

Sound Source Localization and Virtual Testing of Binaural Audio



Hugh O'Dwyer

Department of Electronic Engineering
University of Dublin, Trinity College

This dissertation is submitted for the degree of
Doctor of Philosophy

March 2021

To my friends and family.

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgment.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Hugh O'Dwyer
March 2021

This thesis presents a series of studies concerned with the analysis and assessment of binaural audio. These works are for the most part concerned with the area of sound source localization with machine learning techniques used to determine the cues and features which help the human auditory system perform such tasks. This work aims to bridge the gap in knowledge on how we perceive the height of sound sources in particular. Also presented in this thesis is research which investigates methods of virtual headphone testing and an assessment of the timbral quality and directional encoding accuracy of Ambisonic microphones. Following some introductory chapters on spatial audio and its perception and reproduction, these research topics are discussed.

In the work presented on virtual headphone testing, the results of two virtual headphone test methodologies are compared for similarities and differences. While a virtual headphone test method which uses the transfer function of different headphones to generate artificial stimuli has been widely accepted as accurate in its results compared to conventional headphone testing, a method which uses binaural recordings of stimuli has been disregarded due largely to a small number of inconclusive studies. It has also been shown that the perception of spatial audio in virtual headphone testing correlates strongly with general audio preference. The results of the work verify that when using a transfer function based, 'filtered' virtual headphone test, there is indeed a direct correlation between the rating of general audio quality and spatial audio quality. However, this same result was not found when using binaural recordings to perform virtual headphone testing showing that the two methods produce statistically different results. This furthers the claim that a recorded binaural method is unsuitable for on-ear virtual headphone testing. It also vindicates the use of binauralisation using transfer function based convolution.

A series of machine learning algorithms are also presented which aim to resolve sound source localization with a variety of interaural and spectral cues. As our understanding of the phenomena which allow us to perceive sound source height and distance is limited, the development of physical models which can accurately predict these metrics has been limited. However, with the recent rise in accessibility to machine learning platforms, a unique opportunity to resolve this problem has presented itself. This thesis presents a series of studies which use machine learning algorithms to find a robust feature set capable of performing three-dimensional sound source localization in binaural signals in both the matched and mismatched HRTF condition.

A combination of deep neural networks and random forest algorithms are assessed in their ability to accurately predict source location. Through experimentation it is found that gammatone cepstral coefficients are capable of representing vertical localization cues and are combined with interaural features to create a new feature set for performing sound source localization. Using the adaptive affinity propagation clustering algorithm, generalization of the proposed network is improved for the mismatched HRTF condition.

The proposed network is shown to perform well in the mismatched HRTF condition in both anechoic and adverse environments while the results of the HRTF clustering reveal specific anthropometric features which are most relevant for vertical localization.

Acknowledgements

I would first like to thank my supervisor Professor Frank Boland for offering me the wonderful opportunity to pursue this Ph.D. His enthusiasm and wealth of knowledge were invaluable attributes which kept me going throughout the four years of this endeavour.

I would also like to thank Dr. Enda Bates for providing me with the wonderful experience of assisting him in his research on Ambisonic microphones. This provided me with the confidence and experience in the early stages of my time in academia that undoubtedly enabled me to conduct my own research.

My colleagues in the spatial audio research team have been incredible friends and I would like to thank Luke Ferguson, Sebastian Csadi, Brian Fallon, and Sean Dooney for their continued support. In addition to this I would also like to acknowledge the rest of the Sigmedia team for sharing their time and expertise with me over the years so in no particular order thank you to Matt Roddy, George Sterpu, Colm O'Reilly, Marco Forte, & Daniel Ringis.

The Electronic Engineering Department has an abundance of incredible lecturers and I'd like to thank Naomi Harte, Anil Kokaram, Francois Pitie, Brian Foley for trusting me to teach their labs and tutorials. The staff of the Music and Media Technologies course were also very kind in assisting me in my work and I'd like to express my gratitude to Jenny Kirkwood, Jimmy Eadie, and Dermot Furlong.

Some of the best friends I made along this journey were the technical staff of the Electronic Engineering department with whom I taught classes and played some not-so easy going games of football. Thank you to Conor Nolan, John Squires, Shane Hunt, Eugene O'Rourke, and James McLoughlin.

Lastly I'd like to thank my friends and family for their support throughout the course of this Ph.D. For listening to my ramblings and pretending to understand what I was doing with myself, your love and patience was invaluable.

Abstract

This thesis is concerned with aspects of our perception of spatial audio and with novel technology used to record, analyse and present it. The field of spatial audio has become increasingly prevalent with mediums such as virtual reality and augmented reality becoming readily available to anyone with a smart phone. The sense of immersion when experiencing a virtual environment is not only reliant on visuals but also on audio with an accurate portrayal of an auditory scene crucial to creating a sense of realism. The work presented in this thesis focuses on the analysis of spatial audio including ambisonic and binaural signals. Fundamentally, this research is focused on localization with machine learning tools used to evaluate localization cues in these binaural signals. Preliminary chapters on the perception and reproduction of spatial audio provide a relevant background to the research presented in this thesis. The major contributions of this thesis can be described as follows:

A subjective study comparing the quality of spatial audio over different headphones is presented. Using virtual headphone test methodologies, a series of MuSHRA tests were conducted which examined both the perceived spatial quality and general audio quality preference of a variety of stimuli. The results of the study show that there is a strong correlation between the perception of general audio quality and spatial audio quality presented via headphones. This is true however only when using a transfer function based virtual headphone test methodology is used. The study shows that results found using a binaural recording based approach to virtual headphone testing are inconsistent.

This investigation into the capture and presentation of binaural audio is followed by studies in sound source localization in binaural signals using machine learning. This work explores the relationship between spectral and interaural cues with source direction. These studies focus entirely on binaural audio which has been synthesised using convolution with Head Related Impulse Responses (HRIRs). Novel methods are presented which use a combination of feature extraction with logistic classification to develop a proficient algorithm to localise sound sources in terms of both horizontal and vertical position. While the frequency dependent cues which allow us to detect sound source elevation are difficult to model using traditional methods, the results of these studies show they can be modelled accurately using deep learning and gammatone cepstral coefficients.

To improve the performance of this algorithm in the mismatched Head-Related Transfer Function (HRTF) condition, a clustering algorithm is introduced and used to cluster similar HRTF sets together providing a more robust performance. This clustering approach is evaluated in both anechoic and adverse environments and is used to identify anthropomorphic features which are important for determining our sense of vertical localization.

Table of contents

List of figures	xvii
List of tables	xxiii
Nomenclature	xxv
1 Introduction	1
1.1 Focus of Thesis	3
1.2 Thesis Outline	3
1.3 Contribution of Thesis	6
1.4 Publications	6
2 Perception of Spatial Audio	11
2.1 The Human Auditory System	11
2.2 Sound Source Localization	14
2.2.1 Head Related Coordinate System	14
2.2.2 Head Related Impulse Response	15
2.2.3 Head Related Transfer Function	16
2.2.4 Localization in the Horizontal Plane	18
2.2.5 Localization in the Vertical Plane	20
2.2.6 Distance Perception	29
2.3 Sound Source localization in Reverberant Environments	31
2.3.1 Reverberation and the Room Impulse Response	32
2.3.2 The Precedence Effect	33
2.3.3 Measurement of Room Acoustic Properties	34
2.3.4 Binaural Room Impulse Response	35
2.4 Modelling the Human Auditory System	35
2.4.1 Auditory Scales	36
2.4.2 Auditory Filter Models	38
2.5 Conclusion	38

3	Reproduction of Spatial Audio	41
3.1	Stereophony	41
3.1.1	2-channel Stereophony	42
3.1.2	Multichannel Stereophony	46
3.1.3	Vector Base Amplitude Panning	48
3.2	Digital Sound Projection	48
3.2.1	The Image Source Method (ISM)	52
3.2.2	The PRobabilistic Image Source Method (PRISM)	53
3.3	Ambisonics	54
3.3.1	Encoding Ambisonics	56
3.3.2	Decoding Ambisonics	57
3.3.3	Higher Order Ambisonics	60
3.4	The TCD Audiosphere	61
3.5	Comparing Ambisonic Microphones	63
3.5.1	Review of Directional Analysis Methods	63
3.5.2	Method	64
3.5.3	Results & Analysis	67
3.5.4	Discussion	67
3.6	Binaural Reproduction	70
3.6.1	Virtual Loudspeaker Approach	71
3.6.2	Optimised Virtual Loudspeaker Approach	72
3.6.3	Headphone Design	72
3.7	Conclusion	73
4	Virtual Headphone Testing for Spatial Audio	75
4.1	Virtual Headphone Testing	76
4.2	Methodology	78
4.2.1	Headphones	79
4.2.2	Stimuli	81
4.2.3	Stimuli Preparation - Binaural Recording and Headphone Frequency Response Measurement	83
4.3	Testing	84
4.4	Experimental Results	90
4.5	Discussion	93
4.6	Conclusion	95
5	Machine Learning and its Applications in Audio Engineering	97
5.1	Machine Learning	97

5.1.1	Machine Learning Terminology	98
5.1.2	Generalization	100
5.2	Machine Learning and Audio Engineering	101
5.3	Supervised Learning	102
5.3.1	Multi-Layer Perceptrons: Neural Networks	103
5.3.2	Decision Trees and Random Forests	107
5.3.3	Mutual Information	109
5.4	Conclusion	110
6	Binaural Feature Selection for Sound Source Localization	113
6.1	Introduction	113
6.2	Binaural System Model	114
6.2.1	Synthesising Binaural Signals	115
6.2.2	Binaural Cues and Their Extraction	116
6.3	Single Feature Performance for SSL Tasks	121
6.3.1	Method	123
6.3.2	Feature Performance on the Horizontal Plane	125
6.3.3	Feature Performance on the Median Plane	128
6.3.4	Feature Performance on the Sagittal Plane at 45°	131
6.4	Multi-Feature Performance for SSL Tasks	133
6.4.1	DNN training using CCF and ILD, the Ma Method	134
6.4.2	Random Forest Training using ILD and IPD, the Wu Method	134
6.4.3	Comparing Multi-Feature Localization Performance	135
6.4.4	Discussion	136
6.5	Conclusion	139
7	HRTF Clustering for Robust Full Sphere Sound Source Localization	141
7.1	Introduction	141
7.2	The Affinity Propagation Clustering Algorithm	142
7.2.1	The AP algorithm	142
7.2.2	An adaptive AP algorithm	144
7.2.3	The Silhouette Index	144
7.3	Clustering Analysis	145
7.4	Spherical Localization and Angular Error	147
7.4.1	Localization Performance	149
7.5	Anthropometric Feature Importance	150
7.6	Conclusion	154

8 Conclusion	157
8.1 Summary	158
8.2 Future Work	160
8.2.1 Chapter 4	160
8.2.2 Chapters 6 and 7	161
8.3 Closing Remarks	162
Appendix A Comparing Ambisonic Microphones	165
A.1 Ambisonic Microphones	165
A.2 Subjective Listening Test	168
A.2.1 Experiment A - Timbral Quality	168
A.3 Conclusions From Comparing Ambisonic Microphones	173
References	175

List of figures

1.1	VR headsets commonly used in gaming applications	2
2.1	The anatomy of the ear, reproduced from [128]. Sound waves travel through the ear canal and vibrate the eardrum (tympanic membrane). These vibrations are transmitted to cochlea via three tiny bones. Depending on the frequency of the sound, different parts of the cochlea are excited, sending messages to the brain.	12
2.2	A diagram of the Human Cochlea, reproduced from [170]. The Basilar Membrane (BM) is located inside the cochlea between the scala media and the scala tympani. The BM is moved when sound waves fall on the ear. Different frequencies cause different parts of the BM to move, high frequencies resonate near the base of the cochlea while low frequencies resonate towards its apex.	13
2.3	The Head Related Coordinate System used in this thesis. θ represents the horizontal angle relative to the centre of the listener. ϕ represents the vertical angle relative to the centre of the listener and r represents the distance of the sound source relative to the centre of the listener	14
2.4	HRIR pairs from subject 2 of the SADIE II database [8]. Left HRIRs (orange) and right HRIRs (blue) for 0° Azimuth and 0° Elevation (Left) and for 45° Azimuth and 0° Elevation (Right)	15
2.5	HRTF pairs from subject 2 of the SADIE II database [8]. Left HRTFs (orange) and right HRTFs (blue) for 0° Azimuth and 0° Elevation (Left) and for 45° Azimuth and 0° Elevation (Right)	17
2.6	Phase ambiguities arise when wavelength λ is significantly smaller than the diameter of the head. This occurs at approximately 1500Hz	19
2.7	The cone of confusion in sound source localization caused by similar values for ITD and ILD	20
2.8	Vertical image position compared with loudspeaker position for octave band pink noise. Reproduced after Cabrera [32].	21

2.9	localization blur in the median plane for a continuous speech signal, reproduced after Blauert [20]	22
2.10	HRIRs (left) and HRTFs (right) displayed as images showing the effect that reflections and diffraction cause on the HRIR with increasing elevation. This subsequently results in Pinna nulls and resonances in the HRTF. This particular plot uses HRTFs from subject 10 in the CIPIC database at an azimuth angle of 0°	24
2.11	Raykar's method for extracting the notch frequencies created by head, torso, and pinna reflections from the HRIR [158]. The HRIR can be seen in (a) through (f) as it goes through the process of notch extraction. The corresponding spectrum magnitude can be seen in (f) through (j) with the group delay employed in (k) to extract the notch frequencies.	25
2.12	Notch frequencies in the HRTFs of subject 11 of the CIPIC database. Notches are shown for both ears and at 5 different azimuthal planes. Notches found using Raykar's method as described in [158].	27
2.13	A set of possible paths for a sound wave to take between a source and receiver in a room. These include a direct path along with a first and second order reflection	32
2.14	A typical echogram of a reverberant space. A direct sound is caused by a short, impulsive signal such as handclap, balloon pop or gun-shot. Following the direct sound is a series of 1st and 2nd order early reflections. The reverberation tail is made up of 3rd order and higher reflections. The density at which these reflections arrive makes it impossible to distinguish their exact order	33
2.15	A 32-channel Gammatone Filterbank with center frequencies ranging from 0-8kHz.	37
2.16	The Mel Scale relative to the Hertz Scale	37
3.1	2-channel setup for stereophonic sound reproduction	42
3.2	The AB microphone technique by Fletcher. This technique aims to capture the time differences between two microphone signals	43
3.3	The Blumlein microphone technique. This coincident pair aims to capture amplitude differences between two microphone signals	43
3.4	The XY coincident microphone setup	44
3.5	The Mid-Side (M-S) coincident microphone setup	44
3.6	The near-coincident ORTF microphone configuration	45
3.7	The KEMAR dummy head microphone from GRAS [66]	46

3.8	Recommended 5-channel configuration for the <i>ITU 5.1</i> surround sound reproduction	47
3.9	3-channel stereophonic configuration reformulated with vectors, i.e. VBAP	49
3.10	Beamforming from array of loudspeakers creates virtual loudspeaker positions to emulate a 5.1 surround sound setup	50
3.11	Convergent Audio Beamforming with a line array	51
3.12	Lattice with mirror-image sources used for calculating early reflections in a simple RIR model. The dashed lines represent the natural path of the reflections while the solid arrow tipped lines represent the actual paths between source and receiver. Here blue represents 1st order reflections and green represents 2nd order reflections	53
3.13	The PRobabilistic Image Source Method (PRISM). The uncertain reflectors, W1, W2 and W3, are represented with blurry lines. The uncertainty of first order, R1, R2 and R3, and second order virtual receivers, R12 and R23, are also shown. The plane behind the source array is not considered for reflections. The reflectors and the virtual receiver positions become more spread out as the number of reflections increases	54
3.14	The practical tetrahedral microphone (left) and design from patent document (right), reproduced after Craven [39].	55
3.15	B-format signal configurations, reproduced after Craven [39].	57
3.16	Shown here are the five platonic solids. Placing loudspeakers at their vertices offers the only means of uniformly distributing loudspeakers in 3D space	59
3.17	Spherical Harmonics up to fourth order	61
3.18	Positioning system for Experiment B. The MH acoustics Eigenmike is placed at the centre of the TCD Audiosphere [11]	65
3.19	Histogram of azimuth angles across 25 frequency bands [13]	67
3.20	Mean azimuth and elevation error for 16 sources [13]	69
3.21	Azimuth average error and variance [13]	69
3.22	Elevation average error and variance [13]	70
3.23	Shown here is the setup for binaurally rendering sound, with each sound source processed separately	71
3.24	From left to right - in-ear, over-ear, and on-ear headphones	73
4.1	Beyerdynamic DT-770 headphones. The replicator headphones used in this study for their flat frequency response and high fidelity [148]	79

