences ($p < 0.001$) in distributions between different reverb conditions. Reproduced from Daugintis et al. 2024.

In this study, it was analysed similarly to González-Toledo et al. 2024 by calculating the difference between individual performances in separated vs co-located scenarios.

Figure 8.4 shows SRM in terms of percentage point differences in performances across participants for different reverberation and HRTF conditions. Two-way RM ANOVA was performed to understand the effect of reverb and HRTF conditions on SRM. It revealed significant differences between the reverb conditions ($F(1, 12) = 27.133$, $p < 0.001$). On average, the SRM was higher in the diotic reverberation case: while the performance was similar between the reverberation conditions when the maskers were separated from the target, it was impinged by the diotic reverb more than the dichotic one in the co-located maskers scenario.

On the other hand, the differences between HRTFs didn't meet the significance criteria ($F(2, 24) = 2.889$, $p = 0.075$). A slight trend of lower median SRM with the *worst* HRTF condition (especially in the dichotic case) can be observed in the figure but data from more participants might be required to resolve its significance.

8.6.2 Comparison with Objective Metrics

The presented results indicate that, at a group level, the speech perception performance does not depend on the specific HRTF selection used in this study. However, performance differences

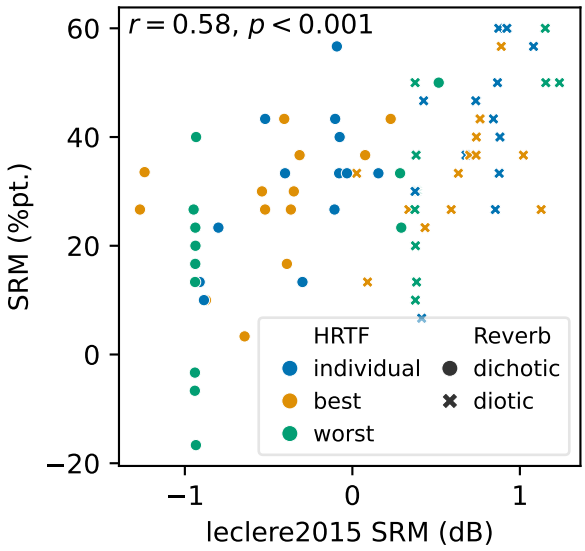


Figure 8.5: Spatial release from masking (SRM) as a percentage-point difference of participants' performance between the co-located and separated maskers position, plotted against the SRM (in dB), calculated from BRIRs using *leclere2015* auditory model. Each point represents an individual result with a specific HRTF and reverb condition. Reproduced from Daugintis et al. 2024.

between HRTF conditions exist on an individual level. To assess if these differences could be explained by some other objective metrics, an auditory speech intelligibility model was used. The model used—*leclere2015* from the Auditory Modelling Toolbox (Majdak et al. 2022)—estimates SRM as an effective ratio (in dB) between a reverberant speech target and a spatially separated static broadband noise masker (Leclère et al. 2015). The model was supplied with the BRIRs from the target and the separated masker positions, which were recorded from the Max patcher output to account for the full audio rendering pipeline. The model relies on a few assumptions about the cut-off between useful early reflections and detrimental late reverberation, which are room-dependent, but the default room-independent parameters were used in this exercise. For each participant, the SRMs with all three HRTFs and two reverb conditions were estimated.

The model estimates in dB are plotted against the measured SRM in percentage points and shown in Figure 8.5. The linear relationship between the two metrics was assessed by calculating Pearson's correlation coefficient. It revealed a moderate positive correlation between the model output and the listening test results ($r = 0.58$, $p < 0.001$). Generally, the model predicted mostly negative SRM values for the dichotic reverb scenarios and positive SRM for the diotic ones, but the absolute values might not be easy to interpret without the model calibration. Furthermore, the *worst* HRTFs are clustered around a few *leclere2015* SRM values. This bias is a result of the HRTF selection procedure, which often favours the same HRTF to be chosen as the *worst* for multiple subjects (Daugintis et al. 2023b). Besides this clustering of the *worst* HRTF, there are no measurable differences between the distributions associated with different HRTF conditions.

8.7 Discussion

The study results suggest that when reproducing a multi-speaker virtual environment, there is some benefit in using fully directional dichotic reverberation as compared to diotic binaurally-correlated reverb of the same strength. These results follow previous findings of speech intelligibility improvement when artificial reverberation is decorrelated between the two ears (Orduña-Bustamante et al. 2018) or increased speech reception threshold in noise and voice interferers when binaural reverberation is used over the diotic one (Lavandier and Culling 2008). However, the effect in this study is only present in the more difficult scenario of the co-located target and maskers. One possible explanation is that the separated condition is easier to perform, resulting in participants not having to rely on subtle spatial reverberation cues. In contrast, the participants found the co-located condition very difficult. In general, for the majority of the incorrect responses (88 % of the cases) the participants selected the colour and/or the number, uttered by one of the two interfering speakers, so 11 % ($1/9^{th}$) correct response rate could be considered a chance level. Median responses for the co-located maskers and diotic reverberation condition balances just above this threshold, indicating that many participants were guessing the answers during this condition. Therefore, adjusting the difficulty level by manipulating the DRR and number of interfering speakers could lead to a more sensitive test and accentuated differences between conditions.

Nonetheless, the diotic reverberation condition used in this test can only be seen as a worst-case baseline since it does not represent a realistic reverberation scenario. Highly correlated binaural reverberant signals that are not filtered through HRTFs offer little externalisation (besides the direct part). It would be interesting to compare these results with a more realistic reverberation that is uncorrelated between the ears, but acoustically isotropic. Furthermore, the test performance may be dependent on the exact acoustics of the measured space, so contrasting the isotropic reverberation with a highly anisotropic one is subject to future research.

On the other hand, the HRTF classification based on modelled sound localisation does not appear to relate to the speech reception performance in a statistically significant way. Nevertheless, a minor pattern, favouring individual and *best* HRTFs over the *worst* one starts to emerge in the SRM results. Thus, data from more subjects might be needed to draw a more robust conclusion on the role of HRTF personalisation. Although the correct perception of the source location has been previously shown to play a role in unmasking the target speech (Brungart 2001b), other factors such as binaural and spectral signal-level differences between the two BRIR positions might have a stronger influence on the performance, irrespective of the

HRTF individualisation. This is somewhat illustrated by the moderate result correlation with the speech intelligibility model, which is only based on signal-level BRIR differences. Although the interpretation of the modelled results in this scenario is limited because the model only considers static energetic masking and doesn't account for the informational masking part, better control of the test conditions, correcting for the signal-level BRIR differences, could provide more meaningful results. Finally, alternative HRTF selection methodologies and a bigger participant sample size might help better separate the HRTF personalisation effect from purely signal-level differences.

Further improvements could be made by administering the test in a more immersive virtual environment, which visually matches the rendered acoustic space. Better visual-acoustic congruence may aid the sense of sound externalisation and may in turn impact the observed effect, providing a more realistic assessment of the speech unmasking phenomena.

8.8 Chapter Summary and Conclusions

This chapter presented a binaurally rendered speech-on-speech masking test, which extended the perceptual evaluation of the auditory-model-based HRTF selection from Chapter 7 to a more realistic listening scenario. The study examined the effect of using *individual*, *best*, and *worst* HRTF as well as realistic dichotic versus diotic reverberation rendering on understanding reverberant masked speech. Listeners were asked to attend to a target talker positioned virtually in front while ignoring interfering maskers that were either co-located with the target or spatially separated and positioned behind the listener. Although a statistically significant improvement in speech perception was found when using fully spatial (dichotic) reverberation compared with diotic rendering, no statistically significant effect of HRTF selection emerged at the group level. This suggests that HRTF individualisation may play a more limited role in complex, reverberant listening environments than in controlled anechoic conditions, investigated in Chapter 7.

Despite the absence of a group-level effect, notable HRTF-related differences persisted across individual participants, and these differences showed moderate correlations with auditory model predictions of the spatial release from masking (SRM). This relationship points to the influence of energetic masking, which is driven primarily by the signal-level characteristics of the BRIRs rather than their degree of individualisation, on listeners' ability to separate target speech from maskers. These findings might be valuable for designing efficient VR/AR applications with a focus on social group interactions and telepresence.

Chapter 9

Conclusions

9.1 Summary of Thesis Achievements

This thesis investigated the link between numerical and perceptual non-individual head-related transfer function (HRTF) similarity. Through the development of an HRTF similarity analysis based on a *barumerli2023* auditory model, its evaluation using perceptual listening tests and, more broadly, its use in various HRTF evaluation scenarios, it addressed the main research question on the ability of the Bayesian auditory model to predict perceptual fit of non-individual HRTFs and the correlation between the HRTF fit and performance in spatial audio tasks.

It began by introducing the HRTF in the context of spatial hearing cues, outlining the fundamentals of human spatial hearing (**Chapter 2**). The chapter also discussed various attempts to assess and simulate human spatial hearing. It introduced the *barumerli2023* Bayesian auditory model for sound localisation, which served as a backbone for the perceptually motivated HRTF analysis throughout the rest of the thesis. The HRTF was then defined from the acoustic and signal processing perspectives, and its role in the binaural spatial audio rendering was discussed in **Chapter 3**. The chapter examined the individual variability of HRTFs and reviewed existing methods for acquiring or personalising them to enhance perceptual rendering quality. In addition, it surveyed common evaluation techniques, highlighting the need to link numerical HRTF differences with perceptual impact on listener experience.

Chapter 4 then presented the measurement and release of a freely available SONICOM HRTF dataset, containing more than 300 high-quality individual HRTF measurements, which provides a key contribution to the field of data-driven HRTF personalisation research (**Contribution 1**). It also successfully demonstrated the use of the Bayesian auditory model to provide a perceptually grounded numerical evaluation for HRTF measurement consistency, fulfilling Thesis **Objective 1**.

Since its release, the SONICOM HRTF dataset has been used in multiple studies researching topics such as audio augmented reality (Bhattacharyya et al. 2024; Martin et al. 2025b), human adaptation to altered hearing cues (Guiraud et al. 2024; Meyer and Picinali 2025b), binaural cue augmentation (Marggraf-Turley et al. 2025a), and neurophysiology of spatial hearing (Marggraf-Turley et al. 2025b).

Chapter 5 expanded on the existing knowledge of HRTF personalisation research, presented in the previous chapters, and identified some current challenges within the field, particularly relevant in the context of increasing use of data-driven approaches. It emphasised the importance of developing perceptually motivated evaluation techniques to assess the HRTFs and discussed the benefits of engaging with the wider research community to address the remaining challenges. These aspirations were illustrated through the organisation of the first Listener Acoustic Personalisation (LAP) challenge with the aim to benchmark algorithms for HRTF dataset harmonisation and HRTF spatial upsampling (**Contribution 2**). It detailed the development of challenge evaluation framework and, expanding on **Objective 1**, presented another application of the Bayesian auditory model as a perceptual validation tool for harmonised HRTFs. It also discussed important considerations and limitations of using auditory models as proxies for human listeners, which provided a valuable background for more in-depth investigation into the use of the model in assessing non-individual HRTF similarity. In general, the LAP challenge successfully engaged with the wider research community by inviting submissions for both tasks, some of which have been published as conference and journal papers (Arevalo and Villegas 2025; Ito et al. 2025; Masuyama et al. 2025; Zhao et al. 2025).

Chapter 6 built on the specific use-cases of the Bayesian auditory model presented in the earlier chapters and offered a more systematic investigation of the model and its application to assessing non-individual HRTF similarity. It addressed **Objective 2** by presenting the calibration of the *barumerli2023* model using prior human sound localisation data from 16 subjects and the development of a new non-individual HRTF similarity metric. Based on distributions of predicted polar and quadrant localisation errors within a limited directional range in front of the listener, the metric classified non-individual HRTFs into *good* and *bad* categories, selecting the *best*/*worst* from each group and investigated the robustness of this classification to model parameters. In this chapter, the classification was also compared to the results of a perceptual HRTF rating study, conducted by Katz and Parseihian (2012). Although the overall correlation between the two methods was low, the analysis identified several factors that may explain the discrepancy, including differences in experimental conditions and the inherent challenges of achieving high repeatability in subjective listening tests. An adjusted analysis that accounted for the specific

stimulus directions used in the perceptual experiment showed improved correlations between some perceptual ratings and the model predictions. This finding suggests that the proposed metric may capture certain aspects of perceived HRTF rendering quality, while also highlighting the need for dedicated perceptual validation, motivating the listening studies presented in the subsequent chapters.

Chapter 7 presented the design and results of the perceptual evaluation, aimed at validating the selection of the *best* and the *worst* non-individual HRTF, according to the proposed metric, and addressing **Objective 3**. It presented static localisation and dynamic spatial audio quality assessments, performed by 17 participants. It revealed that the localisation performance with the *best* HRTF was largely comparable to the *individual* HRTF, while the *worst* HRTF significantly degraded the performance and was qualitatively perceived as more different. The cross-experiment result analysis also demonstrated a link between the HRTF fit in terms of localisation performance and its perceived similarity to the individual HRTF. The consistency of qualitative ratings decreased with more semantically complex sound stimuli and for higher-level cognitive attributes, namely *externalisation* and *naturalness*. These results support the use of the proposed auditory model-based metric as a tool to assess the non-individual HRTF fit without extensive perceptual assessments (**Contribution 3**). It also shows that while individual HRTFs ensure the best localisation performance, a well-fitting non-individual HRTF may be sufficient for many applications, especially when dynamic rendering and more complex sound scenes are considered (**Contribution 4**).

Finally, **Chapter 8** expanded on the perceptual evaluation of the proposed metric (**Objective 3**) to a more complex listening scenario involving speech comprehension in reverberant environments. It presented a behavioural experiment using a headphone-based speech-on-speech masking test with 13 participants. The study examined two factors affecting the comprehension of reverberant masked speech. Firstly, it compared the performance when using individual HRTFs, the *best* non-individual HRTFs, and the *worst* non-individual HRTFs. Secondly, it compared dichotic reverberation, decoded from 4th-order Ambisonic SRIRs, with diotic reverberation derived from their omnidirectional Ambisonic channel. At the group level, no statistically significant effect of HRTF fit on speech comprehension was observed. However, a significant improvement in speech perception was observed when dichotic reverberation was used instead of diotic, but only when the maskers were co-located with the target speaker. Notably, individual performance differences related to HRTFs were present and showed a moderate correlation with SRM predictions from the binaural speech intelligibility model. These findings suggest that the benefits of HRTF personalisation may be context-dependent, highlighting its limited

importance in scenarios involving the comprehension of binaurally rendered reverberant speech (**Contribution 4**).

9.2 Overview of Future Directions

The perceptual evaluation of the proposed auditory model-based HRTF similarity metric opens up several avenues for future research, including:

- use of the metric in studies on **human adaptation to HRTFs**
- use of the metric in **HRTF personalisation** research
- further **improvements to the metric**
- development of **more complex listening experiments**

These directions are discussed in more detail in this section.

In research on **human adaptation to non-individual HRTFs**, this metric could be used to select non-individual HRTFs for training, enabling researchers to evaluate how the initial HRTF fit influences the adaptation process. Several studies have already investigated this effect. Parseihian and Katz (2012) found that well-matched HRTFs gave participants an advantage to achieve better localisation performance after rapid training, while Stitt et al. (2019) observed that training with poorly matched HRTFs was not successful for some participants. Similarly, Trapeau et al. (2016) found that the extent of disturbance introduced by ear moulds partly explained the amount and speed of adaptation to altered hearing cues. However, the perceptual HRTF fitting used by Parseihian and Katz (2012) and Stitt et al. (2019) has been shown to be unreliable for some listeners (Andreopoulou and Katz 2016), indicating the need for more robust HRTF selection methods to further investigate the adaptation performance.

Meyer and Picinali (2025b) has already applied the metric proposed in this thesis to compare sound localisation training performance using the *worst* versus *individual* HRTFs, demonstrating its potential for adaptation studies. While the results showed no significant differences between the two conditions, the training protocol was relatively short and involved a limited number of participants, which likely reduced observable adaptation effects. Preliminary results from an extended study, however, revealed significant differences in localisation performance between the *individual* and the *worst* HRTF (Meyer and Picinali 2025a). Given the growing interest in identifying a single HRTF suitable for widespread use across binaural audio applications to promote long-term adaptation (Lladó et al. 2024b; Meyer-Kahlen et al. 2025), the similarity metric proposed in this thesis could help assess the feasibility of such a universal HRTF.

The proposed metric could also support the development and evaluation of **HRTF personalisation** algorithms. Although numerous personalisation methods have been proposed, few have been validated by perceptual listening tests due to the challenges of conducting comprehensive behavioural experiments (Fantini et al. 2025). Instead, most studies rely on objective measures such as spectral differences, which, as discussed in Chapter 3, do not fully capture perceptual aspects of HRTF similarity. Some researchers have employed auditory models, such as the one by Baumgartner et al. (2014), to provide perceptually grounded evaluations of personalisation methods (Fantini et al. 2021; Zhang et al. 2020; Zhao et al. 2024). This approach could be extended by incorporating the *barumerli2023* model for full-sphere localisation performance evaluation. In particular, the metric proposed in this thesis could be adapted as an efficient benchmarking tool for comparing HRTF personalisation algorithms and identifying the most effective strategies.

Currently, the proposed HRTF similarity metric requires a reference individual HRTF for comparison, limiting its direct application in personalisation techniques. Future research could explore **improving the metric** for this purpose. Key questions include determining the minimum number of individual HRTF measurement positions needed to reliably select the *best* non-individual HRTF and assessing the metric's robustness to measurement quality, for example, by introducing background noise or room response. Adapting auditory model-based selection to these limitations could enable the development of an HRTF personalisation method based on a quick, low-quality measurement procedure outside specialist facilities. This setup could resemble the approach proposed by Reijniers et al. (2020), using a single stationary speaker as a sound source, recording a few sample impulse responses along with listener orientation via head-tracked in-ear microphones, and matching them to a high-quality, best-fitting HRTF from a laboratory-measured dataset.

The correspondence between sound localisation performance and qualitative evaluations of non-individual HRTFs, demonstrated in Chapter 7, suggests opportunities to extend the metric beyond predicting localisation performance. While sound localisability is the most straightforward perceptual dimension of spatial audio quality, other perceptual attributes are also important to consider (Pelzer et al. 2020; Simon et al. 2016). Some efforts have already been made to predict additional perceptual qualities associated with binaural listening, such as externalisation (Baumgartner and Majdak 2021), colouration (McKenzie et al. 2022; McKenzie and Brinkmann 2025), and binaural speech intelligibility (Lavandier et al. 2022). Consequently, further research into perceptually motivated HRTF similarity metrics could explore integrating multiple perceptual models into a unified framework. Such a combined metric could be made adaptive

by weighting different perceptual predictions according to specific application needs—for example, minimising perceived colouration for spatial music listening, or maximising speech intelligibility in teleconferencing scenarios.

This thesis primarily focused on identifying perceptually distant HRTFs and examining their impact on spatial audio perception. Limiting the range of HRTF conditions helped optimise the length and complexity of experimental procedures. However, future work could extend the metric to represent a more graded HRTF similarity scale. However, two challenges are important to consider:

1. The complexity of spatial sound perception makes it difficult to define a single, monotonic perceptual scale. Perceived HRTF fit may vary depending on spatial regions of interest, error metrics used, and perceptual attributes considered. While correlations exist across some factors, as shown in the cross-experiment analysis in § 7.3.3, any scale will likely remain context-specific.
2. Prior studies (Andreopoulou and Katz 2016; Kim et al. 2020) indicate that some listeners struggle to perform perceptual HRTF evaluation tasks consistently. Inherent experimental uncertainties impose practical limitations on developing a more finely graded qualitative scale that does not overfit the data from a single experiment.

These limitations suggest that a more detailed perceptual metric would only be meaningful if its intended use case is clearly defined.

In general, the studies presented in this thesis were conducted under controlled laboratory conditions but **more natural listening scenarios** should be investigated in the future. While the controlled experiments revealed some effects of HRTF personalisation, it remains unclear how these findings generalise to more realistic spatial audio scenarios. Authentic VR/AR experiences may involve dynamic room acoustics, six-degree-of-freedom navigation, and interactive audiovisual tasks. Studies such as those by Poirier-Quinot and Katz (2020) and Roßkopf et al. (2024), along with the reverberant speech experiment in Chapter 8, demonstrate a common failure to determine clear benefits of personalised HRTFs in complex listening environments. However, these studies also highlight the difficulty of distinguishing a true null effect of HRTF personalisation from insufficient test sensitivity, complicating generalisation to broader applications. This reflects a wider concern about the specificity of so-called 'ecologically valid' experiments in psychology (Holleman et al. 2020). Future research should therefore aim to design more complex yet robust experimental protocols to clarify the benefits and limitations of HRTF personalisation in contexts that reflect specific real-world use cases of spatial audio technology.

Some of these future research aims can benefit from engagement with a wider research community. Organisation of the Listener Acoustic Personalisation (LAP) challenge, discussed in Chapter 5, demonstrated that engaging with other researchers—both internally during challenge development and externally when anticipating and evaluating proposed solutions—helps identify domain challenges, benchmark potential solutions, uncover limitations, and advance the state of the art. This format could therefore be applied to a broader range of topics, including the development of comprehensive listening experiments and the design and evaluation of auditory models.

9.3 Personal Concluding Remarks

Looking ahead, I believe it is important to reflect on the ultimate goal of binaural spatial audio personalisation research, which is to create accessible technologies that adapt to individual needs. Spatial audio can enhance presence and collective embodiment in virtual social experiences (Dicke et al. 2010; Kukshinov and Nacke 2025; Nowak et al. 2023; Yang et al. 2020), but it also introduces novel challenges around user equity, safety, and privacy. For example, a spatialised voice can be used as a more natural way to connect across distances. However, in the context of emerging social virtual networks, it can facilitate new forms of embodied harassment (Freeman et al. 2022, 2024) when used as a tool to frighten and intimidate other users (MacArthur et al. 2024; Zheng et al. 2023). When designing such virtual spaces, it is therefore crucial to consider the potential for misuse and implement user-centric safeguarding tools (Zheng et al. 2023). Similarly, the difficulty in judging conversation privacy in VR (Williamson et al. 2021) requires technological solutions to facilitate the user experience (Wang et al. 2025). Also, personalising spatial audio relies on sensitive biometric data, underscoring the importance of robust privacy protections (Mate et al. 2025). Therefore, as researchers strive for perceptual realism in virtual environments, it is essential to acknowledge the social implications of these technologies and address emerging challenges to ensure they drive positive, inclusive change.

Bibliography

3DTune-In (2023). *Unity Wrapper for 3D-Tune-In Toolkit (version 3.1)*. (Last viewed May 8, 2025). URL: github.com/3DTune-In/3dti_AudioToolkit_UnityWrapper/tree/v3.1.

Aggius-Vella, E. et al. (2020). 'Comparison of Auditory Spatial Bisection and Minimum Audible Angle in Front, Lateral, and Back Space'. *Sci. Rep.* 10.1, p. 6279. DOI: 10.1038/s41598-020-62983-z.

Ahrens, A., Cuevas-Rodriguez, M. and Brimijoin, W. O. (2021). 'Speech Intelligibility with Various Head-Related Transfer Functions: A Computational Modelling Approach'. *JASA Express lett. (JASA-EL)* 1.3, p. 034401. DOI: 10.1121/10.0003618.

Alary, B., Massé, P., Schlecht, S. J., Noisternig, M. and Välimäki, V. (2021). 'Perceptual Analysis of Directional Late Reverberation'. *J. Acoust. Soc. Am.* 149.5, pp. 3189–3199. DOI: 10.1121/10.0004770.

Algazi, V. R., Avendano, C. and Duda, R. O. (2001a). 'Estimation of a Spherical-Head Model from Anthropometry'. *J. Audio Eng. Soc. (AES)* 49.6.

Algazi, V. R., Avendano, C. and Duda, R. O. (2001b). 'Elevation Localization and Head-Related Transfer Function Analysis at Low Frequencies'. *J. Acoust. Soc. Am.* 109.3, pp. 1110–1122. DOI: 10.1121/1.1349185.

Algazi, V., Duda, R., Thompson, D. and Avendano, C. (2001c). 'The CIPIC HRTF Database'. In: *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*. New Platz, NY, USA, pp. 99–102. DOI: 10.1109/ASPAA.2001.969552.

Ananthabhotla, I., Ithapu, V. K. and Brimijoin, W. O. (2021). 'A Framework for Designing Head-Related Transfer Function Distance Metrics That Capture Localization Perception'. *JASA Express lett. (JASA-EL)* 1.4, p. 044401. DOI: 10.1121/10.0003983.

Andreopoulou, A., Begault, D. R. and Katz, B. F. G. (2015). 'Inter-Laboratory Round Robin HRTF Measurement Comparison'. *IEEE J. Sel. Topics Signal Process.* 9.5, pp. 895–906. DOI: 10.1109/JSTSP.2015.2400417.

Andreopoulou, A. and Katz, B. F. G. (2016). 'Investigation on Subjective HRTF Rating Repeatability'. In: *Proc. Audio Eng. Soc. (AES) Conv.* Paris, France, p. 9597.

Andreopoulou, A. and Katz, B. F. G. (2017). 'Identification of Perceptually Relevant Methods of Inter-Aural Time Difference Estimation'. *J. Acoust. Soc. Am.* 142.2, pp. 588–598. DOI: 10.1121/1.4996457.

Andreopoulou, A. and Roginska, A. (2011). 'Towards the Creation of a Standardized HRTF Repository'. In: *Proc. Audio Eng. Soc. (AES) Conv.* New York, NY, USA.

Arbel, L., Ananthabhotla, I., Ben-Hur, Z., Alon, D. L. and Rafaely, B. (2024). 'On HRTF Notch Frequency Prediction Using Anthropometric Features and Neural Networks'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Seoul, Korea, pp. 816–820. DOI: 10.1109/ICASSP48485.2024.10447610.

Arend, J. M., Brinkmann, F. and Pörschmann, C. (2021a). 'Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions'. *J. Audio Eng. Soc. (AES)* 69.1/2, pp. 104–117. DOI: 10.17743/jaes.2020.0070.

Arend, J. M., Ramírez, M., Liesefeld, H. R. and Pörschmann, C. (2021b). 'Do Near-Field Cues Enhance the Plausibility of Non-Individual Binaural Rendering in a Dynamic Multimodal Virtual Acoustic Scene?' *Acta Acust.* 5, p. 55. DOI: 10.1051/aacus/2021048.

Arevalo, C. and Villegas, J. (2025). 'Spatial Upsampling of Head-Related Impulse Responses via Elevation-Wise Encoder-Decoder Networks'. *IEEE Open J. Signal Process.*, pp. 1–9. DOI: 10.1109/OJSP.2025.3613209.

Armstrong, C., Thresh, L., Murphy, D. and Kearney, G. (2018). 'A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database'. *Appl. Sci.* 8.11, p. 2029. DOI: 10.3390/app8112029.

Asano, F., Suzuki, Y. and Sone, T. (1990). 'Role of Spectral Cues in Median Plane Localization'. *J. Acoust. Soc. Am.* 88.1, pp. 159–168. DOI: 10.1121/1.399963.