4.2	The frequency response of the Beyerdynamic DT-770 headphones measured on the Neumann KU-100 Dummy Head Microphone and its inverse filter. A total of 31 measurements were taken in 1/3rd octave bands from 20Hz to 20kHz.	80
4.3	A variety of headphones mounted on a Neumann KU-100 dummy head microphone. Top-left: Sennheiser HD 650 (Circumaural, Open-Back), Top-Right: Sennheiser PX-100-II (Circumaural, Open-Back), Bottom-Left: Koss KSC 75 (Circumaural, Closed-Back), Bottom-Right: Philips SH5200 (Circumaural, Closed-Back) [148]	81
4.4	Frequency Response of the headphones used in the study. Solid curves, left channel; dashed curve, right channel [148]	82
4.5	Spectrograms of the original raw recording of the 'Funk' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.	85
4.6	Spectrograms of the original raw recording of the 'Nature' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.	86
4.7	Spectrograms of the original raw recording of the 'Shaker' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.	87
4.8	Spectrograms of the original raw recording of the 'Surround' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.	88
4.9	The experimental results for the first MUSHRA test where subjects ranked the general and spatial quality of the audio signals obtained using the 'filtered' method [148]	91
4.10	The experimental results for the first MUSHRA test where subjects ranked the general and spatial quality of the audio signals obtained using the 'recorded' method [148]	92
5.1	A neural network with an input layer, hidden layer and an output layer . .	104
5.2	A Decision Tree	108
5.3	The variation in mutual information for ILD and IPD up to 20kHz for binaural speech signals with SNR: Inf and SNR: 30dB.	111

5.4	The variation in mutual information for ILD and IPD up to 20kHz for binaural speech signals with SNR: 20dB and SNR: 10dB.	112
6.1	Examples of CCF plots between binaural signals at $-80^\circ, 0^\circ, 80^\circ$ on the horizontal plane. These measurements were made using random speech signals from the TSP database convolved with pairs of HRIRs from CIPIC subject 165.	118
6.2	Finding W_{IACC} , an approximation for ASW. Reproduced after Sato [167].	119
6.3	The center frequencies used to calculate MFCC and GTCC values. . . .	122
6.4	A flowchart of the DNN based localization method.	123
6.5	The architecture of the DNN used to perform sound source localization using single feature inputs.	124
6.6	The performance of each feature along the Horizontal Plane, correctly classified. A single training HRTF is used to train the DNN and subsequently tested on each of the 45 HRTFs from the CIPIC dataset.	126
6.7	Horizontal plane feature performance to within 5° . Accuracy is shown as the average performance in both the matched and mismatched HRTF condition.	127
6.8	The performance of each feature along the Median Plane, correctly classified to within 6° . A single training HRTF is used to train the DNN and subsequently tested on each of the 45 HRTFs from the CIPIC dataset. . .	129
6.9	Median plane feature performance for estimating elevation to within 6° s. Accuracy is shown as the average performance in both the matched and mismatched HRTF condition.	130
6.10	The performance of each feature along the Sagittal Plane at 45° , correctly classified to within 6° . A single training HRTF is used to train the DNN and subsequently tested on each of the 45 HRTF sets from the CIPIC dataset.	132
6.11	Sagittal plane at 45° feature performance for estimating elevation to within 6° . Accuracy is shown as the average performance in both the matched and mismatched HRTF condition.	133
6.12	Horizontal Plane accuracy in the matched and mismatched condition for azimuth estimation to within 5 degrees.	137
6.13	Median plane accuracy in the matched and mismatched condition for elevation estimation to within 6 degrees.	137
6.14	Sagittal plane at 45° accuracy in the matched and mismatched condition for elevation estimation to within 6 degrees.	138

6.15	Sagittal plane at 45° method performance for estimating elevation to within 6°. A single training HRTF is used to train the algorithm and subsequently tested on each of the 45 HRTFs from the CIPIC dataset. Two popular methods of binaural localization are compared to two new proposed methods.	139
7.1	The Similarity Matrix for the proposed method with performance scores based on sagittal plane localization accuracy to within 6 °.	146
7.2	The Silhouette Index values for 2 to 10 clusters found using the Adaptive Affinity Propagation algorithm.	147
7.3	The performance of the proposed feature set along the Saggital Plane at 45°, correctly classified to within 6°. Performance is compared for training on a single HRTF set versus training on 4 central HRTF sets.	148
7.4	Source direction estimates from noisy signals.	148
7.5	Pinna geometry measurements of the CIPIC database, reproduced after Algazi [5].	152
7.6	Box plots of cavum concha width in both left and right ears for subjects within each cluster.	153
A.1	The Ambisonic microphones compared in Part 1 of the comparing Ambisonic microphones study (from left to right): DPA 4006, Core Sound TetraMic, Soundfield MKV, MH Acoustics Eigenmike, and Zoom H2n [11]	166
A.2	The Sennheiser Ambeo microphone which was included in Part 2 of the comparing Ambisonic microphones study [13]	166
A.3	Quartet Recording Setup [13]	169
A.4	Listening Test User Interface [13]	169
A.5	Listening Test Results: Q1 Bright/Dull [13]	172
A.6	Listening Test Results: Q2 Full/Thin [13]	172

List of tables

3.1	The distances and locations of each loudspeaker relative to the centre of the TCD Audiosphere	62
3.2	Reverberation time per octave band after treatment of room.	63
3.3	Mean Azimuth (AZ) offset and Elevation (EL) offset error $\pm 95\%$ Confidence Interval for 16 loudspeaker locations	68
4.1	The headphones examined in this study:	80
4.2	The stimuli used in this study:	83
4.3	Correlation coefficients (R) [148]	90
4.4	Correlation <i>p-values</i> [148]	90
4.5	General quality preference. Test method - <i>filtered</i>	93
4.6	Spatial quality preference. Test method - <i>filtered</i>	93
4.7	General quality preference. Test method - <i>recorded</i>	94
4.8	Spatial quality preference. Test method - <i>recorded</i>	94
6.1	Performance of Individual Features for Localization on the Horizontal Plane	125
6.2	Performance of Individual Features at Vertical Localization on the Median Plane.	128
6.3	Performance of Individual Features for Localization on the Sagittal Plane at 45°	131
7.1	Clustering result with 4 clusters	147
7.2	Average Angular Error in Noisy environments with and without HRTF Clustering	150
7.3	Average Azimuth and Elevation Accuracy with and without HRTF Clustering	151
7.4	Pinna Feature Performance	152

Nomenclature

Roman Symbols

r sound source distance

Greek Symbols

λ the wavelength of a sound wave

ϕ vertical location of a sound source in degrees

Θ The joint position of a sound source in horizontal and vertical space

θ azimuthal location of a sound source in degrees

Acronyms / Abbreviations

AAP Adaptive Affinity Propagation

AP Affinity Propagation

AR Augmented Reality

ASW Apparent Source Width

BRIR Binaural Room Impulse Response

CCF Cross-Correlation Function

DDR Direct to Reverberant Ratio

DOA Direction Of Arrival

DSP Digital Signal Processing

ERB Equivalent Rectangular Bandwidth

FVP Free View-Point

GTCCs Gammatone Cepstral Coefficients

HAS Human Auditory System

HpTF Headphone Related Transfer Function

HRIR Head Related Impulse Response

HRTF Head Related Transfer Function

IACC Interaural Cross-Correlation Coefficient

ILD Interaural Level Difference

IPD Interaural Phase Difference

ISM Image Source Method

ITD Interaural Time Difference

MFCCs Mel Frequency Cepstral Coefficients

PRISM PRobabilistic Image Source Method

RF Random Forest

RIR Room Impulse Response

RTF Relative Transfer Function

SSL Sound Source Localization

VAE Virtual Auditory Environment

VBAP Vector Base Amplitude Panning

VR Virtual Reality

Chapter 1

Introduction

This thesis is concerned with spatial audio and in particular with Sound Source Localization (SSL). As humans, our ability to localise sound sources in three-dimensional space has developed throughout the course of our evolution as a species. And while our understanding of the phenomena which enable us to perform this task is broad, there are still certain factors which we have not fully realised. One such aspect of sound source localization which we do not fully comprehend is our ability to localise sound in the vertical plane, a phenomenon for which we have no accurate predictive model. It is known however, that this task is performed using sound wave reflections from the head, shoulders, and ears (pinnae). These reflections vary depending on the direction of the source toward the individual and as such can cause interference to the signal perceived at the ear. This happens particularly at frequencies greater than 3kHz, this is where the wavelength, λ of the sound source becomes comparable to the pinna size [2]. However, as the distances and dimensions of our bodies are unique to each individual, so too is our sense of elevation perception. As a result, modelling the relationship between reflected sound waves and perceived elevation is an incredibly difficult task.

An increased interest in spatial hearing for Virtual Auditory Environments (VAEs) as well as the prevalence of machine learning in recent years has presented a unique opportunity to explore this phenomenon in greater detail. VAEs are presented to listeners through the mediums of Virtual Reality (VR), Augmented Reality (AR), video games, and in educational tools to name a few. While these sound scenes can sometimes be pre-rendered, they are usually generated in real time using Digital Signal Processing (DSP) techniques. VAEs can consist of many sound sources positioned at various angles and distances relative to the user and it's imperative to accurately simulate these positions to present a realistic sound-field to the listener. If a VAE presented to a user is not perceptually comparable to a real world auditory scene, it can break the sense of immersion intended by the experience and thus render the medium lacking in its execution [80].



Fig. 1.1 VR headsets commonly used in gaming applications

VR has seen somewhat of a resurgence in the last five years with many manufacturers releasing their own VR headsets. These include the HTC Vive [83], the Oculus Rift from Facebook [141], and Sony’s Playstation VR [180] which can be seen in Figure 1.1. While these devices are somewhat costly, there exist cheaper alternatives such as the Google Cardboard [65], and Samsung’s Gear VR [165] which bring VAEs to a much larger audience.

Beyond entertainment, VAEs have proven to be a successful tool in the field of education. A 2017 study by Lui & Hoy found that surgeons who trained using virtual reality simulation of temporal bone surgery showed an improvement in mastoidectomy performance [116]. VAEs can also be used to treat patients suffering from anxiety disorders as shown in a 2019 study by Denbow and Anderson which demonstrated that virtual reality exposure therapy was as effective as in-vivo exposure therapy [16].

With recent advancements in computational power and access to a wider range of databases it has been possible to employ neural networks and other machine learning techniques to better understand how our perception of sound works. For instance there have been successful studies performed by the author where neural networks have been used to analyse the direction and distance of a sound source in a binaural audio scene. A better understanding of our perception of spatial audio is fundamental to the improvement of *auralisation* in VAEs and can even be used to better synthesise sound sources for these applications.

In essence the main hypothesis of this thesis can be formulated as follows:

Modern tools such as deep neural networks can be employed to help us better understand and model the perception of sound source localization, a phenomenon which is highly unique to the individual. This thesis proposes that in using these tools, we can gain a deeper understanding of the cues which allow us to perform

sound source localization in three dimensional space. This in turn can allow us to understand the individuality of the HRTF. As a majority of spatial audio is synthesised using transfer functions, there is a need to fully comprehend it both in terms of subjective quality and directional encoding accuracy. In doing so, we may therefore enhance the sense of immersion experienced in virtual auditory environments.

In addition to this hypothesis, this thesis also contains experimental work which analyses the tools we use to capture and present audio in VAEs, i.e. omnidirectional microphones and headphones. These analyses are built on existing frameworks, specifically, the MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA) listening test and Directional Audio Coding (DirAC). As such, this thesis is indebted to the developers of these methods and hopes to expand on their work by further developing these techniques.

1.1 Focus of Thesis

This thesis focuses on the areas of spatial audio, sound source localization and its applications. Several experiments are described in chapters 3 and 4 which investigate the effect that 3-dimensional microphones and headphones have on the perceived quality of spatial audio. In chapters 5,6, and 7 a series of experiments using neural networks and other machine learning techniques are described which allow for the development of a robust algorithm capable of performing sound source localization in both the matched and mismatched HRTF condition.

1.2 Thesis Outline

The previous part of this chapter defined the context and focus of this work. The remainder of this thesis is organised as follows:

Chapter 2: Perception of Spatial Audio

This chapter discusses the mechanisms behind the perception of sound in humans. Beginning with the physiology behind the human auditory systems, the fundamental cues and phenomena responsible for the detection of sound source direction, distance, and motion are described. In addition to describing the human auditory system, a description of some of the popular ways in which it is modelled using digital filters is discussed. The cues and limitations of sound localization in the horizontal and vertical planes is explored. Our ability to localise sound with varying distances and in reverberant environments is also discussed. To explain these mechanisms, a head related coordinate system is introduced which is used throughout this thesis.

Chapter 3: Reproduction of Spatial Audio

Spatial audio can be reproduced using a variety of methods. In this chapter we describe some of the most common means of presenting spatial audio to a listener. Beginning with 2-channel stereophony, the fundamentals of loudspeakers and the production of phantom sources is described accompanied by some common stereo recording techniques. Following this is the expansion into multi channel stereophony and some of the common systems used to present it such as 5.1 home speaker setups. This includes a section on Vector Base Amplitude Panning (VBAP). Moving on from stereophony, an exploration of digital sound projection is given with a particular emphasis on sound bar technology. Within this subsection is a description of the PRobabilistic Image Source Method (PRISM). This is a novel theoretical method for calibrating digital sound projectors in rooms with uncertain geometries which was contributed to by the author. An in-depth discussion of ambisonics is included which features work performed by the author on the evaluation of directional encoding of ambisonic microphones. Finally, the binaural reproduction of spatial audio over headphones is discussed in anticipation of chapter 4 of this thesis on virtual headphone testing.

Chapter 4: Virtual Headphone Testing for Spatial Audio

This chapter explores the merits of two common methods for performing Virtual Headphone Testing. By far the most common means of reproducing spatial audio is via headphones; therefore it is imperative to evaluate the means by which we compare the quality of spatial audio presented over headphones. While many biases arise when using a traditional swap and compare method to headphone comparison such as fit, branding, etc., virtual headphone testing provides an alternative whereby the headphones under test are simulated using digital signal processing techniques. The first virtual headphone testing method used in this comparison is a *filtered* method whereby the Headphone Related Transfer Function (HpTF) of each headphone is measured using a dummy head microphone and applied to a variety of sound stimuli to recreate the effect of each headphone. The second method is a *recorded* method, where the same set of stimuli are recorded from each pair of headphones directly onto a dummy head microphone. Both the general and spatial audio quality of stimuli prepared using these two methods is evaluated using a MUSHRA test. While this study validates the consistency of results found using a *filtered* test methodology, inconsistent results found via the *recorded* method fail to validate it as a reliable means of examine headphone quality preference.

Chapter 5: Machine Learning and its Applications in Audio Engineering

This chapter introduces the reader to the fundamentals of machine learning and the relationship between machine learning and audio engineering. Beginning by explaining

the terminology and core concepts behind machine learning techniques, this chapter lays the groundwork for the experimental work performed in chapters 6 and 7. An overview of popular machine learning studies related to audio engineering is given as well an insight to the techniques used in these works. This is followed by an in-depth explanation of supervised learning techniques; specifically, multi-layer perceptrons or neural networks and the decision tree based random forest algorithm are discussed. These are the techniques which are used in later chapters. While on the subject of decision trees, the concept of information gain or *mutual information* is introduced and used to visualise the reliance of vertical localization on spectral cues above 7kHz. It is also shown how vertical localization diminishes in the presence of noise.

Chapter 6: Binaural Feature Selection for Sound Source Localization

This chapter focuses on the area of machine learning for binaural sound source localization. It contains the results of several experiments which seek to evaluate the ability of binaural cues to determine sound source azimuth and elevation using machine learning algorithms. While it is shown that interaural cues such as interaural time difference and interaural level difference are proficient in determining source azimuth, these cues perform poorly in terms of vertical localization and are out-performed by cues which make use of spectral information. Specifically it is shown that gammatone cepstral coefficients are the best performing feature for elevation estimation. A new composite feature set is introduced and its performance is compared to two state of the art localization methods. While it's shown that this novel method doesn't perform as well as the other methods in horizontal localization, it outperforms both methods with relation to vertical localization. Importantly it is shown to outperform the other methods in the mismatched HRTF condition, a finding which is the basis for the subsequent chapter.

Chapter 7: HRTF Clustering for Robust Spherical Sound Source Localization

In the penultimate chapter of the thesis, the adaptive affinity propagation algorithm is introduced and is used to cluster similar HRTF sets together based on their performance in terms of vertical localization using the novel method described in chapter 6. In doing this it is shown that HRTF clustering results in more robust performance in terms of localization in the mismatched HRTF condition and in the presence of noise. Finally, these HRTF clusters along with anthropometric data are used to determine the significance of pinna features in terms of vertical localization. This reveals the importance of the cavum concha width as an anatomical feature which informs our sense of vertical sound source localization.

Chapter 8: Conclusions

The final chapter in the thesis discusses the contribution of the novel work presented while outlining the possible directions for future work.

1.3 Contribution of Thesis

The work described in this thesis can be summarised by the following list:

- A comprehensive examination of binaural features and their application to sound source localization tasks.
- A novel method of performing sound source localization in binaural signals using machine learning.
- An improvement of source localization in the mismatched HRTF condition using HRTF clustering.
- A machine learning based metric of anatomical feature importance for vertical localization
- An investigation into headphone comparison techniques and the accuracy of their performance.
- An investigation into the quality of a variety of ambisonic microphones in capturing three dimensional sound fields.

1.4 Publications

Elements of the work described in this thesis have appeared in the following publications:

Conference Publications:

- C1: "Machine Learning for Sound Source Elevation Detection" by Hugh O'Dwyer, Enda Bates, and Frank Boland, in *Proceedings of the Workshop on Intelligent Music Production*, Huddersfield, UK, September 2018
- C2: "A Machine Learning Approach to Detecting Sound-Source Elevation in Adverse Environments" by Hugh O'Dwyer, Enda Bates, and Frank Boland, in *Proceedings of the Audio Engineering Society Convention 144*, Milan, Italy, May 2018
- C3: "Virtual Headphone Testing for Spatial Audio" by Hugh O'Dwyer, Marcin Gorzel, Luke Ferguson, Enda Bates, and Frank Boland, in *Proceedings of the Interactive Audio Systems Symposium*, York, UK, September 2016

- C4: "A Study in Machine Learning Applications for Sound Source Localization with Regards to Distance" by Hugh O'Dwyer, Sebastian Csadi, Enda Bates, and Frank Boland, in *Proceedings of the Audio Engineering Society Convention 146*, Dublin, Ireland, March 2019
- C5: "A Recording Technique for 6 Degrees of Freedom VR" by Enda Bates, Hugh O'Dwyer, Karl-Philipp Flachsbarth and Frank Boland, in *Proceedings of the Audio Engineering Society Convention 144*, Milan, Italy, May 2018
- C6: "Comparing Ambisonic Microphones — Part 1" by Enda Bates, Marcin Gorzel, Hugh O'Dwyer, Luke Ferguson, and Frank Boland, in *Proceedings of the Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*, Guildford, UK, July 2016
- C7: "Recording and Composing Site-Specific Spatial Music for 360 Video" by Enda Bates, Sebastian Csadi, Hugh O'Dwyer, Luke Ferguson, and Frank Boland, in *Proceedings of the Audio Engineering Society Convention 146*, Dublin, Ireland, March 2019
- C8: "Calibration of Digital Sound Projectors with Scene Uncertainty" by Luke Ferguson, Enda Bates, Hugh O'Dwyer, Sebastian Csadi, and Frank Boland, in *Proceedings of the Audio Engineering Society Convention 146*, Dublin, Ireland, March 2019
- C9: "Direct to Reverberant Ratio Measurements in Small and Mid-Sized Rooms" by Sebastian Csadi, Frank Boland, Luke Ferguson, Hugh O'Dwyer, and Enda Bates, in *Proceedings of the Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, York, UK, March 2019
- C10: "Comparing Ambisonic Microphones — Part 2" by Enda Bates, Sean Dooney, Marcin Gorzel, Hugh O'Dwyer, Luke Ferguson, and Frank Boland, in *Proceedings of the Audio Engineering Society Convention 142*, Berlin, Germany, May 2017

In C1 and C2 the author designed machine learning algorithms capable of estimating the elevation of a binaural speech signal. In C6 this algorithm operated as a logistic classifier predicting one of 10 classes of elevation using a combination of interaural and monaural cues. This algorithm could predict the correct class of elevation angle to within $\pm 10^\circ$ in reverberant environments with an accuracy of approximately 80%. In C5, a similar algorithm was developed which used non-linear regression to give an estimate of source elevation in a binaural signal. This algorithm performed better than the previous one with an average accuracy of 85% when estimating to within $\pm 10^\circ$ across 3 separate reverberant environments.

In C3 the author conducted a series of virtual headphone comparison tests to evaluate the consistency of each test. Two sets of testing stimuli were generated by first recording a set of audio signals onto a binaural head microphone using different headphones (*recorded methodology*). The second set of test stimuli were obtained by capturing the frequency response of the test headphones and applying them to the raw audio test signals (*filtered methodology*). The author also designed a MUSHRA type experimental procedure for the subjective evaluation of these two test methods.

In C4 the author designed a machine learning algorithm capable of estimating the distance of a binaural speech signal to within 0.5 metres up to a distance of 4 metres. This required the measurement of a series of Binaural Room Impulse Responses (BRIRs) at varying distances in a variety of spaces.

In C5 the author performed a directional analysis on recordings made using a novel multi channel microphone technique and reproduction system intended to support six degrees of freedom of listener movement. This method utilizes four Ambisonic microphones in a near-coincident arrangement with a 50cm spacing. The results of this analysis indicated that there are some inconsistencies in the on and off-axis response, but suggest that the technique can potentially support six degrees of freedom in a recorded audio scene using a compact microphone array that is well suited to Virtual Reality (VR), particularly in Free View Point (FVP) applications.

In C6 and C10 the author successfully implemented the Directional Audio Coding (DirAC) method of spatial sound representation and analysis in MATLAB [155]. This code was used to perform an analysis of the quality of directional encoding in B-format recordings made using various ambisonic microphones. The author recorded these audio signals within a 360° 16-channel loudspeaker array. The results of this DirAC analysis formed the results of the objective analysis of the quality of the ambisonic microphones under test. A subjective analysis of the audio quality of the microphones was assessed using a series of MUSHRA tests. Some of this testing was carried out by the author. In C8, the same objective analysis was performed as in C10 but with one additional microphone. The subjective testing in C8 however incorporated head rotation via a virtual reality headset which allowed the spatial quality of the audio recordings to be analysed as well as the general quality.

In C7 the author assisted the lead author in recording site-specific spatial music for 360° video. For this project, a piece of music composed for saxophone quintet was recorded in four different locations using a Soundfield ST450 Mark 2 microphone system. For exterior locations, a Rycote Windjammer and windshield were also used. This experimental work investigated the influence of distance, space, movement, and spill on spatial music.

In C8 the author assisted the studies lead author in deriving the mathematical formulae necessary to extend the Image Source Method (ISM) of digital sound projector calibration

to account for room geometry uncertainty. This resulted in the foundations of a novel method for calibrating digital sound projectors referred to as the Probabilistic Image Source Method (PRISM).

In C9 the author made measurements of Room Impulse Responses (RIRs) and Binaural Room Impulse Responses (BRIRs) in three rooms of various sizes. These measurements were used by the papers lead author to develop a measurement of Direct to Reverberant Ratio (DRR) from BRIRs. This method presents a trend between DRR and source distance that is similar to that of a DRR measurement taken from an omnidirectional RIR.

In addition to these works, it is hoped that the results shown here in chapters 6 and 7 will be published in an academic journal shortly.

Chapter 2

Perception of Spatial Audio

This chapter introduces some of the fundamental concepts that underpin the research reported in this thesis. Beginning with an overview of the human auditory system, the mechanisms which allow us to process sound are described. These include interaural cues such as the time of arrival of sound and the difference in sound intensity perceived at both ears, monaural cues such as spectral interference caused by the pinna, head, and torso, and dynamic cues related to the motion of a sound source relative to a listener's head. The influence that factors such as sound source direction, distance, and space can have on the perception of sound waves can be examined using the Head-Related Transfer Function (HRTF) and its time domain equivalent, the Head-Related Impulse Response (HRIR). The measurement of HRTFs and HRIRs are described in this chapter along with an examination of how they are influenced by different localization cues. The chapter concludes with a section on how the human auditory system can be modelled through auditory scales and auditory filter models.

2.1 The Human Auditory System

The Human Auditory System (HAS) is made up of a complex network of internal and external organs which allow us to perceive sound. It can be broken down into three major sections: the outer, middle and inner ear. A diagram of the entire human auditory system can be seen in Figure 2.1. The outer ear consists of the ear canal and the external cartilaginous flaps - known as the pinna. As the pinna is the only external part of the HAS, it is commonly referred to as the ear. However, this is incorrect as it constitutes just one part of the ear with the remaining sections located internally. The pinna collects incoming sound waves and transmits them towards the ear canal. While some sound waves directly enter the ear canal, others are reflected off the folds and curves of the pinna resulting in constructive and destructive interference of the sound waves. This translates

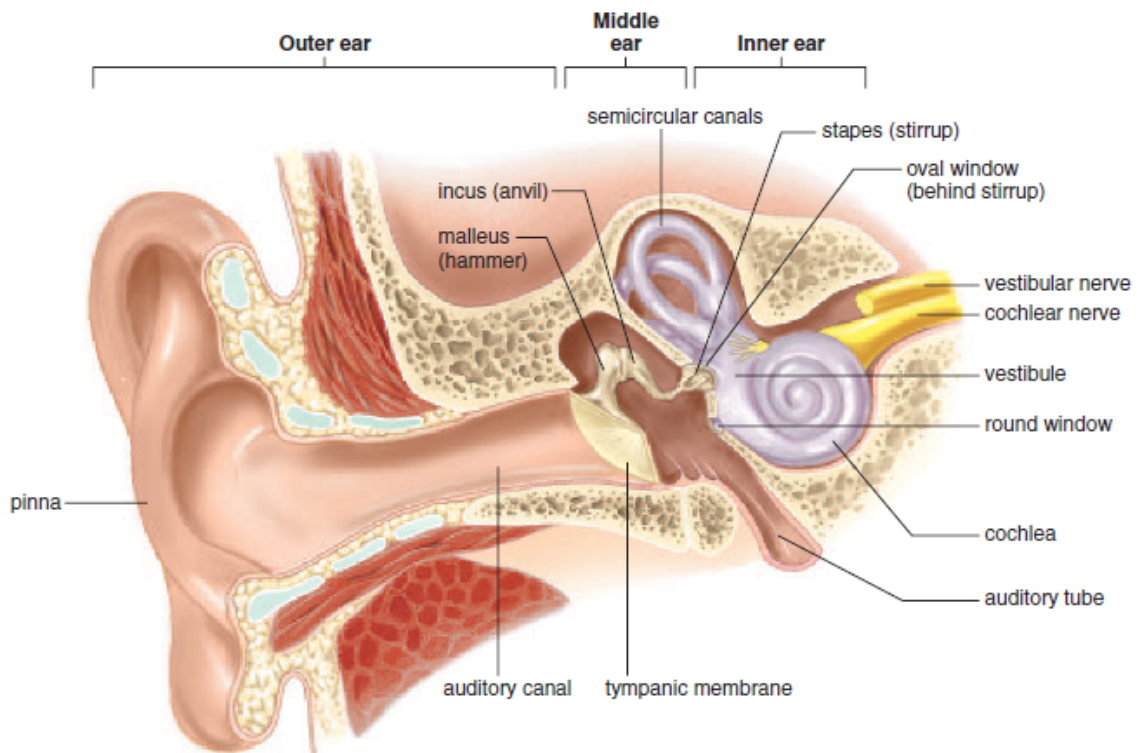


Fig. 2.1 The anatomy of the ear, reproduced from [128]. Sound waves travel through the ear canal and vibrate the eardrum (tympanic membrane). These vibrations are transmitted to cochlea via three tiny bones. Depending on the frequency of the sound, different parts of the cochlea are excited, sending messages to the brain.

to amplification and attenuation of particular frequencies, allowing filtering and spectral shaping of the incoming sound to occur. As the shape of each pinna is unique to the individual, sound is perceived differently by the ears of each person. It should also be noted that on average the male pinna is larger than that of the female [106]. This frequency dependent amplitude modulation provides useful cues to the individual as to the location of a sound source, particularly in the vertical plane [2]. Additional amplitude modulation occurs in the ear canal in the range of 3kHz to 12kHz, the frequency range used for speech communication [37]. The ear canal has an average width of approximately 0.7 centimetres and an average length of approximately 2.5 centimeters [51].

After travelling through the ear canal, sound waves are passed onto the middle ear via the tympanic membrane or eardrum. The membrane vibrates with the oncoming waves and these vibrations are passed through to the air-filled tympanic cavity. Here lie the ossicles, three tiny bones which act as a lever transmitting the low-pressure vibrations of the eardrum onto a smaller membrane called the oval or round window. The ossicles are the smallest bones in the human body and are individually called the malleus (hammer), incus (anvil) and stapes (stirrup). The action of the middle ear acts as a sort of mechanical

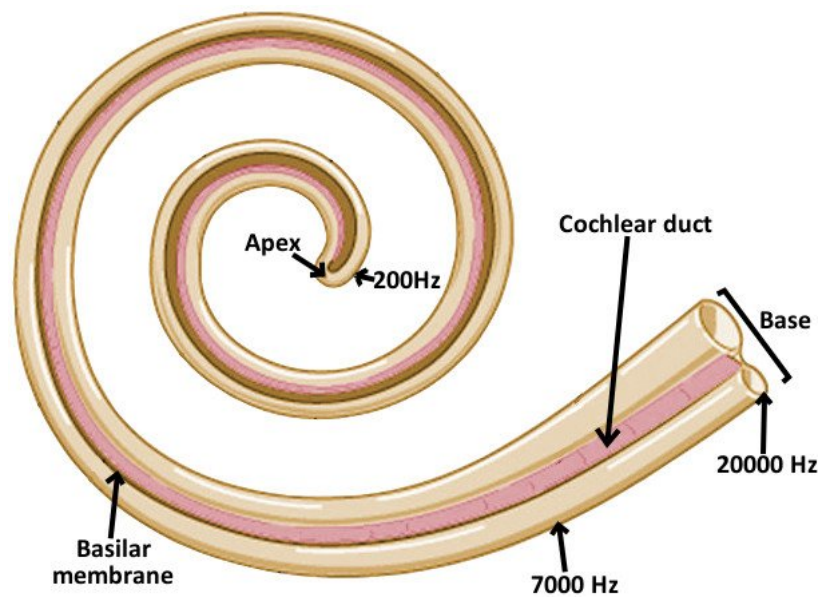


Fig. 2.2 A diagram of the Human Cochlea, reproduced from [170]. The Basilar Membrane (BM) is located inside the cochlea between the scala media and the scala tympani. The BM is moved when sound waves fall on the ear. Different frequencies cause different parts of the BM to move, high frequencies resonate near the base of the cochlea while low frequencies resonate towards its apex.

amplifier or impedance transformer, ensuring that the sound waves can be interpreted by the fluid filled inner ear whose impedance is much higher than that of air.

Beyond the oval window is the inner ear comprised principally of the cochlea, a coiled tube with an approximate length of 3.5 centimeters when unrolled. It contains three fluid-filled chambers - the scala vestibuli, scala tympani and scala media, which are separated by two membranes - the basilar membrane and Reissner's membrane. A diagram of the cochlea can be seen in Figure 2.2. The Basilar Membrane (BM) separates the scala media and scala tympani and is lined with tiny hair cells which resonate with particular frequencies. At its base, the membrane is stiff and narrow becoming flexible and wider toward its apex.