Ashida, G. and Carr, C. E. (2011). 'Sound Localization: Jeffress and Beyond'. *Curr. Opin. Neurobiol.* 21.5, pp. 745–751. DOI: 10.1016/j.conb.2011.05.008.

Bahu, H., Carpentier, T., Noisternig, M. and Warusfel, O. (2016). 'Comparison of Different Egocentric Pointing Methods for 3D Sound Localization Experiments'. *Acta Acust. united with Acust.* 102.1, pp. 107–118. DOI: 10.3813/AAA.918928.

Bakdash, J. Z. and Marusich, L. R. (2017). 'Repeated Measures Correlation'. *Front. Psychol.* 8, p. 456. DOI: 10.3389/fpsyg.2017.00456.

Balachandar, K. and Carlile, S. (2019). 'The Monaural Spectral Cues Identified by a Reverse Correlation Analysis of Free-Field Auditory Localization Data'. *J. Acoust. Soc. Am.* 146.1, pp. 29–40. DOI: 10.1121/1.5113577.

Barumerli, R., Geronazzo, M. and Avanzini, F. (2018a). 'Localization in Elevation with Non-Individual Head-Related Transfer Functions: Comparing Predictions of Two Auditory Models'. In: *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pp. 2539–2543. DOI: 10.23919/EUSIPCO.2018.8553320.

Barumerli, R., Geronazzo, M. and Avanzini, F. (2018b). 'Round Robin Comparison of Inter-Laboratory HRTF Measurements – Assessment with an Auditory Model for Elevation'. In: *IEEE 4th VR Workshop Sonic Interact. Virtual Environ. (SIVE)*, pp. 1–5. DOI: 10.1109/SIVE.2018.8577091.

Barumerli, R. and Majdak, P. (2025). 'FrAMBI: A Software Framework for Auditory Modeling Based on Bayesian Inference'. *Neuroinform.* 23.2, p. 20. DOI: 10.1007/s12021-024-09702-5.

Barumerli, R., Majdak, P., Geronazzo, M., Meijer, D., Avanzini, F. and Baumgartner, R. (2023). 'A Bayesian Model for Human Directional Localization of Broadband Static Sound Sources'. *Acta Acust.* 7, p. 12. DOI: 10.1051/aacus/2023006.

Barumerli, R., Majdak, P., Reijniers, J., Baumgartner, R. and Geronazzo, M. (2020). 'Predicting Directional Sound-Localization of Human Listeners in Both Horizontal and Vertical Dimensions'. In: *Proc. Audio Eng. Soc. (AES) Conv.* Online, p. 10360.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). 'Fitting Linear Mixed-Effects Models Using Lme4'. *J. Stat. Soft.* 67.1. DOI: 10.18637/jss.v067.i01.

Baumgartner, R. and Majdak, P. (2021). 'Decision Making in Auditory Externalization Perception: Model Predictions for Static Conditions'. *Acta Acust.* 5, p. 59. DOI: 10.1051/aacus/2021053.

Baumgartner, R., Majdak, P. and Laback, B. (2013). 'Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications'. In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer, pp. 93–119. DOI: 10.1007/978-3-642-37762-4_4.

Baumgartner, R., Majdak, P. and Laback, B. (2014). 'Modeling Sound-Source Localization in Sagittal Planes for Human Listeners'. *J. Acoust. Soc. Am.* 136.2, pp. 791–802. DOI: 10.1121/1.4887447.

Begault, D. R. (2001). 'Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source'. *J. Audio Eng. Soc. (AES)* 49.10, p. 13.

Berebi, O., Ben-Hur, Z., Alon, D. L. and Rafaely, B. (2025). 'BSM-iMagLS: ILD Informed Binaural Signal Matching for Reproduction With Head-Mounted Microphone Arrays'. *IEEE Trans. Audio, Speech, Language Process.* 33, pp. 2705–2718. DOI: 10.1109/TASLPRO.2025.3581096.

Berger, C. C., Gonzalez-Franco, M., Tajadura-Jiménez, A., Florencio, D. and Zhang, Z. (2018). 'Generic HRTFs May Be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity'. *Front. Neurosci.* 12, p. 21. DOI: 10.3389/fnins.2018.00021.

Bernstein, L. R. and Trahiotis, C. (1994). 'Detection of Interaural Delay in High-Frequency Sinusoidally Amplitude-Modulated Tones, Two-Tone Complexes, and Bands of Noise'. *J. Acoust. Soc. Am.* 95.6, pp. 3561–3567. DOI: 10.1121/1.409973.

Best, V., Baumgartner, R., Lavandier, M., Majdak, P. and Kopčo, N. (2020). 'Sound Externalization: A Review of Recent Research'. *Trends Hear.* 24. DOI: 10.1177/2331216520948390.

Bhattacharyya, J., Picinali, L., Vinciarelli, A. and Brewster, S. (2024). 'Investigating the Influence of Environmental Acoustics and Playback Device for Audio Augmented Reality Applications'. In: *AES Audio for Games*. Tokyo, Japan.

Blauert, J. (1969). 'Sound Localization in the Median Plane'. *Acta Acust. united with Acust.* 22, pp. 205–213.

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. 2nd ed. The MIT Press. DOI: 10.7551/mitpress/6391.001.0001.

Blauert, J., ed. (2013). *The Technology of Binaural Listening*. 1st ed. Modern Acoustics and Signal Processing. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-37762-4.

Blauert, J. and J. Braasch, eds. (2020). *The Technology of Binaural Understanding*. Modern Acoustics and Signal Processing. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-00386-9.

Bolia, R. S., D'Angelo, W. R. and McKinley, R. L. (1999). 'Aurally Aided Visual Search in Three-Dimensional Space'. *Hum. Factors* 41.4, pp. 664–669. DOI: 10.1518/001872099779656789.

Bolia, R. S., D'Angelo, W. R., Mishler, P. J. and Morris, L. J. (2001). 'Effects of Hearing Protectors on Auditory Localization in Azimuth and Elevation'. *Hum. Factors* 43.1, pp. 122–128. DOI: 10.1518/001872001775992499.

Bolia, R. S., Nelson, W. T., Ericson, M. A. and Simpson, B. D. (2000). 'A Speech Corpus for Multitalker Communications Research'. *J. Acoust. Soc. Am.* 107.2, pp. 1065–1066. DOI: 10.1121/1.428288.

Brandenburg, K., Klein, F., Neidhardt, A., Sloma, U. and Werner, S. (2020). 'Creating Auditory Illusions with Binaural Technology'. In: *The Technology of Binaural Understanding*. Ed. by J. Blauert and J. Braasch. Cham: Springer International Publishing, pp. 623–663. DOI: 10.1007/978-3-030-00386-9_21.

Braren, H. S. and Fels, J. (2020a). *A High-Resolution Head-Related Transfer Function Data Set and 3D-scan of KEMAR*. RWTH Aachen University. DOI: 10.18154/RWTH-2020-11307.

Braren, H. S. and Fels, J. (2020b). *A High-Resolution Individual 3D Adult Head and Torso Model for HRTF Simulation and Validation: HRTF Measurement*. RWTH Aachen University. DOI: 10.18154/RWTH-2020-06761.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press. DOI: 10.7551/mitpress/1486.001.0001.

Brimijoin, W. O., Hassager, H. G., Ithapu, V. K. and Robinson, P. (2021). 'Individualization of Head Related Transfer Functions for Presentation of Audio Content'. U.S. pat. 10,932,083 B2. Meta Platforms Technologies LLC.

Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M. and Weinzierl, S. (2019a). 'A Round Robin on Room Acoustical Simulation and Auralization'. *J. Acoust. Soc. Am.* 145.4, pp. 2746–2760. DOI: 10.1121/1.5096178.

Brinkmann, F., Dinakaran, M., Pelzer, R., Grosche, P., Voss, D. and Weinzierl, S. (2019b). 'A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses'. *J. Audio Eng. Soc. (AES)* 67.9, pp. 705–718. DOI: 10.17743/jaes.2019.0024.

Brinkmann, F., Lindau, A. and Weinzierl, S. (2017). 'On the Authenticity of Individual Dynamic Binaural Synthesis'. *J. Acoust. Soc. Am.* 142.4, pp. 1784–1795. DOI: 10.1121/1.5005606.

Brinkmann, F. and Weinzierl, S. (2023). 'Audio Quality Assessment for Virtual Reality'. In: *Sonic Interactions in Virtual Environments*. Ed. by M. Geronazzo and S. Serafin. Cham: Springer, pp. 145–178. DOI: 10.1007/978-3-031-04021-4_5.

Brinkmann, F. et al. (2023). 'Recent Advances in an Open Software for Numerical HRTF Calculation'. *J. Audio Eng. Soc. (AES)* 71.7/8, pp. 502–514. DOI: 10.17743/jaes.2022.0078.

Bronkhorst, A. W. and Plomp, R. (1988). 'The Effect of Head-Induced Interaural Time and Level Differences on Speech Intelligibility in Noise'. *J. Acoust. Soc. Am.* 83.4, pp. 1508–1516. DOI: 10.1121/1.395906.

Bronkhorst, A. W. (2015). 'The Cocktail-Party Problem Revisited: Early Processing and Selection of Multi-Talker Speech'. *Atten. Percept. Psychophys.* 77.5, pp. 1465–1487. DOI: 10.3758/s13414-015-0882-9.

Brughera, A., Dunai, L. and Hartmann, W. M. (2013). 'Human Interaural Time Difference Thresholds for Sine Tones: The High-Frequency Limit'. *J. Acoust. Soc. Am.* 133.5, pp. 2839–2855. DOI: 10.1121/1.4795778.

Brungart, D. S. (2001a). 'Evaluation of Speech Intelligibility with the Coordinate Response Measure'. *J. Acoust. Soc. Am.* 109.5, pp. 2276–2279. DOI: 10.1121/1.1357812.

Brungart, D. S. (2001b). 'Informational and Energetic Masking Effects in the Perception of Two Simultaneous Talkers'. *J. Acoust. Soc. Am.* 109.3, pp. 1101–1109. DOI: 10.1121/1.1345696.

Brungart, D. S., Durlach, N. I. and Rabinowitz, W. M. (1999). 'Auditory Localization of Nearby Sources. II. Localization of a Broadband Source'. *J. Acoust. Soc. Am.* 106.4, pp. 1956–1968. DOI: 10.1121/1.427943.

Burkhard, M. D. and Sachs, R. M. (1975). 'Anthropometric Manikin for Acoustic Research'. *J. Acoust. Soc. Am.* 58.1, pp. 214–222. DOI: 10.1121/1.380648.

Callan, A., Callan, D. E. and Ando, H. (2013). 'Neural Correlates of Sound Externalization'. *NeuroImage* 66, pp. 22–27. DOI: 10.1016/j.neuroimage.2012.10.057.

Carlile, S., Balachandar, K. and Kelly, H. (2014). 'Accommodating to New Ears: The Effects of Sensory and Sensory-Motor Feedback'. *J. Acoust. Soc. Am.* 135.4, pp. 2002–2011. DOI: 10.1121/1.4868369.

Carlile, S. and Blackman, T. (2014). 'Relearning Auditory Spectral Cues for Locations inside and Outside the Visual Field'. *J. Assoc. Res. Otolaryngol.* 15.2, pp. 249–263. DOI: 10.1007/s10162-013-0429-5.

Carlile, S. and Leung, J. (2016). 'The Perception of Auditory Motion'. *Trends Hear.* 20, pp. 1–19. DOI: 10.1177/2331216516644254.

Carlini, A., Bordeau, C. and Ambard, M. (2024). 'Auditory Localization: A Comprehensive Practical Review'. *Front. Psychol.* 15, p. 1408073. DOI: 10.3389/fpsyg.2024.1408073.

Cherry, E. C. (1953). 'Some Experiments on the Recognition of Speech, with One and with Two Ears'. *J. Acoust. Soc. Am.* 25.5, pp. 975–979. DOI: 10.1121/1.1907229.

Cox, T. J. et al. (2023). 'Overview of the 2023 ICASSP SP Clarity Challenge: Speech Enhancement for Hearing Aids'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Rhodes, Greece, pp. 1–2. DOI: 10.1109/ICASSP49357.2023.10433922.

Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Reyes-Lecuona, A. and Picinali, L. (2021). 'Impact of Non-Individualised Head Related Transfer Functions on Speech-in-Noise Performances within a Synthesised Virtual Environment'. *J. Acoust. Soc. Am.* 149.4, pp. 2573–2586. DOI: 10.1121/10.0004220.

Cuevas-Rodríguez, M. et al. (2019). '3D Tune-In Toolkit: An Open-Source Library for Real-Time Binaural Spatialisation'. *PLOS ONE* 14.3, e0211899. DOI: 10.1371/journal.pone.0211899.

Culling, J. F., Lavandier, M. and Jelfs, S. (2013). 'Predicting Binaural Speech Intelligibility in Architectural Acoustics'. In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 427–447. DOI: 10.1007/978-3-642-37762-4_16.

Culling, J. F., Hawley, M. L. and Litovsky, R. Y. (2004). 'The Role of Head-Induced Interaural Time and Level Differences in the Speech Reception Threshold for Multiple Interfering Sound Sources'. *J. Acoust. Soc. Am.* 116.2, pp. 1057–1065. DOI: 10.1121/1.1772396.

Culling, J. F. and Stone, M. A. (2017). 'Energetic Masking and Masking Release'. In: *The Auditory System at the Cocktail Party*. Ed. by J. C. Middlebrooks, J. Z. Simon, A. N. Popper and R. R. Fay. Vol. 60. Cham: Springer International Publishing, pp. 41–73. DOI: 10.1007/978-3-319-51662-2_3.

Dabike, G. R. et al. (2024). *The First Cadenza Challenges: Using Machine Learning Competitions to Improve Music for Listeners with a Hearing Loss*. arXiv: 2409.05095.

Daugintis, R., Alary, B., Geronazzo, M. and Picinali, L. (2024). 'Effects of Binaural Rendering Personalisation and Reverberation on Speech-on-Speech Masking'. In: *Proc. Audio Eng. Soc. (AES) Conf. Audio Virtual & Augmented Reality*. Redmond, WA, USA. URL: https://aes2.org/publications/elibrary-page/?id=22671.

Daugintis, R., Barumerli, R., Geronazzo, M., Pauwels, J., Picinali, L. and Poole, K. C. (2025a). 'Listener Acoustic Personalisation Challenge—LAP24: Head-Related Transfer Function Dataset Harmonization'. *IEEE Open J. Signal Process.* 6, pp. 950–964. DOI: 10.1109/OJSP.2025.3592601.

Daugintis, R., Barumerli, R., Geronazzo, M. and Picinali, L. (2023a). 'Initial Evaluation of an Auditory-Model-Aided Selection Procedure for Non-Individual HRTFs'. In: *Proc. Conv. EAA Forum Acusticum*. Turin, Italy. DOI: 10.61782/fa.2023.0489.

Daugintis, R., Barumerli, R., Picinali, L. and Geronazzo, M. (2023b). 'Classifying Non-Individual Head-Related Transfer Functions with a Computational Auditory Model: Calibration and Metrics'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Rhodes, Greece. DOI: 10.1109/ICASSP49357.2023.10095152.

Daugintis, R., Geronazzo, M. and Picinali, L. (2025b). 'On Comparing Auditory Models and Perceptual Assessment When Rating Head-Related Transfer Functions'. In: *Proc. Conv. EAA Forum Acusticum EuroNoise*. Málaga, Spain, pp. 3421–3424. DOI: 10.61782/fa.2025.0643.

Daugintis, R., Geronazzo, M., Poole, K. C. and Picinali, L. (2026). 'Perceptual evaluation of an auditory model–based similarity metric for head-related transfer functions'. Under review.

Dicke, C., Aaltonen, V., Rämö, A. and Vilermo, M. (2010). 'Talk to Me: The Influence of Audio Quality on the Perception of Social Presence'. In: *Proc. HCI*. DOI: 10.14236/ewic/HCI2010.36.

Dietrich, P., Masiero, B. and Vorländer, M. (2013). 'On the optimization of the multiple exponential sweep method'. *J. Audio Eng. Soc. (AES)* 61.3, pp. 113–124.

Dietz, M., Ewert, S. D. and Hohmann, V. (2011). 'Auditory Model Based Direction Estimation of Concurrent Speakers from Binaural Signals'. *Speech Commun.* 53.5, pp. 592–605. DOI: 10.1016/j.specom.2010.05.006.

Dinakaran, M. et al. (2018). 'Perceptually Motivated Analysis of Numerically Simulated Head-Related Transfer Functions Generated by Various 3D Surface Scanning Systems'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 551–555. DOI: 10.1109/ICASSP.2018.8461789.

Djelani, T., Pörschmann, C., Sahrhage, J. and Blauert, J. (2000). 'An Interactive Virtual-Environment Generator for Psychoacoustic Research II: Collection of Head-Related Impulse Responses and Evaluation of Auditory Localization'. *Acta Acust. united with Acust.* 86, pp. 1046–1053.

Doma, S., Brožová, N. and Fels, J. (2023a). 'Examining the Interrelation Behavior of Distance Metrics for Head-Related Transfer Function Evaluation: A Case Study'. *Acta Acust.* 7, p. 34. DOI: 10.1051/aacus/2023028.

Doma, S., Ermert, C. A. and Fels, J. (2023b). 'A Magnitude-Based Parametric Model Predicting the Audibility of HRTF Variation'. *J. Audio Eng. Soc. (AES)* 71.4, pp. 155–172. DOI: 10.17743/jaes.2022.0080.

Ege, R., Opstal, A. J. V. and Van Wanrooij, M. M. (2018). 'Accuracy-Precision Trade-off in Human Sound Localisation'. *Sci. Rep.* 8.16399. DOI: 10.1038/s41598-018-34512-6.

Engel, I., Henry, C., Amengual Garí, S. V., Robinson, P. W. and Picinali, L. (2021). 'Perceptual Implications of Different Ambisonics-based Methods for Binaural Reverberation'. *J. Acoust. Soc. Am.* 149.2, pp. 895–910. DOI: 10.1121/10.0003437.

Engel, I., Henry, C., Garı, S. V. A., Robinson, P., Poirier-Quinot, D. and Picinali, L. (2019). 'Perceptual Comparison of Ambisonics-based Reverberation Methods in Binaural Listening'. In: Paris, France.

Engel, I. et al. (2023). 'The SONICOM HRTF Dataset'. *J. Audio Eng. Soc. (AES)* 71.5, pp. 241–253. DOI: 10.17743/jaes.2022.0066.

Enzner, G. (2009). '3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering'. In: *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*, pp. 325–328. DOI: 10.1109/ASPAA.2009.5346532.

Fantini, D., Avanzini, F., Ntalampiras, S. and Presti, G. (2021). 'HRTF Individualization Based on Anthropometric Measurements Extracted from 3D Head Meshes'. In: *Proc. Immersive and 3D Audio (I3DA)*. Bologna, Italy: IEEE, pp. 1–10. DOI: 10.1109/I3DA48870.2021.9610904.

Fantini, D., Geronazzo, M., Avanzini, F. and Ntalampiras, S. (2025). 'A Survey on Machine Learning Techniques for Head-Related Transfer Function Individualization'. *IEEE Open J. Signal Process.* 6, pp. 30–56. DOI: 10.1109/OJSP.2025.3528330.

Farahikia, M. and Su, Q. T. (2017). 'Optimized Finite Element Method for Acoustic Scattering Analysis With Application to Head-Related Transfer Function Estimation'. *J. Vib. Acoust.* 139.3, p. 034501. DOI: 10.1115/1.4035813.

Farina, A. (2007). 'Advancements in Impulse Response Measurements by Sine Sweeps'. In: *Proc. Audio Eng. Soc. (AES) Conv.* Vienna, Austria, p. 7121.

Folkerts, M. L. and Stecker, G. C. (2022). 'Spectral Weighting Functions for Lateralization and Localization of Complex Sound'. *J. Acoust. Soc. Am.* 151.5, pp. 3409–3425. DOI: 10.1121/10.0011469.

Freeman, G., Zamanifard, S., Maloney, D. and Acena, D. (2022). 'Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality'. *Proc. ACM Hum.-Comput. Interact.* 6 (CSCW1), pp. 1–30. DOI: 10.1145/3512932.

Freeman, G. et al. (2024). 'Understanding and Mitigating New Harms in Immersive and Embodied Virtual Spaces: A Speculative Dystopian Design Fiction Approach'. In: *Compan. Publ. Conf. Comput.-Support. Coop. Work Soc. Comput. (CSCW)*. San Jose, Costa Rica: ACM, pp. 288–295. DOI: 10.1145/3678884.3681865.

Freyman, R. L., Helfer, K. S., McCall, D. D. and Clifton, R. K. (1999). 'The Role of Perceived Spatial Separation in the Unmasking of Speech'. *J. Acoust. Soc. Am.* 106.6, pp. 3578–3588. DOI: 10.1121/1.428211.

Friedrich, P. and Schönwiesner, M. (2025). 'Adaptation Rate and Persistence across Multiple Sets of Spectral Cues for Sound Localization'. *J. Acoust. Soc. Am.* 157.3, pp. 1543–1553. DOI: 10.1121/10.0036056.

Gamper, H. (2013). 'Head-Related Transfer Function Interpolation in Azimuth, Elevation, and Distance'. *J. Acoust. Soc. Am.* 134.6, EL547–EL553. DOI: 10.1121/1.4828983.

Gardner, W. G. and Martin, K. D. (1995). 'HRTF Measurements of a KEMAR'. *J. Acoust. Soc. Am.* 97.6, pp. 3907–3908. DOI: 10.1121/1.412407.

Gari, S. V. A., Schissler, C. and Robinson, P. (2024). 'Perceptual Comparison of Efficient Real-Time Geometrical Acoustics Engines in Virtual Reality'. In: *AES Audio for Games*. Tokyo, Japan.

Geronazzo, M. (2025). 'Strong and Weak Head-Related Transfer Functions: The eHRTF Analytical Framework'. *JASA Express lett. (JASA-EL)* 5.8, p. 087202. DOI: 10.1121/10.0038961.

Geronazzo, M., Peruch, E., Prandoni, F. and Avanzini, F. (2019). 'Applying a Single-Notch Metric to Image-Guided Head-Related Transfer Function Selection for Improved Vertical Localization'. *J. Audio Eng. Soc. (AES)* 67.6, pp. 414–428. DOI: 10.17743/jaes.2019.0010.

Geronazzo, M., Spagnol, S. and Avanzini, F. (2018). 'Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric'. *IEEE/ACM Trans. Audio, Speech, Language Process.* 26.7, pp. 1247–1260. DOI: 10.1109/TASLP.2018.2821846.

Geronazzo, M., Spagnol, S., Bedin, A. and Avanzini, F. (2014). 'Enhancing Vertical Localization with Image-Guided Selection of Non-Individual Head-Related Transfer Functions'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4463–4467. DOI: 10.1109/ICASSP.2014.6854446.

Geronazzo, M., Tissieres, J. Y. and Serafin, S. (2020). 'A Minimal Personalization of Dynamic Binaural Synthesis with Mixed Structural Modeling and Scattering Delay Networks'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 411–415. DOI: 10.1109/ICASSP40776.2020.9053873.

Geronazzo, M. et al. (2024). *Technical Report: SONICOM / IEEE Listener Acoustic Personalisation (LAP) Challenge - 2024*. DOI: 10.36227/techrxiv.173153187.72930965/v1.

Ghorbal, S., Bonjour, X. and Séguier, R. (2020). 'Computed Hrirs and Ears Database for Acoustic Research'. In: *Proc. Audio Eng. Soc. (AES) Conv.* Online. URL: https://aes2.org/publications/elibrary-page/?id=20778.

Gilkey, R. H., Good, M. D., Ericson, M. A., Brinkman, J. and Stewart, J. M. (1995). 'A Pointing Technique for Rapidly Collecting Localization Responses in Auditory Research'. *Behav. Res. Methods Instrum. Comput.* 27.1, pp. 1–11. DOI: 10.3758/BF03203614.

Glasberg, B. R. and Moore, B. C. (1990). 'Derivation of Auditory Filter Shapes from Notched-Noise Data'. *Hear. Res.* 47.1–2, pp. 103–138. DOI: 10.1016/0378-5955(90)90170-T.

Goertzel, B. (2014). 'Artificial General Intelligence: Concept, State of the Art, and Future Prospects'. *J. Artif. Gen. Intell.* 5.1, pp. 1–48. DOI: 10.2478/jagi-2014-0001.

González-Toledo, D., Cuevas-Rodríguez, M., Vicente, T., Picinali, L., Molina-Tanco, L. and Reyes-Lecuona, A. (2024). 'Spatial Release from Masking in the Median Plane with Non-Native Speakers Using Individual and Mannequin Head Related Transfer Functions'. *J. Acoust. Soc. Am.* 155.1, pp. 284–293. DOI: 10.1121/10.0024239.

Graetzer, S. et al. (2020). *Clarity: Machine Learning Challenges to Revolutionise Hearing Device Processing*. DOI: 10.48550/arXiv.2006.11140. arXiv: 2006.11140.

Grey, J. M. and Gordon, J. W. (1978). 'Perceptual Effects of Spectral Modifications on Musical Timbres'. *J. Acoust. Soc. Am.* 63.5, pp. 1493–1500. DOI: 10.1121/1.381843.

Grijalva, F., Martini, L., Florencio, D. and Goldenstein, S. (2016). 'A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features'. *IEEE/ACM Trans. Audio, Speech, Language Process.* 24.3, pp. 559–570. DOI: 10.1109/TASLP.2016.2517565.

Guezenoc, C. and Seguier, R. (2018). 'HRTF Individualization: A Survey'. In: *Proc. Audio Eng. Soc. (AES) Conv.* New York, p. 10129.

Guezenoc, C. and Séguier, R. (2020). 'A Wide Dataset of Ear Shapes and Pinna-Related Transfer Functions Generated by Random Ear Drawings'. *J. Acoust. Soc. Am.* 147.6, pp. 4087–4096. DOI: 10.1121/10.0001461.

Guiraud, P., Sum, K., Pontoppidan, N., Poole, K. and Picinali, L. (2024). 'Adaptation to Altered Interaural Time Differences in a Virtual Reality Environment'. In: *Proc. Conv. EAA Forum Acusticum*. Turin, Italy, pp. 1773–1780. DOI: 10.61782/fa.2023.0412.

Halekoh, U. and Højsgaard, S. (2014). 'A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package Pbkrtest'. *J. Stat. Soft.* 59.9. DOI: 10.18637/jss.v059.i09.

Heller, L. M., Elizalde, B., Raj, B. and Deshmukh, S. (2023). *Synergy between Human and Machine Approaches to Sound/Scene Recognition and Processing: An Overview of ICASSP Special Session*. arXiv: 2302.09719.

Hofman, P. and Van Opstal, A. (2003). 'Binaural Weighting of Pinna Cues in Human Sound Localization'. *Exp. Brain Res.* 148.4, pp. 458–470. DOI: 10.1007/s00221-002-1320-5.