The movement of the stapes in the oval window changes the pressure in the fluid of the scala vestibuli. This in turn changes the pressure along the basilar membrane in response to different frequencies which create peaks in displacement along the membrane at different locations. The increments in frequency along the length of the membrane are approximately logarithmic meaning that humans have a greater resolution of tones at lower frequencies while higher frequencies are progressively more coarse. The hair cells which line the organ of corti have a set of cilia organelles on their surface which move most at the point where the basilar membrane is excited by the pressure waves. The movement of these hair cells is then transcoded into neural signals which are passed onto the brain.

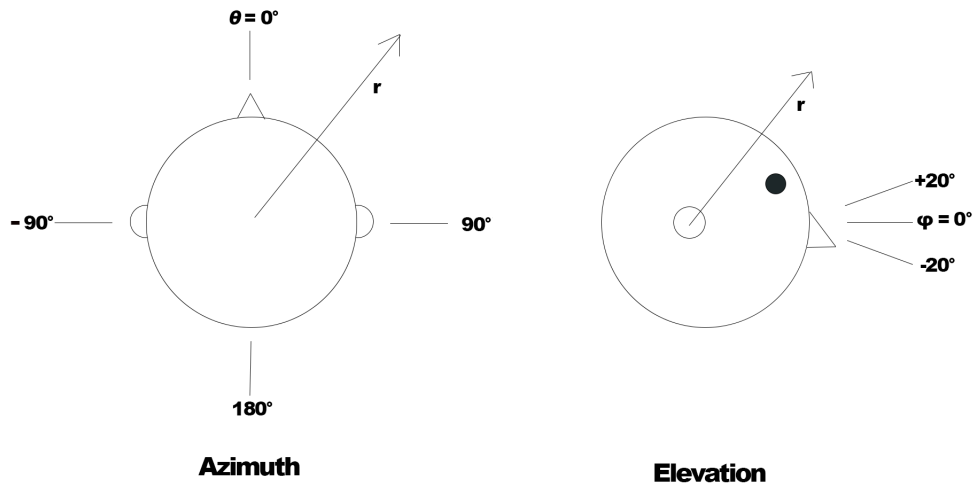


Fig. 2.3 The Head Related Coordinate System used in this thesis. θ represents the horizontal angle relative to the centre of the listener. ϕ represents the vertical angle relative to the centre of the listener and r represents the distance of the sound source relative to the centre of the listener

2.2 Sound Source Localization

Beyond interpreting the loudness, pitch, and timbre of sound, the human auditory system is also capable of locating a sound source in three-dimensional space. Our ability to localise sound is dependent on a variety of interaural (between two ears) and monaural cues which can often be subtle and sometimes contradictory. When reproducing auditory scenes it is crucial that these cues are preserved to create a truly realistic and immersive auralisation. We rely on different variations of cues whether localization is performed in the horizontal plane, vertical plane, or in terms of distance. These cues are also employed differently depending on whether the source is stationary or in motion. Our accuracy in localising sound varies upon many factors but most importantly it depends on the Direction of Arrival (DOA) of the sound. Early studies in sound source localization were summarised in Jens Blauerts' essential text 'Spatial Hearing - The Psychophysics of Human Sound Localization' first published in 1974 and later translated to English in 1983 [19, 20]. Including both studies by Blauert himself and other early works by spatial audio pioneers, the book details many psychoacoustical phenomena related to spatial hearing.

2.2.1 Head Related Coordinate System

At this point it is necessary to introduce a head related coordinate system to help define the terms azimuth and elevation. A diagram of the head-related coordinate system used in

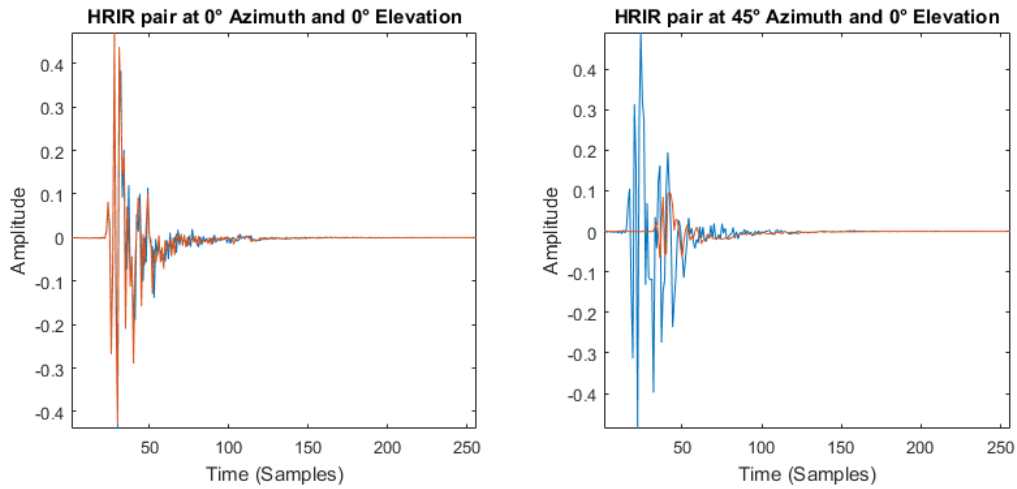


Fig. 2.4 HRIR pairs from subject 2 of the SADIE II database [8]. Left HRIRs (orange) and right HRIRs (blue) for 0°Azimuth and 0°Elevation (Left) and for 45°Azimuth and 0°Elevation (Right)

this thesis can be seen in Figure 2.3. The origin of this coordinate system is taken to be the central position of the head directly between the left and right ears. Source positions can be described relative to the head such that $s = (r, \theta, \phi)$. Here, r represents the distance between the source and the centre of the listeners head. θ is the horizontal angle of the sound source relative to the front of the listener's head. ϕ is the angle between the source and the horizontal plane, i.e. the location along the vertical plane (elevation).

2.2.2 Head Related Impulse Response

The way in which sound is perceived at the ears is influenced not only by the environment through which the sound travels but also by the anatomy of the individual. Sound wave reflections from our head, shoulders, torso, and the folds of our pinnae are unique to each person and as such, sound waves present themselves differently to each individual [2]. The influence that an individuals anatomy has on a sound wave can be captured through what is referred to as a Head-Related Impulse Response (HRIR). The HRIR is a time-domain filter characterised as an acoustic impulse (Dirac delta signal). HRIRs are typically measured in left and right pairs and are a measure of the times at which reflections arrive at the ear based on the initial arrival of a direct sound wave. Examples of measured HRIRs can be seen in Figure 2.4.

HRIRs are usually captured under anechoic conditions using real subjects or dummy head microphones. When measuring HRIRs of human subjects, small in-ear microphones are placed at the entrance of both ear canals which are typically blocked (blocked *meatus*) to avoid any additional resonance. Dummy head (or binaural) microphones have microphone

capsules built into each ear canal which capture sound similarly to how it is perceived by human ears. HRIRs are typically measured by playing a short impulsive sound such as a balloon pop through a loudspeaker at a particular location relative to the centre of the head. The recording of this sound at the ears is then representative of the spectral and temporal effects caused by both the individual and their position relative to the source. However, it has been shown that a better response can be captured using deconvolutional [53]. In this method, an exponentially swept sinusoidal tone is played through a loudspeaker and recorded using in-ear microphones or a dummy head microphone. The recording is then deconvolved with the original swept sinusoidal signal to obtain an impulse response. This method eliminates non-linear distortions in the audio signal path and subsequently the final filters. A major downside to this method however is that it takes much more time, meaning that the measurement of an entire HRIR database could potentially take hours. This is also problematic as HRIR measurement requires subjects to remain absolutely still and even the slightest adjustment during the playback can render the measured HRIR incorrect. There is a large body of research suggesting solutions to this problem via simplification and parallelisation, i.e. by assuming head symmetry and only recording HRIRs for one side of the head and swapping left and right HRIRs to make up the opposite side [121, 156, 215]. Another way to simplify this process is to model and simulate HRIR filters. However, the results of such studies have shown an inferiority to measured HRIRs [3, 9, 71].

Convolving HRIRs with anechoic audio signals results in a binaural signal which gives a faithful impression of sound coming from a particular direction, provided that the left and right signals are delivered using headphones. Using headphones ensures that there is no cross-talk between the ear signals.

2.2.3 Head Related Transfer Function

The Head-Related Transfer Function (HRTF) is the frequency domain counterpart to the HRIR filter and is obtained by applying the Fourier transform to a HRIR filter. Examples of HRTFs can be seen in Figure 2.5. HRTFs allow for a visualisation of the spectral peaks and notches present at different frequencies in a HRIR filter. They were first introduced in a 2001 tutorial by Cheng and Wakefield in the Journal of the Audio Engineering Society (AES) [35] where their role in the synthesis of spatial sound over headphones was explored. HRTF filters contain all of the necessary cues for the localization of sound sources around the sphere surrounding the human head such as interaural level difference, interaural time difference, and the spectral cues which inform our sense of elevation. HRTF filters therefore can provide very convincing binaural signals which avoid front-back confusion and even allow listeners to determine source elevation [73]. This has only been shown to be true for listeners using HRTF filters captured through their own ears however. When using

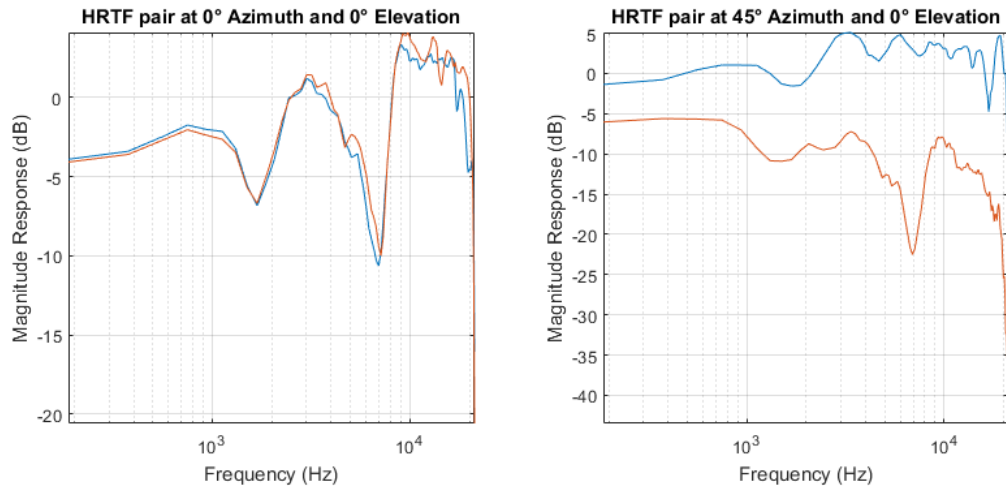


Fig. 2.5 HRTF pairs from subject 2 of the SADIE II database [8]. Left HRTFs (orange) and right HRTFs (blue) for 0°Azimuth and 0°Elevation (Left) and for 45°Azimuth and 0°Elevation (Right)

HRTF filters captured on another person, an effect similar to "listening through someone else's ears" is reported whereby subjects fail to perceive elevation and are often unable to resolve front-back confusion [200]. Kistler and Wightman showed in a 1992 study that HRTFs could be synthesised as a linear combination of five basic spectral shapes that account for approximately 90% of the variance in HRTF magnitude functions [96]. Their study showed that subjects' judgement of the apparent direction of headphone presented sound sources was almost identical for synthesised HRTFs using their model as it was for measured HRTFs.

There are a variety of freely available HRIR/HRTF databases online which have been used extensively to research the area of sound source localization. These include the CIPIC database [5], the MIT KEMAR database [60], and the SADIE database [91]. These publicly available datasets consist of a varying number of subjects and measurements. These datasets have proven to be invaluable tools in spatial audio studies and experiments. The details of some of them are as follows:

- The SADIE II dataset [8] contains 256 sample HRIRs with sampling rates of 44.1KHz - 16bit, 48KHz - 24bit and 96KHz - 24bit recorded at 2818 locations for 18 human subjects and 8802 locations for both a KEMAR binaural mannequin and a KU100 binaural head microphone.
- The SADIE I dataset [91] contains 256 sample HRIRs with sampling rates of 44.1KHz - 16bit, 48KHz - 24bit and 96KHz - 24bit recorded at 170 locations for 18 human subjects and 1550 locations for both a KEMAR binaural mannequin and a KU100 binaural head microphone.

- The CIPIC dataset [5] contains 200 sample HRIRs with sampling rates of 44.1KHz - 16bit at 1250 locations for 43 human subjects and two KEMAR binaural mannequin subjects.
- The MIT KEMAR dataset [60] contains 512 sample HRIRs with sampling rates of 44.1KHz - 16bit at 710 locations for a single KEMAR dummy head microphone.

2.2.4 Localization in the Horizontal Plane

As humans, our ability to localise sound varies considerably depending on the location of the sound as well the nature of the source. We use a number of monaural and interaural cues to perform localization. The most prevalent of these cues being Interaural Time Difference (ITD) and Interaural Level Difference (ILD). These cues are measures of the difference in time of arrival of a source at each ear and the difference in perceived volume between ears. They were first described by Lord Rayleigh in his 1907 paper 'On our Perception of Sound Direction' [159]. Here he presented a theory referred to as the "duplex theory". Rayleigh came to understand that for a sound source localised on one side of the head, the pressure of the sound wave arriving at the ear on the opposite side of the head would be significantly lower due to shadowing and absorption from the head. In addition to this, the phase of the signal arriving at the opposite ear would be different due to the additional distance travelled by the sound wave, this is referred to as Interaural Phase Difference (IPD). These phase ambiguities can be visualized in Figure 2.6. Rayleigh's work concluded that for frequencies below 1500Hz, we perform localization using ITD while we use ILD to localise sounds at higher frequencies. While this generalisation is not entirely accurate, it provided a basis for localization which furthered research in the area.

In the median plane (i.e. the plane bisecting directly in front of and behind of a listener), ITD and ILD values are approximately 0. This is because the path length that a sound must take along this plane towards both ears is equal. Sound which originates off the median plane will have to travel different path lengths to reach both ears resulting in either a positive or negative value for both ITD and ILD. If the human head is modelled as a curved spherical surface, an estimate of ITD can be calculated as follows [203]:

$$ITD(\theta) = \frac{r}{c}(\sin \theta + \theta) \quad (2.1)$$

where θ is the angle of the sound source relative to the median plane having a value of $0 \leq \theta \leq \frac{\pi}{2}$, c is the speed of sound in air (approximately $340m/s$), and r is the radius of the human head (approximately $8.75cm$). This formula is generally known as the Woodworth formula [203]. According to this formula, a maximum ITD value of $661 \mu s$ can be observed at 90° and 270° .

Sound sources which are not positioned on the median plane present much lower sound pressures to the contralateral (far-side to the source) ear than to the ipsilateral (near-side to the source) ear. This is due to sound waves being absorbed by the head, a phenomenon known as head shadowing which is most prevalent for high frequencies. ILD is thus dependent on both source position and frequency and can be defined as follows:

$$ILD(r, \theta, \phi, f) = 20 \log_{10} \left| \frac{P_R(r, \theta, \phi, f)}{P_L(r, \theta, \phi, f)} \right| (dB) \quad (2.2)$$

where P_R and P_L are the frequency domain sound pressures at the left and right ears for a source at (r, θ, ϕ) [206].