Hofman, P. M. and Van Opstal, A. J. (1998). 'Spectro-Temporal Factors in Two-Dimensional Human Sound Localization'. *J. Acoust. Soc. Am.* 103.5, pp. 2634–2648. DOI: 10.1121/1.422784.

Hofman, P. M., Van Riswick, J. G. and Van Opstal, A. J. (1998). 'Relearning Sound Localization with New Ears'. *Nat. Neurosci.* 1.5, pp. 417–421. DOI: 10.1038/1633.

Hogg, A., Liu, H., Jenkins, M. and Picinali, L. (2023). 'Exploring the Impact of Transfer Learning on GAN-based HRTF Upsampling'. In: *Proc. Conv. EAA Forum Acusticum*. Turin, Italy, pp. 2323–2328. DOI: 10.61782/fa.2023.0266.

Hogg, A. O. T., Jenkins, M., Liu, H., Squires, I., Cooper, S. J. and Picinali, L. (2024). 'HRTF Upsampling with a Generative Adversarial Network Using a Gnomonic Equiangular Projection'. *IEEE/ACM Trans. Audio, Speech, Language Process.*, pp. 2085–2099. DOI: 10.1109/TASLP.2024.3375635.

Hogg, A. O. T. et al. (2025). 'Listener Acoustic Personalisation Challenge - LAP24: Head-Related Transfer Function Upsampling'. *IEEE Open J. Signal Process.* 6, pp. 926–941. DOI: 10.1109/OJSP.2025.3588776.

Holleman, G. A., Hooge, I. T. C., Kemner, C. and Hessels, R. S. (2020). 'The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity'. *Front. Psychol.* 11, p. 721. DOI: 10.3389/fpsyg.2020.00721.

Iida, K. and Oota, M. (2018). 'Median Plane Sound Localization Using Early Head-Related Impulse Response'. *Appl. Acoust.* 139, pp. 14–23. DOI: 10.1016/j.apacoust.2018.03.027.

Ito, Y., Nakamura, T., Koyama, S., Sakamoto, S. and Saruwatari, H. (2025). 'Spatial Upsampling of Head-Related Transfer Function Using Neural Network Conditioned on Source Position and Frequency'. *IEEE Open J. Signal Process.*, pp. 1–16. DOI: 10.1109/OJSP.2025.3613132.

Iwaya, Y. (2006). 'Individualization of Head-Related Transfer Functions with Tournament-Style Listening Test: Listening with Other's Ears'. *Acoust. Sci. Technol.* 27.6, pp. 340–343. DOI: 10.1250/ast.27.340.

Jeffress, L. A. (1948). 'A Place Theory of Sound Localization.' *J. Comp. Physiol. Psychol.* 41.1, pp. 35–39. DOI: 10.1037/h0061495.

Jelfs, S., Culling, J. F. and Lavandier, M. (2011). 'Revision and Validation of a Binaural Model for Speech Intelligibility in Noise'. *Hear. Res.* 275.1–2, pp. 96–104. DOI: 10.1016/j.heares.2010.12.005.

Jenny, C. and Reuter, C. (2020). 'Usability of Individualized Head-Related Transfer Functions in Virtual Reality: Empirical Study With Perceptual Attributes in Sagittal Plane Sound Localization'. *JMIR Serious Games* 8.3, e17576. DOI: 10.2196/17576.

Jenny, C. and Reuter, C. (2021). 'Can I Trust My Ears in VR? Head-related Transfer Functions and Valuation Methods with Descriptive Attributes in Virtual Reality'. *Int. J. Virtual Reality* 21.2, pp. 29–43. DOI: 10.20870/IJVR.2021.21.2.4831.

Jiang, J., Xie, B., Mai, H., Liu, L., Yi, K. and Zhang, C. (2019). 'The Role of Dynamic Cue in Auditory Vertical Localisation'. *Appl. Acoust.* 146, pp. 398–408. DOI: 10.1016/j.apacoust.2018.12.002.

Jin, C., Schenkel, M. and Carlile, S. (2000). 'Neural System Identification Model of Human Sound Localization'. *J. Acoust. Soc. Am.* 108.3, pp. 1215–1235. DOI: 10.1121/1.1288411.

Johansson, J., Karjalainen, A., Malinen, M., Tikkanen, J., Saari, V. and Vosough, P. (2020). 'System and Method for Generating Head -Related Transfer Function'. U.S. pat. 10,743,128 B1. Genelec Oy.

Joris, P. X., Smith, P. H. and Yin, T. C. (1998). 'Coincidence Detection in the Auditory System'. *Neuron* 21.6, pp. 1235–1238. DOI: 10.1016/S0896-6273(00)80643-1.

Joyner, M. S., Brandmeyer, A., Daly, S., Baker, J. R., Fanelli, A. and Crum, P. A. C. (2023). 'Personalized HRTFs via Optical Capture'. U.S. pat. 11,778,403 B2. Dolby Laboratories Licensing Corporation.

Katz, B. F. G. and Noisternig, M. (2014). 'A Comparative Study of Interaural Time Delay Estimation Methods'. *J. Acoust. Soc. Am.* 135.6, pp. 3530–3540. DOI: 10.1121/1.4875714.

Katz, B. F. G. and Parseihian, G. (2012). 'Perceptually Based Head-Related Transfer Function Database Optimization'. *J. Acoust. Soc. Am.* 131.2, EL99–EL105. DOI: 10.1121/1.3672641.

Kenward, M. G. and Roger, J. H. (1997). 'Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood'. *Biometrics* 53.3, p. 983. DOI: 10.2307/2533558. JSTOR: 2533558.

Kidd, G. and Colburn, H. S. (2017). 'Informational Masking in Speech Recognition'. In: *The Auditory System at the Cocktail Party*. Ed. by J. C. Middlebrooks, J. Z. Simon, A. N. Popper and R. R. Fay. Vol. 60. Cham: Springer International Publishing, pp. 75–109. DOI: 10.1007/978-3-319-51662-2_4.

Kim, C., Lim, V. and Picinali, L. (2020). 'Investigation into Consistency of Subjective and Objective Perceptual Selection of Non-Individual Head-Related Transfer Functions'. *J. Audio Eng. Soc. (AES)* 68.11, pp. 819–831. DOI: 10.17743/jaes.2020.0053.

Klingel, M., Singhal, U., Seitz, A. R. and Kopčo, N. (2025). 'Binaural-Cue Reweighting Induced by Discrimination Training'. *Atten. Percept. Psychophys.* 87.5, pp. 1788–1800. DOI: 10.3758/s13414-025-03082-x.

Klockgether, S. and Van De Par, S. (2016). 'Just Noticeable Differences of Spatial Cues in Echoic and Anechoic Acoustical Environments'. *J. Acoust. Soc. Am.* 140.4, EL352–EL357. DOI: 10.1121/1.4964844.

Kondo, K., Chiba, T., Kitashima, Y. and Yano, N. (2010). 'Intelligibility Comparison of Japanese Speech with Competing Noise Spatialized in Real and Virtual Acoustic Environments'. *Acoust. Sci. Technol.* 31.3, pp. 231–238. DOI: 10.1250/ast.31.231.

Koo, T. K. and Li, M. Y. (2016). 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research'. *J. Chiropr. Med.* 15.2, pp. 155–163. DOI: 10.1016/j.jcm.2016.02.012.

Kukshinov, E. and Nacke, L. E. (2025). 'Collective Embodiment, or the Social Nature of the Sense of Embodiment in Social VR'. In: *Proc. ACM Int. Conf. Interact. Media Exp. (IMX)*. Niterói, Brazil, pp. 187–199. DOI: 10.1145/3706370.3727895.

Kulkarni, A. and Colburn, H. S. (1998). 'Role of Spectral Detail in Sound-Source Localization'. *Nature* 396.6713, pp. 747–749. DOI: 10.1038/25526.

Kuttruff, H. (2016). *Room Acoustics*. 6th ed. Taylor & Francis Group.

Kutz, J. N. et al. (2025). *Accelerating Scientific Discovery with the Common Task Framework*. DOI: 10.48550/arXiv.2511.04001. arXiv: 2511.04001 [cs].

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B. (2017). 'lmerTest Package: Tests in Linear Mixed Effects Models'. *J. Stat. Soft.* 82.13. DOI: 10.18637/jss.v082.i13.

Langendijk, E. H. A. and Bronkhorst, A. W. (1999). 'Fidelity of three-dimensional-sound reproduction using a virtual auditory display'. *J. Acoust. Soc. Am.* 107.1, pp. 528–537. DOI: 10.1121/1.428321.

Langendijk, E. H. A. and Bronkhorst, A. W. (2002). 'Contribution of Spectral Cues to Human Sound Localization'. *J. Acoust. Soc. Am.* 112.4, pp. 1583–1596. DOI: 10.1121/1.1501901.

Lavandier, M. and Best, V. (2020). 'Modeling Binaural Speech Understanding in Complex Situations'. In: *The Technology of Binaural Understanding*. Ed. by J. Blauert and J. Braasch. Cham: Springer International Publishing, pp. 547–578. DOI: 10.1007/978-3-030-00386-9_19.

Lavandier, M. and Culling, J. F. (2008). 'Speech Segregation in Rooms: Monaural, Binaural, and Interacting Effects of Reverberation on Target and Interferer'. *J. Acoust. Soc. Am.* 123.4, pp. 2237–2248. DOI: 10.1121/1.2871943.

Lavandier, M., Vicente, T. and Prud'homme, L. (2022). 'A Series of SNR-based Speech Intelligibility Models in the Auditory Modeling Toolbox'. *Acta Acust.* 6, p. 20. DOI: 10.1051/aacus/2022017.

Leclère, T., Lavandier, M. and Culling, J. F. (2015). 'Speech Intelligibility Prediction in Reverberation: Towards an Integrated Model of Speech Transmission, Spatial Unmasking, and Binaural de-Reverberation'. *J. Acoust. Soc. Am.* 137.6, pp. 3335–3345. DOI: 10.1121/1.4921028.

Li, S. and Peissig, J. (2020). 'Measurement of Head-Related Transfer Functions: A Review'. *Appl. Sci.* 10.14, p. 5014. DOI: 10.3390/app10145014.

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F. and Weinzierl, S. (2014). 'A Spatial Audio Quality Inventory (SAQI)'. *Acta Acust. united with Acust.* 100.5, pp. 984–994. DOI: 10.3813/AAA.918778.

Lindau, A., Kosanke, L. and Weinzierl, S. (2012). 'Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses'. *J. Audio Eng. Soc. (AES)* 60.11, pp. 887–898.

Lindau, A. and Weinzierl, S. (2012). 'Assessing the Plausibility of Virtual Acoustic Environments'. *Acta Acust. united with Acust.* 98.5, pp. 804–810. DOI: 10.3813/AAA.918562.

Lladó, P., Hyvärinen, P. and Pulkki, V. (2024a). 'The Impact of Head-Worn Devices in an Auditory-Aided Visual Search Task'. *J. Acoust. Soc. Am.* 155.4, pp. 2460–2469. DOI: 10.1121/10.0025542.

Lladó, P., Majdak, P., Barumerli, R. and Baumgartner, R. (2025). 'Spectral Weighting of Monaural Cues for Auditory Localization in Sagittal Planes'. *Trends Hear.* 29. DOI: 10.1177/23312165251317027.

Lladó, P., Pollack, K. and Meyer-Kahlen, N. (2024b). 'Toward a Standard Listener-Independent HRTF to Facilitate Long-Term Adaptation'. *J. Audio Eng. Soc. (AES)* 72.4, pp. 188–192. DOI: 10.17743/jaes.2022.0134.

Lokki, T. and Pätynen, J. (2020). 'Auditory Spatial Impression in Concert Halls'. In: *The Technology of Binaural Understanding*. Ed. by J. Blauert and J. Braasch. Cham: Springer International Publishing, pp. 173–202. DOI: 10.1007/978-3-030-00386-9_7.

Lu, X., Wang, Y., Sang, J. and Zheng, C. (2025). 'BiCG: Binaural Cue Generation from Unified HRTF Datasets'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Hyderabad, India. DOI: 10.1109/ICASSP49660.2025.10890756.

Lübeck, T., Weber, T. and Pörschmann, C. (2024). 'A Perceptual Study on the Plausibility of Binaural Rendering in Anechoic Environments'. *J. Audio Eng. Soc. (AES)* 72.12, pp. 846–859. DOI: 10.17743/jaes.2022.0180.

Luo, Y., Zotkin, D. N. and Duraiswami, R. (2013). 'Gaussian Process Data Fusion for Heterogeneous HRTF Datasets'. In: *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*. New Paltz, NY, USA. DOI: 10.1109/WASPAA.2013.6701842.

MacArthur, C., Kukshinov, E., Harley, D., Pawar, T., Modi, N. and Nacke, L. E. (2024). 'Experiential Disparities in Social VR: Uncovering Power Dynamics and Inequality'. *Front. Virtual Real.* 5, p. 1351794. DOI: 10.3389/frvir.2024.1351794.

Macé, M. J.-M., Dramas, F. and Jouffrais, C. (2012). 'Reaching to Sound Accuracy in the Peri-Personal Space of Blind and Sighted Humans'. In: *Computers Helping People with Special Needs, ICCHP 2012*. Vol. 7383, pp. 636–643. DOI: 10.1007/978-3-642-31534-3_93.

Macpherson, E. A. and Middlebrooks, J. C. (2000). 'Localization of Brief Sounds: Effects of Level and Background Noise'. *J. Acoust. Soc. Am.* 108.4, pp. 1834–1849. DOI: 10.1121/1.1310196.

Macpherson, E. A. and Middlebrooks, J. C. (2002). 'Listener Weighting of Cues for Lateral Angle: The Duplex Theory of Sound Localization Revisited'. *J. Acoust. Soc. Am.* 111.5, pp. 2219–2236. DOI: 10.1121/1.1471898.

Macpherson, E. A. and Sabin, A. T. (2007). 'Binaural Weighting of Monaural Spectral Cues for Sound Localization'. *J. Acoust. Soc. Am.* 121.6, pp. 3677–3688. DOI: 10.1121/1.2722048.

Macpherson, E. A. and Sabin, A. T. (2013). 'Vertical-Plane Sound Localization with Distorted Spectral Cues'. *Hear. Res.* 306, pp. 76–92. DOI: 10.1016/j.heares.2013.09.007.

Majdak, P. (2007). 'Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions'. *J. Audio Eng. Soc. (AES)* 55.7.

Majdak, P., Baumgartner, R. and Jenny, C. (2020). 'Formation of Three-Dimensional Auditory Space'. In: *The Technology of Binaural Understanding*. Ed. by J. Blauert and J. Braasch. Cham: Springer International Publishing, pp. 115–149. DOI: 10.1007/978-3-030-00386-9_5.

Majdak, P., Baumgartner, R. and Laback, B. (2014). 'Acoustic and Non-Acoustic Factors in Modeling Listener-Specific Performance of Sagittal-Plane Sound Localization'. *Front. Psychol.* 5. DOI: 10.3389/fpsyg.2014.00319.

Majdak, P., Goupell, M. J. and Laback, B. (2010). '3-D Localization of Virtual Sound Sources: Effects of Visual Environment, Pointing Method, and Training'. *Atten. Percept. Psychophys.* 72.2, pp. 454–469. DOI: 10.3758/APP.72.2.454.

Majdak, P., Hollomey, C. and Baumgartner, R. (2022). 'AMT 1.x: A Toolbox for Reproducible Research in Auditory Modeling'. *Acta Acust.* 6.19. DOI: 10.1051/aacus/2022011.

Majdak, P., Walder, T. and Laback, B. (2013). 'Effect of Long-Term Training on Sound Localization Performance with Spectrally Warped and Band-Limited Head-Related Transfer Functions'. *J. Acoust. Soc. Am.* 134.3, pp. 2148–2159. DOI: 10.1121/1.4816543.

Makous, J. C. and Middlebrooks, J. C. (1990). 'Two-Dimensional Sound Localization by Human Listeners'. *J. Acoust. Soc. Am.* 87.5, pp. 2188–2200. DOI: 10.1121/1.399186.

Marggraf-Turley, N., Lovedee-Turner, M. and De Sena, E. (2024). 'HRTF Recommendation Based on the Predicted Binaural Colouration Model'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Seoul, Korea, Republic of, pp. 1106–1110. DOI: 10.1109/ICASSP48485.2024.10448092.

Marggraf-Turley, N., Pontoppidan, N. H. and Picinali, L. (2025a). 'The Impact of Individual Head-Related Transfer Function Augmentation on Spatial Release from Masking'. *Hear. Res.* 466, p. 109402. DOI: 10.1016/j.heares.2025.109402.

Marggraf-Turley, N., Shiell, M. M., Pontoppidan, N. H., Cappotto, D. and Picinali, L. (2025b). 'Electroencephalographic Decoding of Sound Location: Comparing Free-Field to Headphone-Based Non-Individual Head-Related Transfer Functions'. *J. Acoust. Soc. Am.* 158.2, pp. 859–870. DOI: 10.1121/10.0038635.

Martin, R. L., McAnally, K. I., Bolia, R. S., Eberle, G. and Brungart, D. S. (2012). 'Spatial Release from Speech-on-Speech Masking in the Median Sagittal Plane'. *J. Acoust. Soc. Am.* 131.1, pp. 378–385. DOI: 10.1121/1.3669994.

Martin, V., Engel, I. and Picinali, L. (2025a). 'Effects of Geometrical Acoustics Model Simplification on Binaural Reverberation'. *IEEE Trans. Audio, Speech, Language Process.* 33, pp. 3480–3493. DOI: 10.1109/TASLPRO.2025.3597463.

Martin, V., Meyer, J. and Picinali, L. (2025b). 'Exploring Dynamic Binaural Rendering Approaches in 6DoF Acoustic Augmented Reality'. In: *Proc. Annu. Meet. Acoust. DAS/DAGA*. Copenhagen, Denmark.

Massé, P., Carpentier, T., Warusfel, O. and Noisternig, M. (2020). 'Denoising Directional Room Impulse Responses with Spatially Anisotropic Late Reverberation Tails'. *Appl. Sci.* 10.3, p. 1033. DOI: 10.3390/app10031033.

Masuyama, Y., Wichern, G., Germain, F. G., Ick, C. and Roux, J. L. (2025). 'Retrieval-Augmented Neural Field for HRTF Upsampling and Personalization'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Hyderabad, India, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10889481.

Mate, S. S., Eronen, A. J., Lehtiniemi, A. J. and Laaksonen, L. J. (2025). 'Privacy Protection in Spatial Audio Capture'. U.S. pat. 12279090B2. Nokia Technologies Oy.

McGraw, K. O. and Wong, S. P. (1996). 'Forming Inferences About Some Intraclass Correlation Coefficients'. *Psychol. Methods* 1.1, pp. 30–46. DOI: 10.1037/1082-989X.1.1.30.

McKenzie, T., Armstrong, C., Ward, L., Murphy, D. T. and Kearney, G. (2022). 'Predicting the Colouration between Binaural Signals'. *Appl. Sci.* 12.5, p. 2441. DOI: 10.3390/app12052441.

McKenzie, T. and Brinkmann, F. (2025). 'Toward an Improved Auditory Model for Predicting Binaural Coloration'. *J. Audio Eng. Soc. (AES)* 73.3, pp. 115–126. DOI: 10.17743/jaes.2022.0192.

McLachlan, G., Majdak, P., Reijniers, J., Mihocic, M. and Peremans, H. (2023). 'Dynamic Spectral Cues Do Not Affect Human Sound Localization during Small Head Movements'. *Front. Neurosci.* 17, p. 1027827. DOI: 10.3389/fnins.2023.1027827.

McLachlan, G., Majdak, P., Reijniers, J., Mihocic, M. and Peremans, H. (2025). 'Bayesian Active Sound Localisation: To What Extent Do Humans Perform like an Ideal-Observer?' *PLoS Comput Biol* 21.1. Ed. by A. E. Martin, e1012108. DOI: 10.1371/journal.pcbi.1012108.

McLachlan, G., Majdak, P., Reijniers, J. and Peremans, H. (2021). 'Towards Modelling Active Sound Localisation Based on Bayesian Inference in a Static Environment'. *Acta Acust.* 5, p. 45. DOI: 10.1051/aacus/2021039.

McMullen, K., Roginska, A. and Wakefield, G. H. (2012). 'Subjective Selection of Head-Related Transfer Functions (HRTFs) Based on Spectral Coloration and Interaural Time Differences (ITD) Cues'. In: *Proc. Audio Eng. Soc. (AES) Conv.* San Francisco, CA, USA.

Meddis, R., E. Lopez-Poveda, R. R. Fay and A. N. Popper, eds. (2010). *Computational Models of the Auditory System*. 1st ed. Springer Handbook of Auditory Research 35. New York, NY, USA: Springer. URL: https://ebookcentral.proquest.com/lib/imperial/detail.action?docID=645350.

Mendonça, C. (2014). 'A Review on Auditory Space Adaptations to Altered Head-Related Cues'. *Front. Neurosci.* 8. DOI: 10.3389/fnins.2014.00219.

Mendonça, C., Campos, G., Dias, P. and Santos, J. A. (2013). 'Learning Auditory Space: Generalization and Long-Term Effects'. *PLOS ONE* 8.10. Ed. by M. S. Malmierca, e77900. DOI: 10.1371/journal.pone.0077900.

Mendonça, C., Campos, G., Dias, P., Vieira, J., Ferreira, J. P. and Santos, J. A. (2012). 'On the Improvement of Localization Accuracy with Non- Individualized HRTF-based Sounds'. *J. Audio Eng. Soc. (AES)* 60.10, pp. 821–830.

Mercugliano, A., Corbani, A., Bigozzi, L., Vettori, G. and Incognito, O. (2025). 'The Effects of Classroom Acoustic Quality on Student Perception and Wellbeing: A Systematic Review across Educational Levels'. *Front. Psychol.* 16, p. 1586997. DOI: 10.3389/fpsyg.2025.1586997.

Meteyard, L. and Davies, R. A. (2020). 'Best Practice Guidance for Linear Mixed-Effects Models in Psychological Science'. *J. Mem. Lang.* 112, p. 104092. DOI: 10.1016/j.jml.2020.104092.

Meyer, J. and Picinali, L. (2025a). 'Observing long-term adaptation and generalization to head-related transfer functions'. Under review.

Meyer, J. and Picinali, L. (2025b). 'On the Generalization of Accommodation to Head-Related Transfer Functions'. *J. Acoust. Soc. Am.* 157.1, pp. 420–432. DOI: 10.1121/10.0034858.

Meyer-Kahlen, N., Pollack, K. and Llado, P. (2025). 'Comparing Measures of Information in Head-Related Transfer-Functions'. In: *Proc. Conv. EAA Forum Acusticum*. Málaga, Spain.

Meyer-Kahlen, N., Schlecht, S. J. and Lokki, T. (2022). 'Clearly Audible Room Acoustical Differences May Not Reveal Where You Are in a Room'. *J. Acoust. Soc. Am.* 152.2, pp. 877–887. DOI: 10.1121/10.0013364.

Meyer-Kahlen, N., Schlecht, S. J., Amengual Garí, S. and Lokki, T. (2024). 'Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility, and Authenticity'. *J. Audio Eng. Soc. (AES)* 72.11, pp. 797–812. DOI: 10.17743/jaes.2022.0178.

Middlebrooks, J. C. (1992). 'Narrow-Band Sound Localization Related to External Ear Acoustics'. *J. Acoust. Soc. Am.* 92.5, pp. 2607–2624. DOI: 10.1121/1.404400.

Middlebrooks, J. C. (1999a). 'Individual Differences in External-Ear Transfer Functions Reduced by Scaling in Frequency'. *J. Acoust. Soc. Am.* 106.3, pp. 1480–1492. DOI: 10.1121/1.427176.

Middlebrooks, J. C. (1999b). 'Virtual Localization Improved by Scaling Nonindividualized External-Ear Transfer Functions in Frequency'. *J. Acoust. Soc. Am.* 106.3, pp. 1493–1510. DOI: 10.1121/1.427147.

Middlebrooks, J. C. (2015). 'Sound Localization'. In: *Handbook of Clinical Neurology*. Vol. 129. Elsevier, pp. 99–116. DOI: 10.1016/B978-0-444-62630-1.00006-8.

Middlebrooks, J. C. (2021). 'A Search for a Cortical Map of Auditory Space'. *J. Neurosci.* 41.27, pp. 5772–5778. DOI: 10.1523/JNEUROSCI.0501-21.2021.

Middlebrooks, J. C. and Green, D. M. (1991). 'Sound Localization by Human Listeners'. *Annu. Rev. Psychol.* 42.1, pp. 135–159. DOI: 10.1146/annurev.ps.42.020191.001031.

Middlebrooks, J. C., Macpherson, E. A. and Onsan, Z. A. (2000). 'Psychophysical Customization of Directional Transfer Functions for Virtual Sound Localization'. *J. Acoust. Soc. Am.* 108.6, pp. 3088–3091. DOI: 10.1121/1.1322026.

Middlebrooks, J. C., Makous, J. C. and Green, D. M. (1989). 'Directional Sensitivity of Sound-Pressure Levels in the Human Ear Canal'. *J. Acoust. Soc. Am.* 86.1, pp. 89–108. DOI: 10.1121/1.398224.

Mills, A. W. (1958). 'On the Minimum Audible Angle'. *J. Acoust. Soc. Am.* 30.4, pp. 237–246. DOI: 10.1121/1.1909553.

Mills, A. W. (1960). 'Lateralization of High-Frequency Tones'. *J. Acoust. Soc. Am.* 32.1, pp. 132–134. DOI: 10.1121/1.1907864.

Minnaar, P., Plogsties, J., Olesen, S. K., Christensen, F. and Moller, H. (2000). 'The Interaural Time Difference in Binaural Synthesis'. In: *Proc. Audio Eng. Soc. (AES) Conv.* 5133. Paris, France.

Møller, H. (1992). 'Fundamentals of Binaural Technology'. *Appl. Acoust.* 36.3, pp. 171–218. DOI: 10.1016/0003-682X(92)90046-U.

Møller, H., Sørensen, M. F., Jensen, C. B. and Hammershøi, D. (1996). 'Binaural Technique: Do We Need Individual Recordings?' *J. Audio Eng. Soc. (AES)* 44.6, pp. 451–469.

Møller, H., Sørensen, M. F., Hammershøi, D. and Jensen, C. B. (1995). 'Head-Related Transfer Functions of Human Subjects'. *J. Audio Eng. Soc. (AES)* 43.5, pp. 300–321.

Moore, B. C. J., Oldfield, S. R. and Dooley, G. J. (1989). 'Detection and Discrimination of Spectral Peaks and Notches at 1 and 8 kHz'. *J. Acoust. Soc. Am.* 85.2, pp. 820–836. DOI: 10.1121/1.397554.

Morimoto, M. and Ando, Y. (1980). 'On the Simulation of Sound Localization.' *J. Acoust. Soc. Jpn.* 1.3, pp. 167–174. DOI: 10.1250/ast.1.167.

Morimoto, M. and Aokata, H. (1984). 'Localization Cues of Sound Sources in the Upper Hemisphere.' *J. Acoust. Soc. Jpn.* 5.3, pp. 165–173. DOI: 10.1250/ast.5.165.