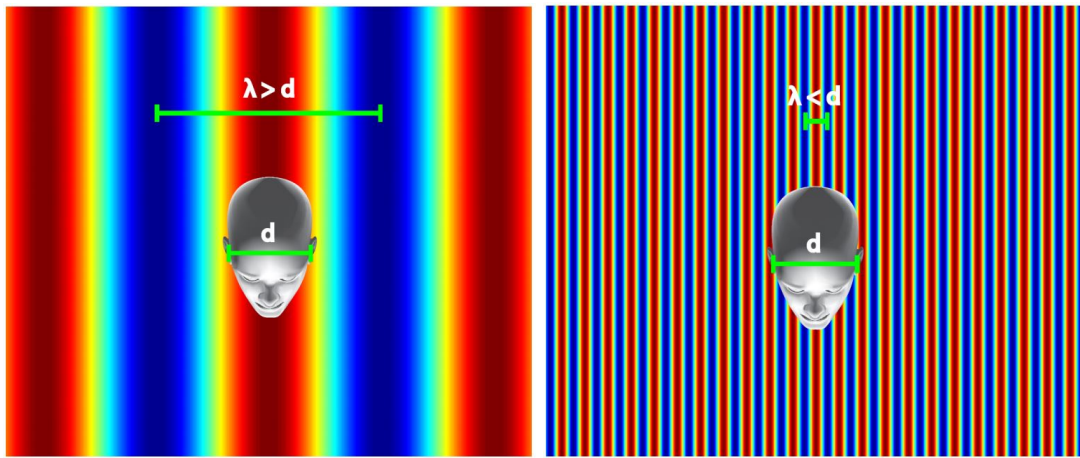


Fig. 2.6 Phase ambiguities arise when wavelength λ is significantly smaller than the diameter of the head. This occurs at approximately 1500Hz

Interaural cues are fundamental in understanding how we localise sound in the horizontal (azimuthal) plane. Studies have shown that localization error can be as low as $1^\circ - 3^\circ$ for auditory events occurring around the vertical or median plane of a listener [21]. However, localization error increases for events which occur laterally to the listener due a phenomenon referred to as the *cone of confusion*. The cone of confusion occurs as a large set of points radiating outwards from the centre of the ear produce almost identical phase delays and transient disparities. The cone of confusion is visible in Figure 2.7, causing ambiguity in both the horizontal and vertical planes. Head movement has been identified as a means to isolate the location of a sound source to a single point on the cone of confusion [192].

A phenomenon known as front-back confusion also arises from similarities in ITD and ILD cues. In this instance, when the head is in a fixed position, sound originating from directly in-front of or behind the listener is indistinguishable as both would result in ITD and ILD values of zero due to symmetry. This occurrence is also remedied using head rotation, thus altering ITD and ILD cues. This dynamic shifting of the head to

resolve sound source localization has also been shown to aid in localization in the vertical plane [151]. This hypothesis was proven in a 1931 study by Young where subjects were unable to perform head movement and were thus unable to localise vertically or resolve front-back confusion [210]. Localization then returned to normal when head movement was permitted.

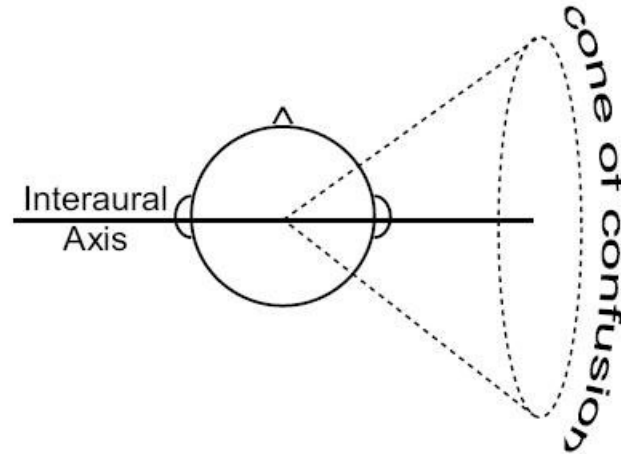


Fig. 2.7 The cone of confusion in sound source localization caused by similar values for ITD and ILD

2.2.5 Localization in the Vertical Plane

Interaural cues, while fundamental to our perception of horizontal localization have little bearing on our ability to locate sound sources in the vertical or elevated plane. Instead, we perceive elevation through spectral cues which are unique to each listener. These spectral changes are caused by reflections from the folds of our pinnae as well as from our heads and shoulders as described by Middlebrooks in [133]. As these anthropometric values can vary to a huge degree depending on the individual, it means that modelling how we perceive elevation is a difficult task. In addition to this, head shadowing causes little high frequency spectral content to be preserved at the contralateral (far-side) ear for sound sources off the median plane. Most spectral cues occur at frequencies above approximately 3kHz and are thus primarily monaural cues as described by Wightman in 1997 [202].

There are however, some phenomena which can help inform the perception of sound in the vertical plane. One such phenomenon is that of the pitch height effect as first described by Pratt in 1930 [153] who stated that *High tones are phenomenologically higher in space than low ones*. The pitch height effect describes how the higher the frequency of a pure tone is, the higher the perceived image position is, regardless of the physical height of

the loudspeaker playing it. This finding was further corroborated by Roffler and Butler in 1968 [161] who showed that for pure tones played through a loudspeaker, perceived image height remains relatively constant while the actual loudspeaker height is increased. In [160], Roffler and Butler then showed that for band-passed noise signals, high frequency components (above 7kHz) are essential for accurate vertical localisation. Cabrera and Tilley showed using octave band pink noise on the median plane that higher frequencies are required for accurate vertical localisation as can be seen in Figure 2.8. [32].

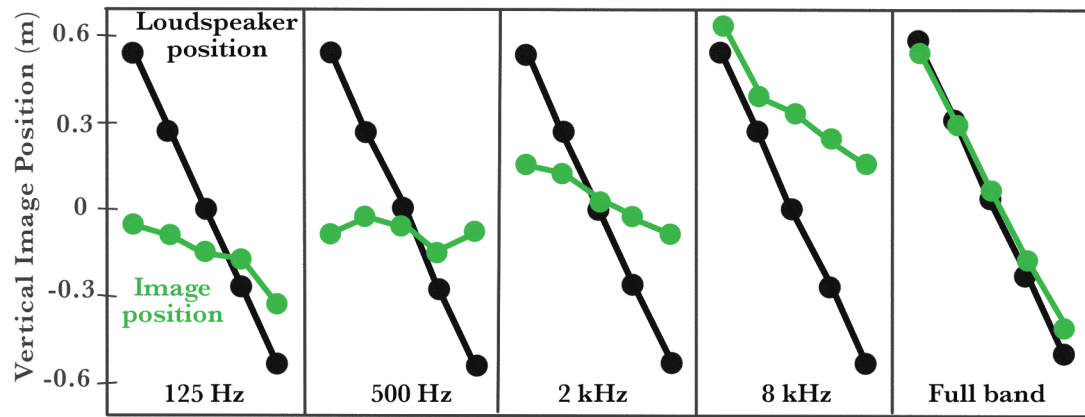


Fig. 2.8 Vertical image position compared with loudspeaker position for octave band pink noise. Reproduced after Cabrera [32].

While localization on the horizontal plane can be performed with a just noticeable difference (JND) less than 5° for sounds occurring close to the median plane, localization in the vertical plane does not have the same resolution and is generally considered poorer. A study on the perception of elevated sound sources was presented by Blauert in [20]. Here, subjects were asked to identify the position of continuous speech across the vertical plane. The results can be seen in Figure 2.9. The standard deviation for the angular source discrimination is much higher for elevated source positions than along the horizontal plane. While Blauert's study used phantom images between two loudspeakers to produce different angles, a study by Barbour showed there was relatively little difference in the perceived location of real or phantom images in elevation detection along the vertical plane [10].

A 2018 study Ege et al. on elevation perception showed how the pinna plays a vital role when localising sound sources in the vertical plane [48]. This work showed how our ability to localise sound in the vertical plane is hindered considerably when a mold is placed around the pinna, preventing reflections of the sound source from occurring. In the same study, it was concluded that horizontal and vertical sound-localization mechanisms are processed by independent neural pathways. Whereas the azimuth coordinate may be

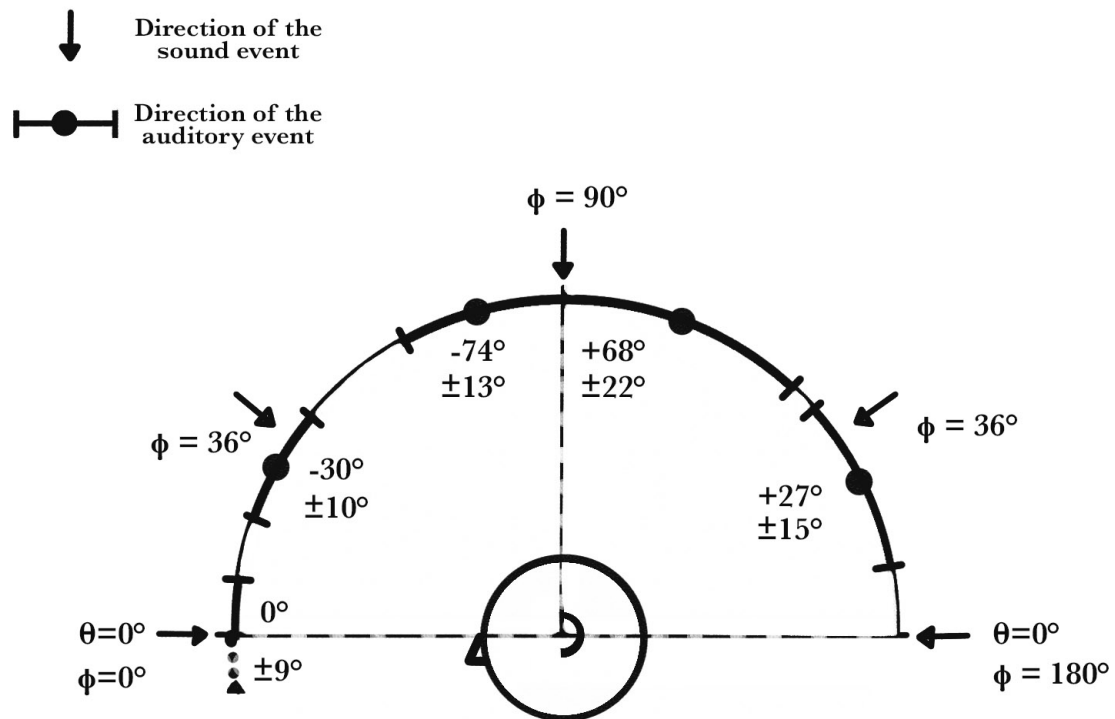


Fig. 2.9 localization blur in the median plane for a continuous speech signal, reproduced after Blauert [20]

extracted by a maximum likelihood estimation, the elevation direction appears to rely strongly on the involvement of a spatial prior, i.e. through head movement.

These findings furthered the results of a previous study by Roffler and Butler where subjects pinnae were covered with a layer of flattened plastic, leaving only a hole at the entrance of the ear canal for sound to pass through [160]. Here it was shown that subjects were virtually incapable of performing vertical localization showing that spectral coloration is in fact a vital cue in terms of elevation. Pinna effects begin to appear at frequencies around 3 kHz, this is where the wavelength, λ becomes comparable to the pinna size [14]. In addition to these pinna effects, there has been shown to be some elevation-dependent features at lower frequencies attributed to head diffraction and torso reflections [2]. In 1968, Shaw and Teranishi proposed a model describing the resonances within the cavities of the pinna and the ear canal, showing that resonant modes were formed at mid to high frequencies between approximately 3kHz and 13kHz [175]. In 1974, Shaw demonstrated that resonant peaks in the pinna moved according to the elevation of a sound source, beginning at approximately 3kHz for sources located below the listener and gradually rising towards 13kHz for sources above the listener [174].

Bloom et al. determined that a sense of elevation in audio signals can be achieved using a manipulation of high-frequency spectral cues [23]. More specifically, Kendall and Rodgers found that 4 specific features of the HRTF were determined to contribute to the sensation of elevation in audio presented over headphones, these being a peak, two notches and attenuation at high frequencies [93]. In Figure 2.10, the variation in peak resonance frequencies and high frequency notch frequencies in the HRTF with Elevation can be seen. The nulls in the HRTF rise with elevation towards a maximum notch frequency around 90 degrees approaching the zenith of the HRTF measurement.

In 2005, Raykar et al. described a method for extracting the notch frequencies created by the reflections from body parts, head diffraction, and the pinna from the HRIR [158]. This 5-step process first finds the linear prediction (LP) residual of the original HRIR to remove the interdependence among adjacent signal samples. A half-hann window of size 1ms is then applied to the residual. Following this, the auto-correlation function of the windowed LP residual is calculated and is windowed using a 0.7ms half Hann window. The log-magnitude spectrum in dB of this signal is then found and its group delay function is calculated. The local minima in the group-delay function are then taken to be the notch frequencies of the HRIR. This process can be seen in Figure 2.11.

By applying Raykar's method of notch extraction to real HRTFs, it can be seen that notches are more prominent in the contralateral HRTF, i.e. in the ear on the opposite side of the head to the HRTF measurement. This is shown in Figure 2.12. Here we see notch frequencies plotted on both left and right ears of the HRTF for 5 azimuthal planes using a single subject from the CIPIC database. At 0°, these notches are roughly equivalent

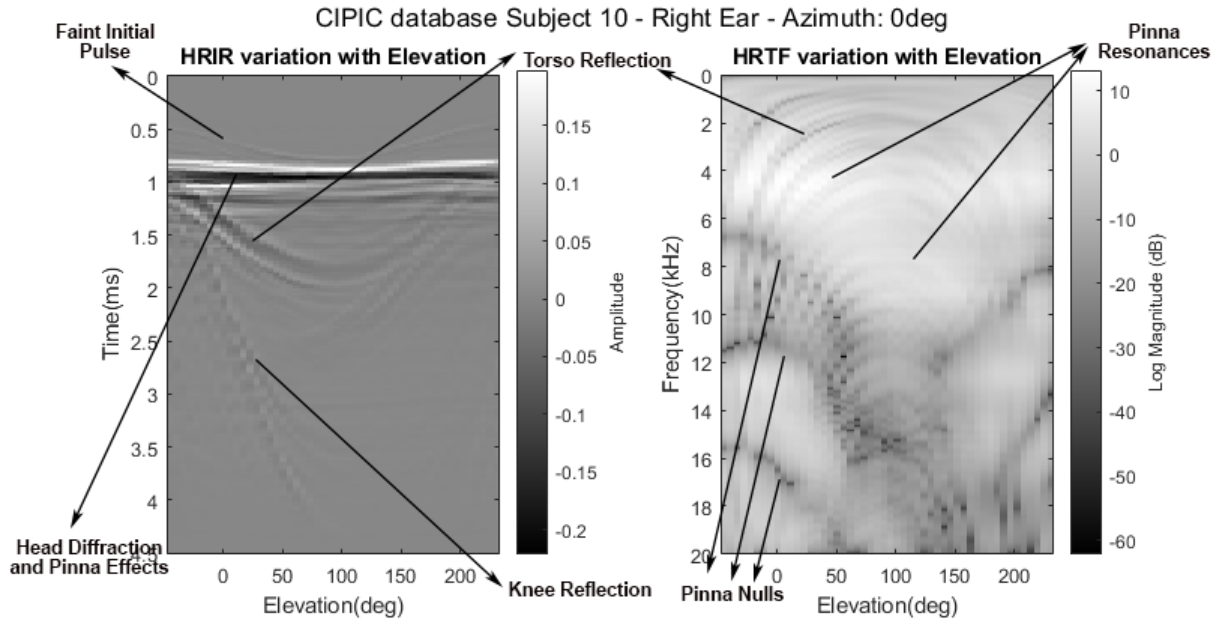


Fig. 2.10 HRIRs (left) and HRTFs (right) displayed as images showing the effect that reflections and diffraction cause on the HRIR with increasing elevation. This subsequently results in Pinna nulls and resonances in the HRTF. This particular plot uses HRTFs from subject 10 in the CIPIC database at an azimuth angle of 0° .

for both ears but as the position of the HRTF moves toward the sides of the head, the rising notch frequencies become less prominent in the ipsilateral HRTF, i.e. the ear on the same side of the head as the HRTF measurement. This signifies that for sound source localisation in the vertical plane, the contralateral HRTF could be just as important if not more important than the ipsilateral HRTF.

Lida et al. described a method whereby a blackman-harris window is ran over a measured HRIR to provide an estimate for the locations of P1, N1 and N2 [84]. They determined that P1, the first peak was mainly distributed in the mid frequency range between 3.4kHz and 4.3kHz. P1 is found in general to be constant with changes in elevation. The two notches, N1 and N2 are located at higher frequencies, between 5.7 and 9.5kHz for N1 and between 8.2 and 13.5kHz for N2. Various methods of extracting these notches have been described across different studies. These methods often incorporate a smoothing of the captured HRIR. Their study found that HRTFs selected from a database based on similar frequencies for P1, N1 & N2 performed approximately as well as the users own HRTF for elevation localization tests in frontal and rear directions. A similar study was conducted using frontal-projection headphones and using Raykar's algorithm to extract values for P1, N1 & N2 from HRIRs in the CIPIC database [182].