Mossop, J. E. and Culling, J. F. (1998). 'Lateralization of Large Interaural Delays'. *J. Acoust. Soc. Am.* 104.3, pp. 1574–1579. DOI: 10.1121/1.424369.

Nakagawa, S., Johnson, P. C. D. and Schielzeth, H. (2017). 'The Coefficient of Determination R^2 and Intra-Class Correlation Coefficient from Generalized Linear Mixed-Effects Models Revisited and Expanded'. *J. R. Soc. Interface.* 14.134, p. 20170213. DOI: 10.1098/rsif.2017.0213.

Neidhardt, A., Fiedler, B. and Heinl, T. (2016). 'Auditory Perception of the Listening Position in Virtual Rooms Using Static and Dynamic Binaural Synthesis'. In: *Proc. Audio Eng. Soc. (AES) Conv.* Paris, France.

Nowak, K., Tankelevitch, L., Tang, J. and Rintel, S. (2023). 'Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings'. In: *Proc. 2nd Annu. Meet. Symp. Hum.-Comput. Interact. Work*. Oldenburg Germany: ACM, pp. 1–10. DOI: 10.1145/3596671.3598578.

Oberkampf, W. L. and Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

Oberkampf, W. L. and Trucano, T. G. (2002). 'Verification and validation in computational fluid dynamics'. *Prog. Aerosp. Sci.* 38.3, pp. 209–272. DOI: 10.1016/S0376-0421(02)00005-2.

Orduña-Bustamante, F., Padilla-Ortiz, A. and Torres-Gallegos, E. A. (2018). 'Binaural Speech Intelligibility through Personal and Non-Personal HRTF via Headphones, with Added Artificial Noise and Reverberation'. *Speech Commun.* 105, pp. 53–61. DOI: 10.1016/j.specom.2018.10.009.

Parseihian, G. and Katz, B. F. G. (2012). 'Rapid Head-Related Transfer Function Adaptation Using a Virtual Auditory Environment'. *J. Acoust. Soc. Am.* 131.4, pp. 2948–2957. DOI: 10.1121/1.3687448.

Pauwels, J. and Picinali, L. (2023). 'On the Relevance of the Differences between HRTF Measurement Setups for Machine Learning'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Rhodes, Greece, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096689.

Pelzer, R., Dinakaran, M., Brinkmann, F., Lepa, S., Grosche, P. and Weinzierl, S. (2020). 'Head-Related Transfer Function Recommendation Based on Perceptual Similarities and Anthropometric Features'. *J. Acoust. Soc. Am.* 148.6, pp. 3809–3817. DOI: 10.1121/10.0002884.

Perrott, D. R. and Saberi, K. (1990). 'Minimum Audible Angle Thresholds for Sources Varying in Both Elevation and Azimuth'. *J. Acoust. Soc. Am.* 87.4, pp. 1728–1731. DOI: 10.1121/1.399421.

Perrott, D. R., Saberi, K., Brown, K. and Strybel, T. Z. (1990). 'Auditory Psychomotor Coordination and Visual Search Performance'. *Percept. Psychophys* 48.3, pp. 214–226. DOI: 10.3758/BF03211521.

Picinali, L. and Katz, B. F. G. (2023). 'System-to-User and User-to-System Adaptations in Binaural Audio'. In: *Sonic Interactions in Virtual Environments*. Ed. by M. Geronazzo and S. Serafin. Cham: Springer, pp. 115–143. DOI: 10.1007/978-3-031-04021-4_4.

Poirier-Quinot, D. and Katz, B. F. G. (2020). 'Assessing the Impact of Head-Related Transfer Function Individualization on Task Performance: Case of a Virtual Reality Shooter Game'. *J. Audio Eng. Soc. (AES)* 68.4, pp. 248–260. DOI: 10.17743/jaes.2020.0004.

Poirier-Quinot, D. and Katz, B. F. G. (2021). 'On the Improvement of Accommodation to Non-Individual HRTFs via VR Active Learning and Inclusion of a 3D Room Response'. *Acta Acust.* 5.25. DOI: 10.1051/aacus/2021019.

Poirier-Quinot, D., S. Lawless, M., Stitt, P. and Katz, B. F. G. (2022). 'HRTF Performance Evaluation: Methodology and Metrics for Localisation Accuracy and Learning Assessment'. In: *Advances in Fundamental and Applied Research on Spatial Audio*. Ed. by B. F. G. Katz and P. Majdak. IntechOpen. DOI: 10.5772/intechopen.104931.

Pollack, K., Kreuzer, W. and Majdak, P. (2022). 'Modern Acquisition of Personalised Head-Related Transfer Functions – an Overview'. In: *Advances in Fundamental and Applied Research on Spatial Audio*. Ed. by B. F. G. Katz and P. Majdak. IntechOpen. DOI: 10.5772/intechopen.102908.

Pöntynen, H. and Salminen, N. H. (2019). 'Resolving Front-Back Ambiguity with Head Rotation: The Role of Level Dynamics'. *Hear. Res.* 377, pp. 196–207. DOI: 10.1016/j.heares.2019.03.020.

Poole, K. C. et al. (2025). 'The Extended SONICOM HRTF Dataset and Spatial Audio Metrics Toolbox'. In: *Proc. Conv. EAA Forum Acusticum EuroNoise*. Málaga, Spain, pp. 6483–6486. DOI: 10.61782/fa.2025.0864.

Prepeliță, S. T., Gómez Bolaños, J., Geronazzo, M., Mehra, R. and Savioja, L. (2019). 'Pinna-Related Transfer Functions and Lossless Wave Equation Using Finite-Difference Methods: Verification and Asymptotic Solution'. *J. Acoust. Soc. Am.* 146.5, pp. 3629–3645. DOI: 10.1121/1.5131245.

Prepeliță, S. T., Gómez Bolaños, J., Geronazzo, M., Mehra, R. and Savioja, L. (2020). 'Pinna-Related Transfer Functions and Lossless Wave Equation Using Finite-Difference Methods: Validation with Measurements'. *J. Acoust. Soc. Am.* 147.5, pp. 3631–3645. DOI: 10.1121/10.0001230.

Pulkki, V. and Karjalainen, M. (2015). *Communication Acoustics: An Introduction to Speech, Audio, and Psychoacoustics*. Chichester, UK: Wiley.

Qi, X. and Wang, L. (2019). 'Parameter-Transfer Learning for Low-Resource Individualization of Head-Related Transfer Functions'. In: *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, pp. 3865–3869. DOI: 10.21437/Interspeech.2019-2558.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Raake, A. and Wierstorf, H. (2020). 'Binaural Evaluation of Sound Quality and Quality of Experience'. In: *The Technology of Binaural Understanding*. Ed. by J. Blauert and J. Braasch. Cham: Springer International Publishing, pp. 393–434. DOI: 10.1007/978-3-030-00386-9_14.

Rafaely, B. et al. (2022). 'Spatial Audio Signal Processing for Binaural Reproduction of Recorded Acoustic Scenes – Review and Challenges'. *Acta Acust.* 6, p. 47. DOI: 10.1051/aacus/2022040.

Raykar, V. C., Duraiswami, R. and Yegnanarayana, B. (2005). 'Extracting the Frequencies of the Pinna Spectral Notches in Measured Head Related Impulse Responses'. *J. Acoust. Soc. Am.* 118.1, pp. 364–374. DOI: 10.1121/1.1923368.

Lord Rayleigh (1907). 'XII. On Our Perception of Sound Direction'. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74, pp. 214–232. DOI: 10.1080/14786440709463595.

Reijniers, J., Partoens, B., Steckel, J. and Peremans, H. (2020). 'HRTF Measurement by Means of Unsupervised Head Movements with Respect to a Single Fixed Speaker'. *IEEE Access* 8, pp. 92287–92300. DOI: 10.1109/ACCESS.2020.2994932.

Reijniers, J., Vanderelst, D., Jin, C., Carlile, S. and Peremans, H. (2014). 'An Ideal-Observer Model of Human Sound Localization'. *Biol. Cybern.* 108.2, pp. 169–181. DOI: 10.1007/s00422-014-0588-4.

Reiss, L. A. J. and Young, E. D. (2005). 'Spectral Edge Sensitivity in Neural Circuits of the Dorsal Cochlear Nucleus'. *J. Neurosci.* 25.14, pp. 3680–3691. DOI: 10.1523/JNEUROSCI.4963-04.2005.

Roache, P. J. (1998). *Verification and Validation in Computational Science and Engineering*. Hermosa.

Rodamar, J. (2018). 'There ought to be a law! Campbell versus Goodhart'. *Significance* 15.6, p. 9. DOI: 10.1111/j.1740-9713.2018.01205.x.

Roginska, A. and P. Geluso, eds. (2018). *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. Audio Engineering Society Presents... New York: Routledge.

Roman, N., Wang, D. and Brown, G. J. (2003). 'Speech Segregation Based on Sound Localization'. *J. Acoust. Soc. Am.* 114.4, pp. 2236–2252. DOI: 10.1121/1.1610463.

Romigh, G. D., Brungart, D. S. and Simpson, B. D. (2015). 'Free-Field Localization Performance With a Head-Tracked Virtual Auditory Display'. *IEEE J. Sel. Topics Signal Process.* 9.5, pp. 943–954. DOI: 10.1109/JSTSP.2015.2421874.

Romigh, G. D. and Simpson, B. D. (2014). 'Do You Hear Where I Hear?: Isolating the Individualized Sound Localization Cues'. *Front. Neurosci.* 8. DOI: 10.3389/fnins.2014.00370.

Roßkopf, S., Kroczek, L. O., Stärz, F., Blau, M., Van De Par, S. and Mühlberger, A. (2024). 'The Impact of Binaural Auralizations on Sound Source Localization and Social Presence in Audiovisual Virtual Reality: Converging Evidence from Placement and Eye-Tracking Paradigms'. *Acta Acust.* 8, p. 72. DOI: 10.1051/aacus/2024064.

Rummukainen, O. S., Robotham, T. and Habets, E. A. P. (2021). 'Head-Related Transfer Functions for Dynamic Listeners in Virtual Reality'. *Appl. Sci.* 11.14, p. 6646. DOI: 10.3390/app11146646.

Saddler, M. R. and McDermott, J. H. (2024). 'Models Optimized for Real-World Tasks Reveal the Task-Dependent Necessity of Precise Temporal Coding in Hearing'. *Nat. Commun.* 15.1, p. 10590. DOI: 10.1038/s41467-024-54700-5.

Sandel, T. T., Teas, D. C., Feddersen, W. E. and Jeffress, L. A. (1955). 'Localization of Sound from Single and Paired Sources'. *J. Acoust. Soc. Am.* 27.5, pp. 842–852. DOI: 10.1121/1.1908052.

Satongar, D. A., Johnson, M. E., Jupin, P. V. and Sheaffer, J. D. (2021). 'System and Method of Determining Head-Related Transfer Function Parameter Based on in-Situ Binaural Recordings'. U.S. pat. 11,190,896 B1. Apple Inc.

Schnupp, J., Nelken, I. and King, A. J. (2010). *Auditory Neuroscience: Making Sense of Sound*. The MIT Press. DOI: 10.7551/mitpress/7942.003.0006.

Schörkhuber, C., Zaunschirm, M. and Höldrich, R. (2018). 'Binaural Rendering of Ambisonic Signals via Magnitude Least Squares'. In: *Proc. Annu. Ger. Conf. Acoust. (DAGA)*. Munich, Germany.

Seeber, B. U. and Fastl, H. (2003). 'Subjective Selection of Non-individual Head-Related Transfer Functions'. In: *Proc. Int. Conf. Auditory Display.* Boston, MA, USA.

Shinn-Cunningham, B. and Ram, S. (2003). 'Identifying Where You Are in a Room: Sensitivity to Room Acoustics'. In: *Proc. Int. Conf. Auditory Display.* Boston, MA, USA.

Simon, L. S. R., Zacharov, N. and Katz, B. F. G. (2016). 'Perceptual Attributes for the Comparison of Head-Related Transfer Functions'. *J. Acoust. Soc. Am.* 140.5, pp. 3623–3632. DOI: 10.1121/1.4966115.

Simpson, B. D., Bolia, R. S., McKinley, R. L. and Brungart, D. S. (2005). 'The Impact of Hearing Protection on Sound Localization and Orienting Behavior'. *Hum. Factors* 47.1, pp. 188–198. DOI: 10.1518/0018720053653866.

Spagnol, S., Geronazzo, M. and Avanzini, F. (2013). 'On the relation between pinna reflection patterns and head-related transfer function features'. *IEEE Trans. Audio, Speech, Language Process.* 21.3, pp. 508–519. DOI: 10.1109/TASL.2012.2227730.

Sridhar, R., Tylka, J. G. and Choueiri, E. Y. (2017). 'A Database of Head-Related Transfer Function and Morphological Measurements'. In: *Proc. Audio Eng. Soc. (AES) Conv.* New York. URL: https://aes2.org/publications/elibrary-page/?id=19308.

Steadman, M. A., Kim, C., Lestang, J.-H., Goodman, D. F. M. and Picinali, L. (2019). 'Short-Term Effects of Sound Localization Training in Virtual Reality'. *Sci. Rep.* 9.1, p. 18284. DOI: 10.1038/s41598-019-54811-w.

Stitt, P., Picinali, L. and Katz, B. F. G. (2019). 'Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues through Active Learning'. *Sci. Rep.* 9.1, p. 1063. DOI: 10.1038/s41598-018-37873-0.

Studebaker, G. A. (1985). 'A "Rationalized" Arcsine Transform'. *J. Speech, Language Hearing Research* 28.3, pp. 455–462. DOI: 10.1044/jshr.2803.455.