In 2014, Geronazzo et al. showed that these notch frequencies could be related to the specific pinna geometry of the individual whereby the path length of reflective sound

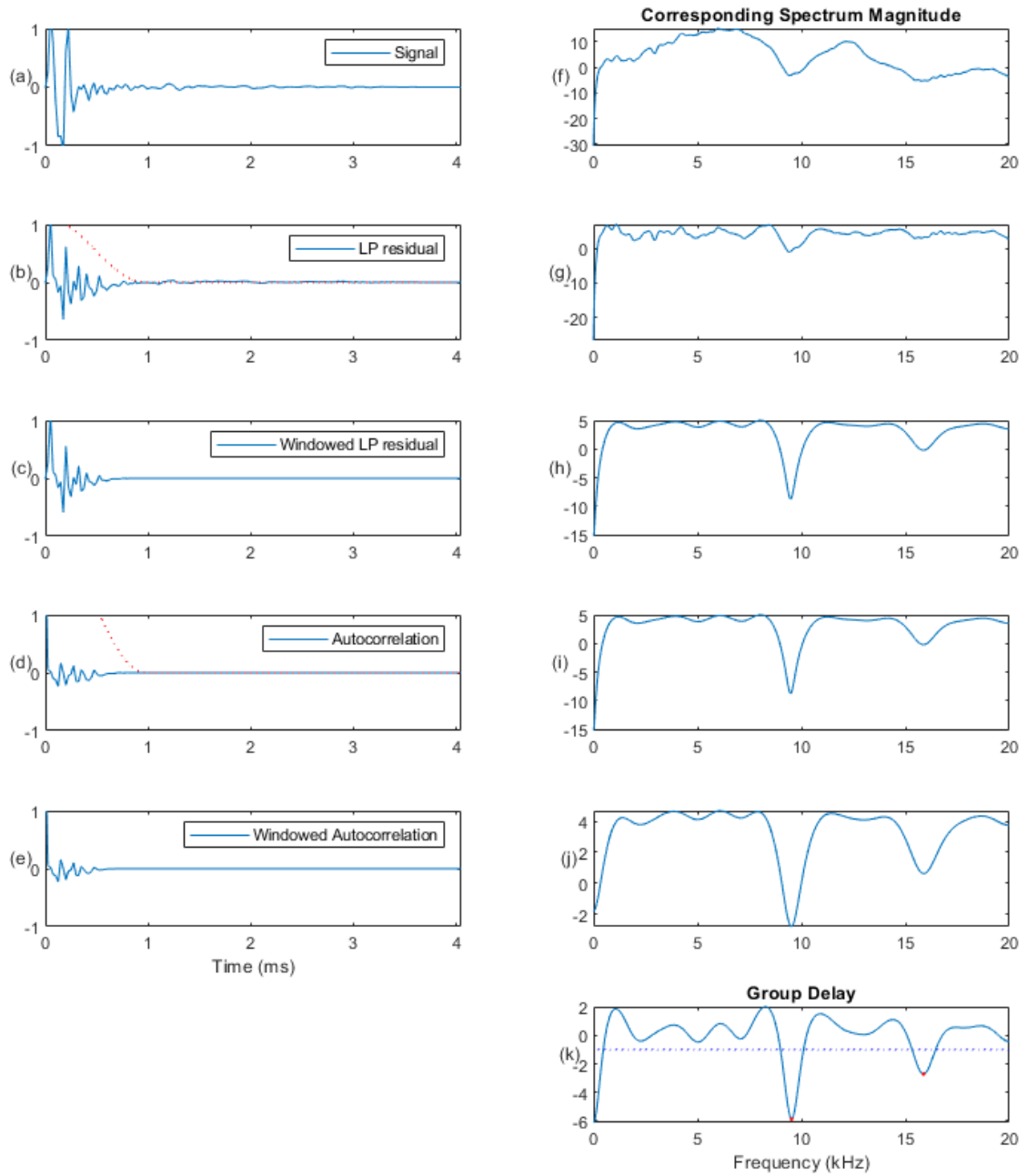
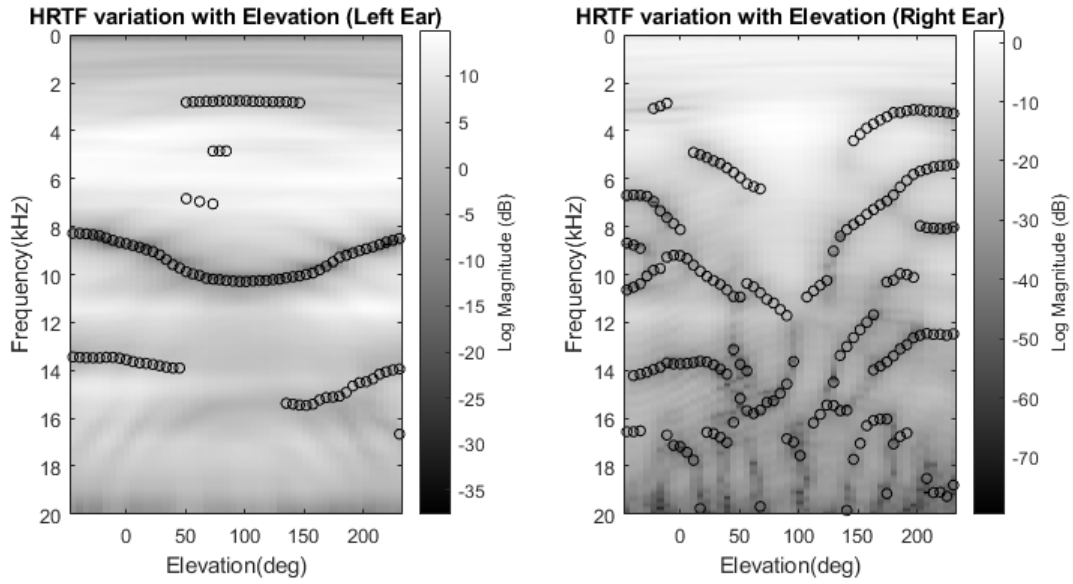
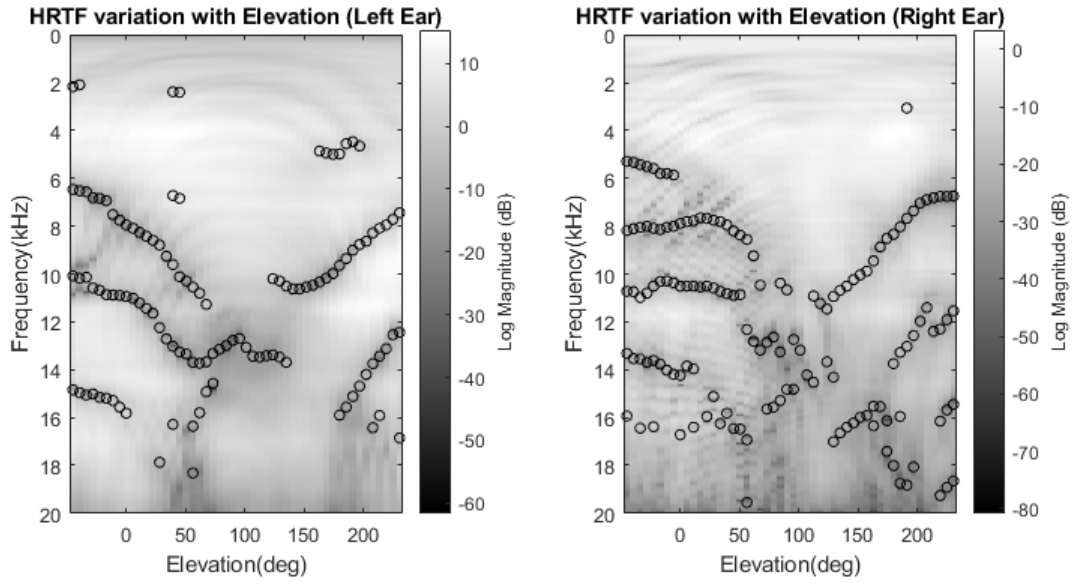


Fig. 2.11 Raykar's method for extracting the notch frequencies created by head, torso, and pinna reflections from the HRIR [158]. The HRIR can be seen in (a) through (f) as it goes through the process of notch extraction. The corresponding spectrum magnitude can be seen in (f) through (j) with the group delay employed in (k) to extract the notch frequencies.

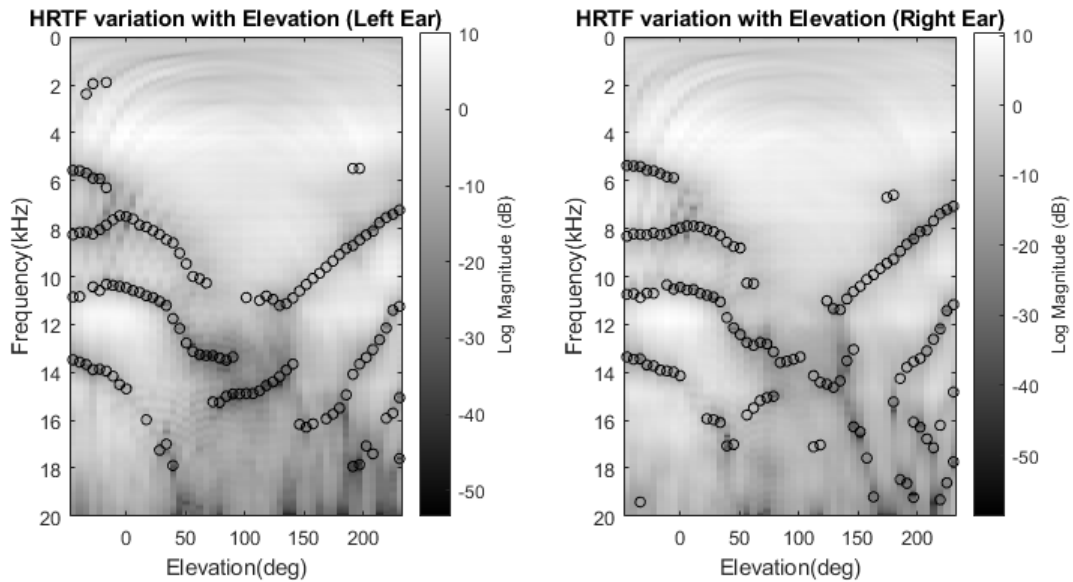
CIPIC database Subject 11 - Azimuth: -80deg



CIPIC database Subject 11 - Azimuth: -45deg



CIPIC database Subject 11 - Azimuth: 0deg



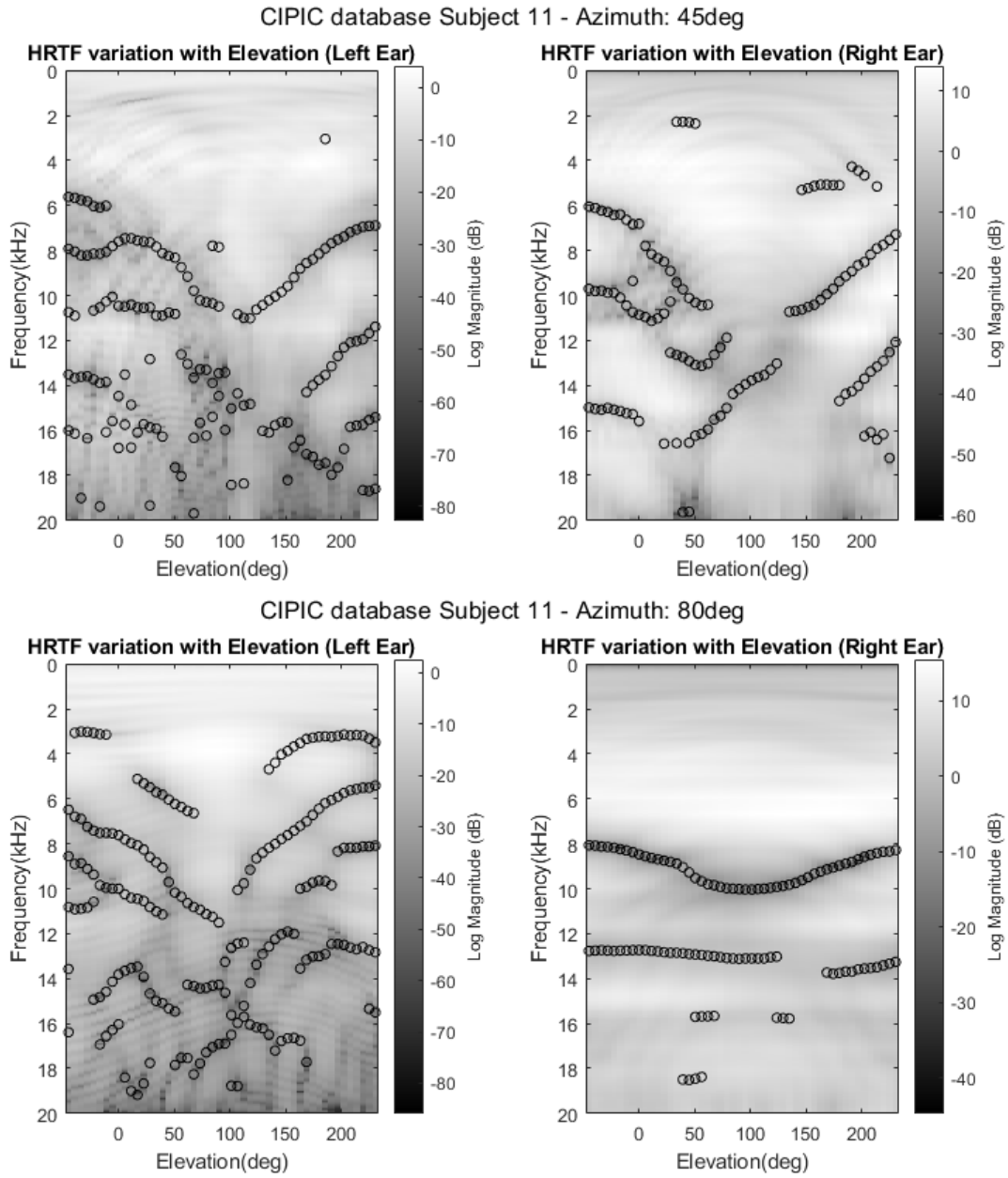


Fig. 2.12 Notch frequencies in the HRTFs of subject 11 of the CIPIC database. Notches are shown for both ears and at 5 different azimuthal planes. Notches found using Raykar's method as described in [158].

waves travelling from the helix border, concha inner wall and the concha outer wall could be used to estimate where notch frequencies would occur in the HRTF [63]. Here the notch frequencies are calculated using the formula, $f = c/2d$ where d is the path length of the reflective sound wave.

2.2.6 Distance Perception

When discussing the perception of sound it is important to make a distinction between the apparent *distance* and *depth* of a sound source. While *distance* describes the physical range between a source and a listener, *depth* describes the listener's perception of an auditory scene as a whole [162]. The perception of distance, unlike source position does not rely on interaural cues such as ITD and ILD. Rather it is influenced by factors such as intensity, spectral and binaural differences, source familiarity, and Direct to Reverberant Ratio (DRR) [98]. DRR is a particularly useful cue as it can be measured and used as a method of simulating different virtual source distances in AR and VR applications. It is defined as the ratio of Direct signal energy to Reverberant signal energy and is calculated by separating a measured impulse response into its direct and reverberant components using an integration window of a particular size. This window is usually applied from the initial onset of the impulse and is defined in [211] as:

$$DRR = 10 \log_{10} \left(\frac{(\int_{T_0}^{T_0+c} x^2(t) dt)}{(\int_{T_0+c}^{\infty} x^2(t) dt)} \right) [dB] \quad (2.3)$$

where $x(t)$ is an omnidirectional room impulse response, T_0 is the initial onset time of the impulse, and c is the chosen window size.

The choice of window size thus has a significant influence on the calculation of DRR and a variety of different sizes have been used for DRR measurements using room impulse responses and binaural room impulse responses. In [211], Zahorik used a 2.5ms window is used to extract the direct sound from BRIRs of room size $830m^3$ where the IR onset time was judged to be 20 dB greater than the noise floor. 2.5 ms is chosen in [211] as it represents the approximate duration of HRIRs measured in anechoic environments [134]. Neidhardt et al. used a 3.5ms window in a room size of $178.7m^3$, but in this case 2.5ms was applied after the peak of the direct sound, and 1ms before the peak [139]. Larsen et al. used a slightly longer window of 3ms following the direct sound onset in the BRIR in a room size of $1000m^3$ [102]. Werner et al instead used a window size of 1.5ms after the initial onset for two rooms of sizes $179m^3$ and $182m^3$ while similarly stating that this window included monaural and binaural cues resulting from reflections from ears, head and torso, but not the room [201]. In contrast, Bronkhorst et al used a window of 6ms in a room size of $37m^3$ [30]. This 6ms window size was chosen due to evidence that the temporal resolution of the human auditory system uses a similar integration window for sound localization based on the precedence effect [82]. Shinn-Cunningham used a longer window of 10ms in a room with $T60 = 450ms$, but unspecified dimensions [176]. Finally, Nielsen used a variety of manually determined window sizes ranging from 4ms, 6ms,

9ms, 10ms, 16ms length with an omnidirectional RIR, in three types of rooms; anechoic ($144m^3$), living room ($90.6m^3$), and classroom ($199.5m^3$) [140].

In a study co-written by the author, a variety of window lengths were compared when calculating DRR for a variety of rooms using BRIRs recorded using a dummy head microphone [40]. Measurements were made at varying distances from an omnidirectional sound source showing that the DRR decreases monotonically with distance regardless of the room size, except when the source is very close to some significant boundary such as a wall or other reflective surface. The results suggest that in smaller rooms, a smaller window size produces more consistent changes in DRR. In addition, a comparison of different source directivities suggests that more consistent DRR values are produced in small rooms when using a directional source, as opposed to an omnidirectional loudspeaker.

The human auditory system interprets sources differently depending on their distance. For near-field sources of less than approximately 1.6m from the listener, the perceived distance is generally overestimated while it is underestimated for distances beyond this. The relationship beyond actual and perceived distance was defined in [211] by Zahorik with the formula:

$$r^* = kr^\alpha \quad (2.4)$$

Where r is the actual source distance, r^* is the perceived distance. k is a constant with an approximate value of 1.32 and α has a value which is dependent on factors such as environmental conditions and the subjects own physical differences. On average however, α has a value of 0.4.

The most basic of distance cues we use is subjective loudness which is in general a monaural cue. A sound source with constant power emits sound pressure waves which have a magnitude inversely proportional to the distance between source and receiver. The result is a decrease of approximately 6dB in sound pressure level for each doubling of distance. This 6dB law holds true only in the free field however. A formula for sound intensity can be derived for a distance r as:

$$I(r) = \frac{p^2}{\rho_o c r^2} \quad (2.5)$$

where p is sound pressure, ρ_o is the acoustic medium density, and c is the speed of sound in that acoustic medium. It can be shown simply then that doubling a source distance results in a 6dB attenuation:

$$10\log_{10}\left(\frac{I(2r)}{I(r)}\right) = 10\log_{10}\left(\frac{\frac{p^2}{\rho_0 c (2r)^2}}{\frac{p^2}{\rho_0 c r^2}}\right) = 10\log_{10}\left(\frac{1}{4}\right) = -6.02dB \quad (2.6)$$

The variability of the power of a sound depending on the source implies that prior knowledge of the source is a strong influence on distance estimation based on loudness cues.

High frequency signals undergo attenuation when travelling through air [88]. While the attenuation of high frequency waves could thus be interpreted as a distance cue, this effect is only noticeable for sound sources that are in the extreme far field. Therefore high-frequency attenuation can be disregarded as a distance cue in indoor environments.

While distance cues are inherently monaural, changes in ILD do occur in the extreme near-field, i.e. within distances of approximately 1m of the listener [31]. This is particularly true for distances shorter than 0.5m where the pressure spectrum at each ear changes significantly.

2.3 Sound Source localization in Reverberant Environments

When considering spatial hearing within a room, it is necessary to appreciate that the sound field in which the listener is immersed will be far more complex than the simple free field case. This is because sound waves undergo many reflections off the multiple surfaces of the room which arrive at different times at the ears of the listener. Reflected sound waves can be defined according to the number of reflections the waves take before reaching the listener. That is, a wave which is reflected once is called a first order reflection while a wave reflected twice en route to the listener is a second order reflection and so on. This complex sound field however can be considered as a superposition of numerous simple sound waves such as plane waves. This collection of waves can be visualised in Figure 2.13. This is an important realisation to be made when considering the generation of convincing virtual auditory environments. When characterising how a sound source will be presented to a listener in a VAE, it's important to consider the size, shape, and materials of the surfaces as well as the positions of the source and the listener. These factors all have significant impact on the transmission of sound within a room. Such a system cannot be considered as an linear, time invariant (LTI) system as there are both non-linear and time variant factors present such as the vibration of surfaces within the room. Normally however, non-linearities are ignored in VAEs and the reflections described

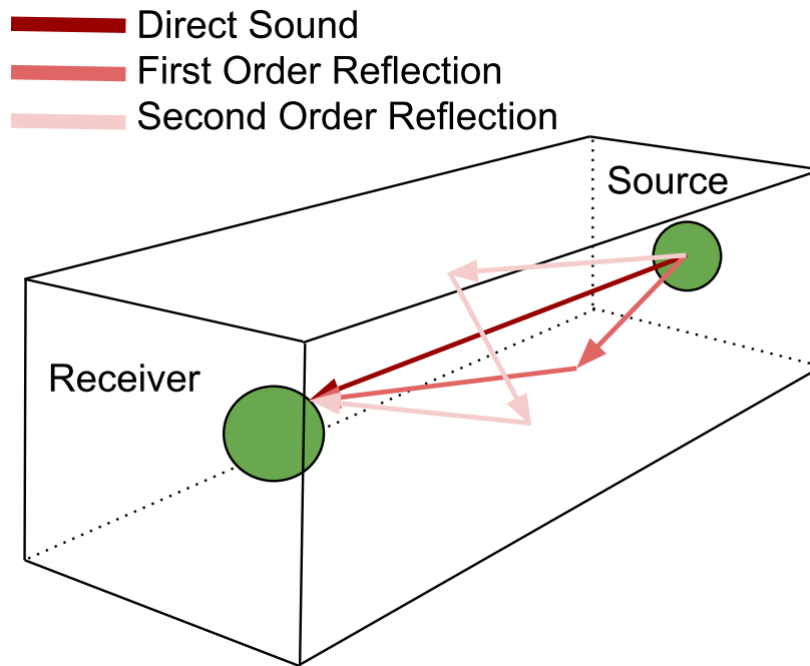


Fig. 2.13 A set of possible paths for a sound wave to take between a source and receiver in a room. These include a direct path along with a first and second order reflection

above are approximated as linear subsystems which comprise an artificial room impulse response.

2.3.1 Reverberation and the Room Impulse Response

Reverberation is the bi-product of room reflections and is the collection of multiple higher order reflections arriving almost simultaneously at the receiver. When the sound source emitted is a single impulse i.e. a Dirac delta signal, a position dependent echogram can be used to measure the magnitude and time of arrival, τ of the initial impulse and subsequent reflections. An Echogram can be generated rather easily by examining the geometry and surface materials of a room, assuming that reflections of sound waves behave similarly to rays of light and obey the same laws of reflection described by Snell's law [177].

In practice, when measuring the acoustic properties of a room, sound pressure level is measured at a receiver point which results in a time sequence known as a Room Impulse Response (RIR). Unlike an echogram, an RIR is made up of both positive and negative pressure values which reflect frequency dependent changes of phase in the signal at each reflection. The boundaries of a reflective surface are given a scalar value which quantifies the reflection (or absorption) coefficients of that surface. In real-life scenarios however, surfaces are rarely homogeneous and cannot be quantified by a single scalar value. This

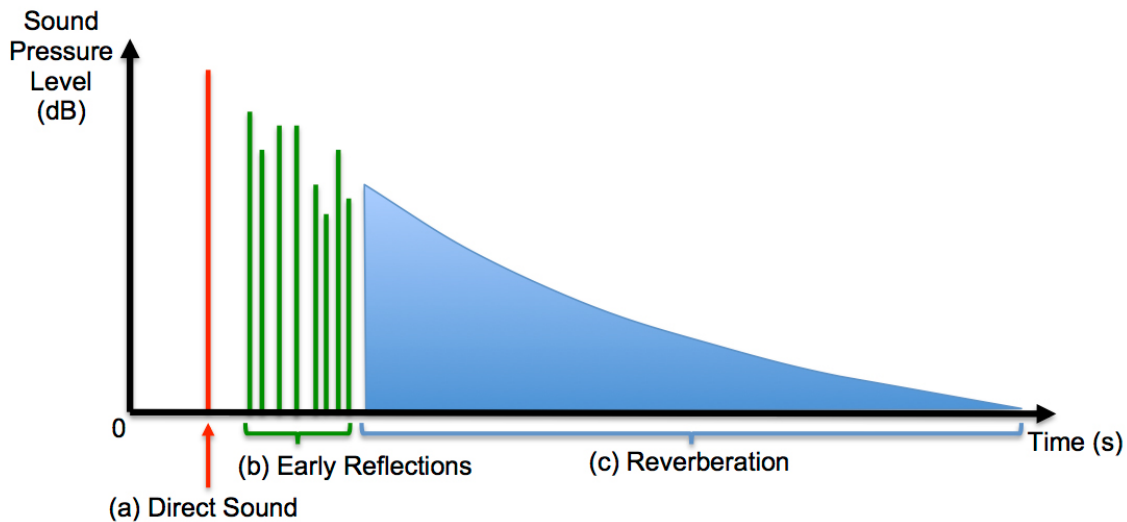


Fig. 2.14 A typical echogram of a reverberant space. A direct sound is caused by a short, impulsive signal such as handclap, balloon pop or gun-shot. Following the direct sound is a series of 1st and 2nd order early reflections. The reverberation tail is made up of 3rd order and higher reflections. The density at which these reflections arrive makes it impossible to distinguish their exact order

means that a simulation of a room acoustic will never be identical to the actual room acoustic obtained by measuring the RIR.

Both Echograms and RIRs are generally considered to consist of three main parts; the direct sound, an array of early reflections and a stochastic tail representative of reverberation. This reverberant tail occurs when the arrival of individual reflections is no longer perceivable. The transition into this diffuse reverberation is referred to as the transition period of the RIR. While there is no proven method of determining the exact transition time from early reflections to reverberation, several methods for different room geometries have been investigated such as those suggested by Jeong et al. [88], Hidaka et al. [77], and Defrance and Polack [42].

By convolving a RIR with an anechoic recording, it is generally believed that the resultant signal will sound as if the recording were made in the same room the RIR was measured in.

2.3.2 The Precedence Effect

The precedence effect is a binaural psychoacoustical effect that defines the perception of the position of a sound source in a reverberant space [110]. The term was originally coined by Wallach et al. in their classic study on localization dominance [193]. While multiple reflections of the sound source arrive in quick succession at the listeners ears, the perceived

spatial location is dominated by the location of the first arriving wave, known as the lead source. This phenomenon has been called the 'law of the first wave' or 'localization dominance'.

Summing Localization refers to a delay between 0 and 1 ms when the waves from the lead and lagging sources are perceptually fused, appearing as a singular sound event. The resultant perceived direction of the sound source is then taken to be a point between the arriving sound waves. In the simplest case of summing localization as presented by Blauert in 1997 [22] when two identical click sounds are played from different loudspeaker locations with a lag time of less than 1 ms, the sounds fuse into a singular image whose location is perceived as the half way point between the loudspeakers. Beyond 1 ms however, these sounds begin to be interpreted as separate sound events from two different directions. At this point, typically between 1 and 5 ms for transient sounds, a perceptual boundary referred to as the 'echo threshold' is reached and the fused image perceptually splits into two images. The echo threshold differs depending on the sound source however and has been shown to be as much as 40 ms for more complex sounds such as speech and music [193].

The term 'lag discrimination suppression' is used to describe the diminished perception of the parameters of the lag stimulus due to the presence of the lead stimulus. The discrimination improves as the delay increases.

2.3.3 Measurement of Room Acoustic Properties

One of the most commonly used factors to describe a room's acoustic properties is the reverberation time referred to as RT_{60} . RT_{60} is defined as the time it takes an initial broadband signal to drop in volume by 60dB after the sound emission has ceased. Reverberation time can be measured across different points in space and also across different frequency bands. In order to measure RT_{60} , a room must first be energized with noise. This is most often done using a short, loud, transient sound such as a balloon pop. It can be difficult to fully stimulate an entire room to measure RT_{60} directly, so often other methods of measuring RT_{60} are used such as T20 and T30. T20 and T30 refer to the time it takes for sound pressure level to decay by 20dB and 30dB respectively. By simply multiplying these times by 3 for T20 and 2 for T30 we can arrive at an estimate for RT_{60} . In both instances, decay time is measured following the initial first 5dB of decay. RT_{60} is usually measured at different frequencies across multiple octaves to get a more detailed frequency dependent measurement. Early Decay Time (EDT) is another metric for describing a room's acoustic properties as simply the time it takes for the signal to decay by 10dB.

While there is a directive for calculating acoustic parameters such as reverberation time in real world spaces - ISO-3383, these parameters can also be obtained based on

geometrical data. Formulae have been suggested by Eyring [50] and Sabine [163] to calculate reverberation time. According to Eyring, RT_{60} can be estimated as:

$$RT_{60} = \frac{0.161V}{S \ln(1 - \alpha)} \quad (2.7)$$

where V is the total volume of the room, S is the total surface area and α is the average absorption coefficient. When α is equal to 1 it means the surface fully absorbs the acoustic energy while an α value equal to 0 means full reflection from the boundary surface. When the absorption coefficient is small, Eyring's equation can be simplified to Sabine's equation which is:

$$RT_{60} = \frac{0.161V}{S\alpha} \quad (2.8)$$

2.3.4 Binaural Room Impulse Response

When a Room Impulse Response is measured using a binaural microphone. The response is called a Binaural Room Impulse Response (BRIR) which is equivalent to the head-related impulse response but with the added effect of the room. BRIRs can be used to simulate the effect of an environment on a sound source by convolving the source sound with a pair of BRIR signals. This practice is the basis for a convincing auralisation of audio in virtual auditory environments (VAEs) such as those found in virtual reality and augmented reality applications. There have been several databases of BRIRs made publicly available included the Aachen impulse response (AIR) database [89] and the Spatial Audio for Domestic Interactive Entertainment (SADIE) database [91]. These databases have allowed research to be conducted as to the variation of BRIRs according to a number of parameters including head size, ear shape, source location and distance.

2.4 Modelling the Human Auditory System

Earlier in this chapter we described the peripheral human auditory system, the mechanisms of which process oncoming sound waves. These manipulated sound waves are then transferred as a signal to the auditory cortex of the brain via electrical signals processed by a series of neurons. The nature in which we process sound is unique to humans and is the culmination of millions of years of evolution whereby sounds and frequencies which are useful to our survival are amplified while others are attenuated. There have been several ways in which the signal processing performed by the HAS has been modelled and in this section we will described some of models as well as giving an insight into some of the research described in later chapters.

2.4.1 Auditory Scales

When modelling the HAS, it is important to understand the way in which different frequencies are perceived by the human ear. The frequency of a sound wave is typically represented as a unit on the hertz (Hz) scale whereby a single hertz is one cycle of a wave per second. While the hertz is the derived unit of frequency in the International System of Units (SI), it does not represent the logarithmic nature of human hearing. The spacing of these frequencies along the length of the basilar membrane can be broken up into different frequency bands, referred to as *critical bands* [41]. The finite frequency resolution of the membrane means that a listener cannot distinguish between two pure tones that are too close in frequency, especially at higher frequencies [69]. When multiple tones from the same critical band are presented simultaneously to a listener it may not be possible to resolve them. This is known as spectral masking.

The logarithmic processing of frequencies along the length of the cochlea has led to the development of several mathematical scales which aim to describe this filtering based on approximation. Two such examples of these scales are the Bark scale [218] and Equivalent Rectangular Bandwidth (ERB) scale [135]. Frequency, f is related to Barks, z , as follows

$$z = \frac{26.8}{1 + \frac{1.96}{f}} - 0.53 \quad (2.9)$$

The ERB scale is related to frequency in kHz as follows

$$E = 21.4 \log_{10}(1 + 4.37f) \quad (2.10)$$

The ERB forms the basis of one of the most popular auditory filtering methods, the Gammatone filter which was designed to model the human auditory system via cochlea simulation. This filterbank transforms complex sounds into a multichannel activity pattern like that observed in the auditory nerve [149]. Gammatone Cepstral Coefficients (GTCCs) have been shown to be effective in non-speech audio recognition and classification [186]. Specifically they have been used successfully implemented in tasks involving emotion recognition [111] and cough recognition [112]. An visualisation of a gammatone filterbank can be seen in Figure 2.15.