Takemoto, H., Mokhtari, P., Kato, H., Nishimura, R. and Iida, K. (2012). 'Mechanism for Generating Peaks and Notches of Head-Related Transfer Functions in the Median Plane'. *J. Acoust. Soc. Am.* 132.6, pp. 3832–3841. DOI: 10.1121/1.4765083.

Tervo, S., Atynen, J. P. and Kuusinen, A. (2013). 'Spatial Decomposition Method for Room Impulse Responses'. *J. Audio Eng. Soc. (AES)* 61.1/2, pp. 17–28.

Thurlow, W. R. and Runge, P. S. (1967). 'Effect of Induced Head Movements on Localization of Direction of Sounds'. *J. Acoust. Soc. Am.* 42.2, pp. 480–488. DOI: 10.1121/1.1910604.

Tollin, D. J. (2003). 'The Lateral Superior Olive: A Functional Role in Sound Source Localization'. *The Neuroscientist* 9.2, pp. 127–143. DOI: 10.1177/1073858403252228.

Trapeau, R., Aubrais, V. and Schönwiesner, M. (2016). 'Fast and Persistent Adaptation to New Spectral Cues for Sound Localization Suggests a Many-to-One Mapping Mechanism'. *J. Acoust. Soc. Am.* 140.2, pp. 879–890. DOI: 10.1121/1.4960568.

Vallat, R. (2018). 'Pingouin: Statistics in Python'. *J. Open Source Softw. (JOSS)* 3.31, p. 1026. DOI: 10.21105/joss.01026.

Van Wanrooij, M. M. and Opstal, A. J. V. (2005). 'Relearning Sound Localization with a New Ear'. *J. Neurosci.* 25.22, pp. 5413–5424. DOI: 10.1523/JNEUROSCI.0850-05.2005.

Vicente, T. and Lavandier, M. (2020). 'Further Validation of a Binaural Model Predicting Speech Intelligibility against Envelope-Modulated Noises'. *Hear. Res.* 390, p. 107937. DOI: 10.1016/j.heares.2020.107937.

Vicente, T., Lavandier, M. and Buchholz, J. M. (2020). 'A Binaural Model Implementing an Internal Noise to Predict the Effect of Hearing Impairment on Speech Intelligibility in Non-Stationary Noises'. *J. Acoust. Soc. Am.* 148.5, pp. 3305–3317. DOI: 10.1121/10.0002660.

Vliegen, J. and Van Opstal, A. J. (2004). 'The Influence of Duration and Level on Human Sound Localization'. *J. Acoust. Soc. Am.* 115.4, pp. 1705–1713. DOI: 10.1121/1.1687423.

Voong, T. M. and Oehler, M. (2019). 'Tournament Formats as Method for Determining Best-fitting HRTF Profiles'. In: *Proc. Int. Congress Acoust. (ICA)*. Aachen, Germany, pp. 4841–4847.

Wallach, H. (1940). 'The Role of Head Movements and Vestibular and Visual Cues in Sound Localization.' *J. Exp. Psychol.* 27.4, pp. 339–368. DOI: 10.1037/h0054629.

Wang, X., Peng, K., Hao, C., Yu, W., Yi, X. and Li, H. (2025). 'VR Whispering: A Multisensory Approach for Private Conversations in Social Virtual Reality'. *IEEE Trans. Visual. Comput. Graphics* 31.5, pp. 3459–3469. DOI: 10.1109/TVCG.2025.3549579.

Warnecke, M., Clapp, S., Ben-Hur, Z., Alon, D. L. and Garí, S. V. A. (2024). 'Sound Sphere 2: A High-resolution HRTF Database'. In: *Proc. Audio Eng. Soc. (AES) Conf. Audio Virtual & Augmented Reality*. Redmond, WA, USA.

Warnecke, M., Jamison, S., Prepeliţă, S., Calamia, P. and Ithapu, V. K. (2022). 'HRTF Personalization Based on Ear Morphology'. In: *Proc. Audio Eng. Soc. (AES) Conf. Audio Virtual & Augmented Reality*.

Warusfel, O. (2003). *Listen HRTF Database*. URL: http://recherche.ircam.fr/equipes/salles/listen/index.html (visited on 24/04/2024).

Watanabe, K., Iwaya, Y., Suzuki, Y., Takane, S. and Sato, S. (2014). 'Dataset of Head-Related Transfer Functions Measured with a Circular Loudspeaker Array'. *Acoust. Sci. Technol.* 35.3, pp. 159–165. DOI: 10.1250/ast.35.159.

Wen, Y., Zhang, Y. and Duan, Z. (2023). 'Mitigating Cross-Database Differences for Learning Unified HRTF Representation'. In: *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*. New Paltz, NY, USA, pp. 1–5. DOI: 10.1109/WASPAA58266.2023.10248178.

Wenzel, E. M., Arruda, M., Kistler, D. J. and Wightman, F. L. (1993). 'Localization Using Nonindividualized Head-Related Transfer Functions'. *J. Acoust. Soc. Am.* 94.1, pp. 111–123. DOI: 10.1121/1.407089.

Wightman, F. L. and Kistler, D. J. (1989a). 'Headphone Simulation of Free-Field Listening. II: Psychophysical Validation'. *J. Acoust. Soc. Am.* 85.2, pp. 868–878. DOI: 10.1121/1.397558.

Wightman, F. L. and Kistler, D. J. (1989b). 'Headphone simulation of free-field listening. II: Psychophysical validation'. *J. Acoust. Soc. Am.* 85.2, pp. 868–878. DOI: 10.1121/1.397558.

Wightman, F. L. and Kistler, D. J. (1992). 'The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization'. *J. Acoust. Soc. Am.* 91.3, pp. 1648–1661. DOI: 10.1121/1.402445.

Williamson, J., Li, J., Vinayagamoorthy, V., Shamma, D. A. and Cesar, P. (2021). 'Proxemics and Social Interactions in an Instrumented Virtual Reality Workshop'. In: *Proc. Conf. Hum. Factors Comput. Syst. (CHI)*. Yokohama, Japan: ACM, pp. 1–13. DOI: 10.1145/3411764.3445729.

Wirler, S. A., Meyer-Kahlen, N. and Schlecht, S. J. (2020). 'Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes'. In: *Proc. Audio Eng. Soc. (AES) Conf. Audio Virtual & Augmented Reality*. Online.

Wisniewski, M. G. et al. (2016). 'Enhanced Auditory Spatial Performance Using Individualized Head-Related Transfer Functions: An Event-Related Potential Study'. *The Journal of the Acoustical Society of America* 140.6, EL539–EL544. DOI: 10.1121/1.4972301.

Woodworth, R. S. (1938). *Experimental Psychology*.

Xie, B. (2013). *Head-Related Transfer Function and Virtual Auditory Display*. 2nd. J. Ross Publishing.

Yamamoto, K. and Igarashi, T. (2017). 'Fully Perceptual-Based 3D Spatial Sound Individualization with an Adaptive Variational Autoencoder'. *ACM Trans. Graph.* 36.6, pp. 1–13. DOI: 10.1145/3130800.3130838.

Yang, J., Sasikumar, P., Bai, H., Barde, A., Sörös, G. and Billinghurst, M. (2020). 'The Effects of Spatial Auditory and Visual Cues on Mixed Reality Remote Collaboration'. *J. Multimodal User Interfaces* 14.4, pp. 337–352. DOI: 10.1007/s12193-020-00331-1.

Yao, D. et al. (2022). 'An Individualization Approach for Head-Related Transfer Function in Arbitrary Directions Based on Deep Learning'. *JASA Express lett. (JASA-EL)* 2.6, p. 064401. DOI: 10.1121/10.0011575.

Yao, D. et al. (2024). 'Perceptually Enhanced Spectral Distance Metric for Head-Related Transfer Function Quality Prediction'. *J. Acoust. Soc. Am.* 156.6, pp. 4133–4152. DOI: 10.1121/10.0034632.

Yu, G., Wu, R., Liu, Y. and Xie, B. (2018). 'Near-Field Head-Related Transfer-Function Measurement and Database of Human Subjects'. *J. Acoust. Soc. Am.* 143.3, EL194–EL198. DOI: 10.1121/1.5027019.

Zagala, F., Noisternig, M. and Katz, B. F. G. (2020). 'Comparison of Direct and Indirect Perceptual Head-Related Transfer Function Selection Methods'. *J. Acoust. Soc. Am.* 147.5, pp. 3376–3389. DOI: 10.1121/10.0001183.

Zahorik, P., Bangayan, P., Sundareswaran, V., Wang, K. and Tam, C. (2006). 'Perceptual Recalibration in Human Sound Localization: Learning to Remediate Front-Back Reversals'. *J. Acoust. Soc. Am.* 120.1, pp. 343–359. DOI: 10.1121/1.2208429.

Zakarauskas, P. and Cynader, M. S. (1993). 'A Computational Theory of Spectral Cue Localization'. *J. Acoust. Soc. Am.* 94.3, pp. 1323–1331. DOI: 10.1121/1.408160.

Zenke, K. and Rosen, S. (2022). 'Spatial Release of Masking in Children and Adults in Non-Individualized Virtual Environments'. *J. Acoust. Soc. Am.* 152.6, pp. 3384–3395. DOI: 10.1121/10.0016360.

Zhang, M., Ge, Z., Liu, T., Wu, X. and Qu, T. (2020). 'Modeling of Individual HRTFs Based on Spatial Principal Component Analysis'. *IEEE/ACM Trans. Audio, Speech, Language Process.* 28, pp. 785–797. DOI: 10.1109/TASLP.2020.2967539.

Zhao, J., Yao, D., Gu, J. and Li, J. (2024). 'Efficient Prediction of Individual Head-Related Transfer Functions Based on 3D Meshes'. *Appl. Acoust.* 219, p. 109938. DOI: 10.1016/j.apacoust.2024.109938.

Zhao, J., Yao, D. and Li, J. (2025). 'Cross-Dataset Head-Related Transfer Function Harmonization Based on Perceptually Relevant Loss Function'. *IEEE Open J. Signal Process.* 6, pp. 865–875. DOI: 10.1109/OJSP.2025.3590248.

Zheng, Q. et al. (2023). 'Understanding Safety Risks and Safety Design in Social VR Environments'. *Proc. ACM Hum.-Comput. Interact.* 7 (CSCW1), pp. 1–37. DOI: 10.1145/3579630.

Zhi, B., Zotkin, D. N. and Duraiswami, R. (2022). 'Towards Fast and Convenient End-to-End HRTF Personalization'. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Singapore, pp. 441–445. DOI: 10.1109/ICASSP43922.2022.9746315.

Zhong, X.-l. and Xie, B.-s. (2014). 'Head-Related Transfer Functions and Virtual Auditory Display'. In: *Soundscape Semiotics - Localisation and Categorisation*. Ed. by H. Glotin. InTech. DOI: 10.5772/56907.

Zonooz, B., Arani, E., Körding, K. P., Aalbers, P. A. T. R., Celikel, T. and Van Opstal, A. J. (2019). 'Spectral Weighting Underlies Perceived Sound Elevation'. *Sci. Rep.* 9.1, p. 1642. DOI: 10.1038/s41598-018-37537-z.

Zotkin, D., Hwang, J., Duraiswaini, R. and Davis, L. (2003). 'HRTF Personalization Using Anthropometric Measurements'. In: *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*. New Paltz, NY, USA, pp. 157–160. DOI: 10.1109/ASPAA.2003.1285855.

Zotkin, D. N., Duraiswami, R. and Gumerov, N. A. (2009). 'Regularized HRTF Fitting Using Spherical Harmonics'. In: *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*, pp. 257–260. DOI: 10.1109/ASPAA.2009.5346521.

Appendix A

Proof of Permission to Reproduce Published Work

4 August 2025

Dear Audio Engineering Society,

I am completing my PhD thesis at Imperial College London entitled ‘Auditory Model-Based Similarity Metric for Head-Related Transfer Functions’.

I seek your permission to reprint, in my thesis, extracts from:

1. Engel, I. et al. (2023). ‘The SONICOM HRTF Dataset’. *J. Audio Eng. Soc.* 71.5, pp. 241–253. DOI: 10.17743/jaes.2022.0066.
2. Daugintis, R., Alary, B., Geronazzo, M. and Picinali, L. (2024). ‘Effects of Binaural Rendering Personalisation and Reverberation on Speech-on-Speech Masking’. In: *Proc. Audio Eng. Soc. Conf. Audio Virtual & Augmented Reality (AES AVAR)*. Redmond, WA, USA. URL: https://aes2.org/publications/elibrary-page/?id=22671.

Extracts and figures authored by me would be incorporated into the thesis chapters. A citation and a link to the published version would be provided in the appropriate sections of the thesis.

The thesis will be added to Spiral, Imperial's institutional repository http://spiral.imperial.ac.uk/ and made available to the public under a CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).

If you are happy to grant me all the permissions requested, please return a signed copy of this letter. If you wish to grant only some of the permissions requested, please list these and then sign.

Yours sincerely,

Rapolas Daugintis

[redacted]

Permission granted for the use requested above:

I confirm that I am the copyright holder of the extract above and hereby give permission to include it in your thesis which will be made available, via the internet, for non-commercial purposes under the terms of the user licence.

[please edit the text above if you wish to grant more specific permission]

Signed: [redacted]

Name: Megan Stewart

Organisation: Audio Engineering Society

Job title: Assocaite Director

Classifying Non-Individual Head-Related Transfer Functions with A Computational Auditory Model: Calibration And Metrics

Conference Proceedings:
ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Author: Rapolas Daugintis

Publisher: IEEE

Date: 04 June 2023

Quick Price Estimate

I would like to... reuse in a thesis/dissertation