Another popular scale for representing these critical bands is the Mel Scale which was first described by Stevens, Volkman, and Newman in 1937 [181]. The Mel Scale is based on subjective pitch perception where observers fractionated tones of 10 different frequencies at a loudness of 60 dB. The scale is related to hertz in that 1000 Hz is equal to 1000 mels. It was from this starting point that subjective measurement of perceived magnitude was measured. the unit of mels can be related to hertz, f as follow:

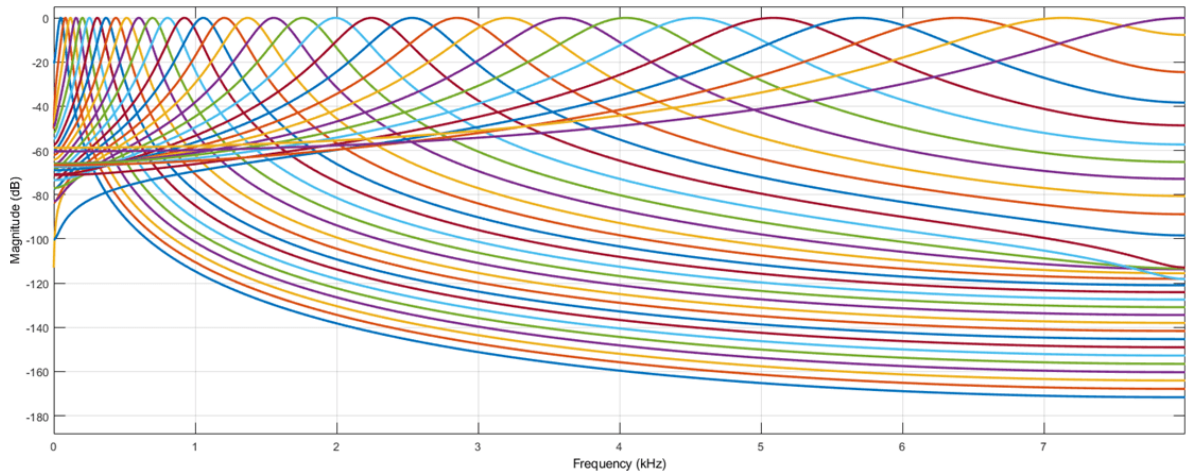


Fig. 2.15 A 32-channel Gammatone Filterbank with center frequencies ranging from 0-8kHz.

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (2.11)$$

The relationship between the two units can be seen in Figure 2.16. In recent years, the Mel scale has increased in use due to the prevalence of Mel-Frequency Cepstral Coefficients (MFCCs) in audio related machine learning applications. They are a common feature used in speech recognition and increasingly so in music information retrieval applications such as genre classification [103].

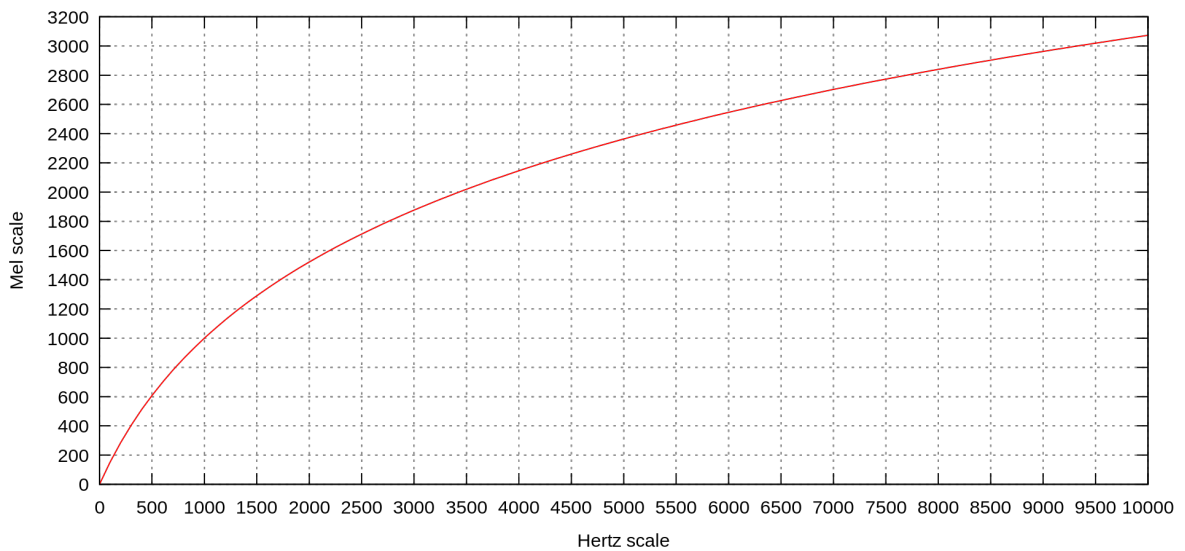


Fig. 2.16 The Mel Scale relative to the Hertz Scale

2.4.2 Auditory Filter Models

Auditory processing performed by the human auditory system is complex and is fundamentally superior to any machine-based hearing system created thus far. This is largely due to the cochlea's ability to decompose, convert and amplify sound waves non-linearly into electrical signals. The cochlea is a causal, active, nonlinear system which can be characterized by its wide dynamic range (0-120 dB SPL) and high frequency selectivity (~ 3 Hz at 1 kHz) [56, 64]. In recent years there have been several attempts to create a machine-hearing system capable of emulating the performance of the cochlea with varying degrees of complexity.

Cochlear models can be divided into one of two classes: transmission-lines (TL) and auditory filterbanks [46]. Transmission line models attempt to represent the physiology of the peripheral ear, treating the partition between the middle and inner ear as a coupled mass-spring-damper system to model wave propagation on the basilar membrane [217]. These models are computationally challenging however and require complicated differential equations in the time domain [7]. Auditory filterbank models use either parallel or cascade filters to model wave propagation on the BM. While parallel filterbank models have little regard for the physiology of the ear, cascade filterbank models take advantage of the way sound propagates in the forward direction as travelling waves in the cochlea. In the cascade of filters, each filter stage models a segment of the nonuniform distributed wave system and its output becomes the input of the next section [117].

An example of such a model is the Cascade of Asymmetric Resonators with Fast-Acting Compression (CAR-FAC) cochlear model by Richard Lyon [118]. The CAR part simulates the basilar membrane's response to sound. The FAC part models the outer hair cell (OHC), the inner hair cell (IHC), and the medial olivocochlear efferent system functions. The FAC feeds back to the CAR by moving the poles and zeros of the CAR resonators automatically. A study by Saremi et al. compared seven computational cochlear models including one cascade filterbank model (CAR-FAC), one transmission-line model, one biophysical model, and four parallel filterbank models in response to a set of common stimuli, which are used in the clinical assessment of human hearing to study their performance [166]. The results show that the CAR-FAC model exhibits an outstanding agreement with the biological data recordings at a reasonably low computational cost. This model was successfully implemented on an FPGA in [209].

2.5 Conclusion

This chapter has provided a description of the human auditory system and the cues related to the perception of sound. A head-related coordinate system has been introduced which is

consistent with the research presented in this thesis. The Head-Related Impulse Response (HRIR) as well as its frequency domain equivalent, the Head-Related Transfer Function (HRTF) have been described in the context of quantifying sound localization in the free field. The cues responsible for the localization of sound in the horizontal and vertical planes as well as with regards to distance are discussed alongside a presentation of the relevant research into these cues. Localization in reverberant environments is also discussed alongside an explanation of Room Impulse Responses and their measurement. Popular auditory scales and filter models are described. Finally, an overview of machine learning and its applications for resolving some of the difficult modelling issues around human hearing is given. These concepts are the fundamental building blocks on which this thesis is based. Furthermore, the distinction between interaural cues relating to sound source localization in the horizontal plane and monaural spectral cues responsible for localization in the vertical plane has been made.

Chapter 3

Reproduction of Spatial Audio

Spatial audio is the term given to any audio delivered in a non-mono format. It is ubiquitous in everyday life from the music we listen to on headphones to the large multi-channel speaker arrays used in cinemas. In the previous chapter, head related filtering and spatial sound localization were discussed as well as the effects that an environment can have on a sound. When reproducing spatial audio, these factors are important to preserve as they are integral to achieving a convincing auralisation. There are many methods of reproducing spatial audio ranging from 2-channel stereophony to higher order Ambisonics. In this chapter, an explanation of some of these methods and how they are employed is given. Beginning with 2-channel stereophony, detailed descriptions are given for multiple reproduction techniques including Vector Base Amplitude Panning (VBAP), digital sound projection, and Ambisonics. A novel method for optimising soundbar calibration in rooms with uncertain geometry is given known as the PRobabalistic Image Source Method (PRISM). Finally, a detailed description of how spatial audio is delivered via headphones is given with reference to headphone design.

3.1 Stereophony

Stereophonic audio systems are those comprised of more than one channel of audio information. Although the term "stereophonic" is most often associated with systems of two loudspeakers, a stereophonic audio system can refer to any system with $N \geq 2$ loudspeakers. As shown in Figure 3.1, Stereophonic sound systems are capable of producing "Phantom" sources between two loudspeakers by adjusting the level and time of arrival of the source from each speaker. Stereophony is used in most commercially available surround-sound systems including the familiar 5.1 home theatre setup.

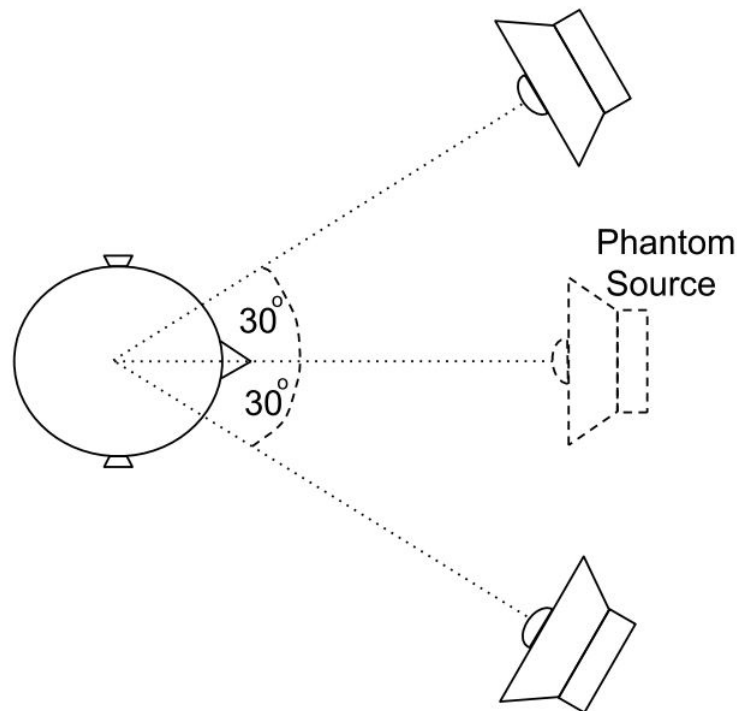


Fig. 3.1 2-channel setup for stereophonic sound reproduction

3.1.1 2-channel Stereophony

Creating a virtual image of a non-existent speaker is done by adjusting the signals arriving at each ear to recreate the natural phenomena of ITD, ILD, or both. This means that either the level of the signals or the time in which they arrive at the listener's ears are altered. This is referred to as either intensity-based stereophony in the case of ILD manipulation or time-based stereophony when ITD is altered. When stereophony is applied to a set of loudspeaker signals, the correct virtual image can only be observed at a particular point known as the "sweet spot". This is assuming that the gain of each loudspeaker is equal and the listener is positioned at an equal distance from both speakers. Movement from the sweet spot towards or away from the speakers changes the level differences and the phase of the signals arriving at the ears. This gives rise to the 'precedence effect' whereby the stereo effect collapses towards the speaker closest to the listener.

Using different microphone arrays, signals for left and right stereo loudspeakers can be recorded which can naturally capture the differences in intensity and time of arrival of a sound source. Various microphone techniques were first described in work presented by Fletcher [57] and Blumlein [25] in the 1930s. In [57], Fletcher describes a microphone technique commonly referred to as the A-B method. Here, two omnidirectional microphones are placed approximately 0.5 - 1m apart as shown in Figure 3.2. This separation allows the microphone signals to capture any difference in phase, loudness, or

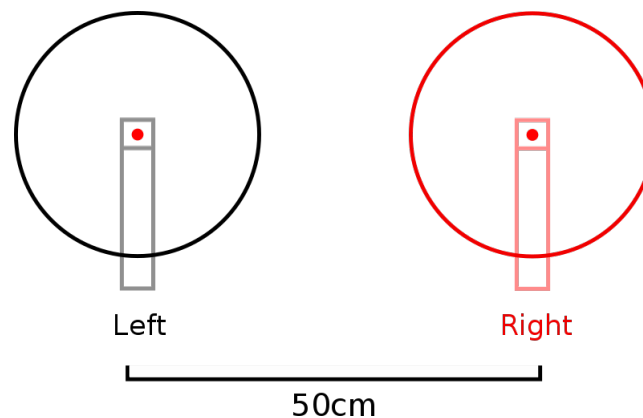


Fig. 3.2 The AB microphone technique by Fletcher. This technique aims to capture the time differences between two microphone signals

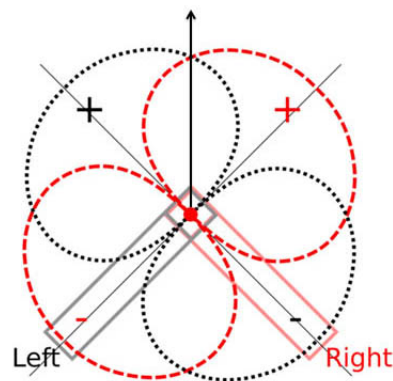


Fig. 3.3 The Blumlein microphone technique. This coincident pair aims to capture amplitude differences between two microphone signals

time of arrival for oncoming sound waves. There are limitations to this method however as signals arriving at the 2 microphones can cause destructive interference if the phase is approximately 180° out of sync. It is also not possible to present signals recorded using the A-B method in mono.

Blumlein's microphone techniques in contrast were solely focused on capturing the difference in sound level for a pair of microphones. As a result, the techniques he developed, known as coincident microphone techniques, were less problematic. The original Blumlein coincident microphone technique consisted of two bi-directional microphones positioned as closely together in space as possible, separated at a 90° angle. This can be seen in Figure 3.3. A difference in amplitude is then guaranteed as one microphone will consistently be on-axis while the other is off-axis to oncoming sound waves.

There are two popular adaptations of the original Blumlein technique. The first of which is the XY coincident pair which is the same as the initial Blumlein technique

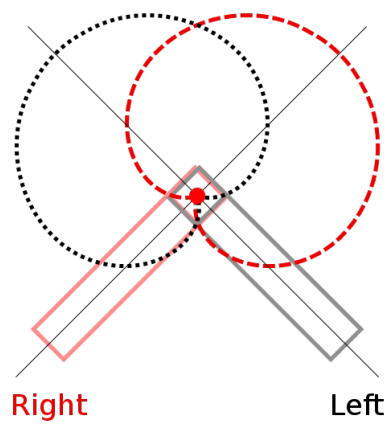


Fig. 3.4 The XY coincident microphone setup

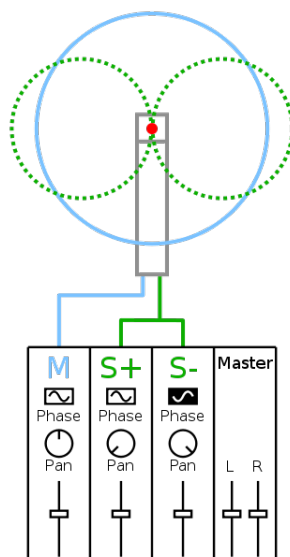


Fig. 3.5 The Mid-Side (M-S) coincident microphone setup

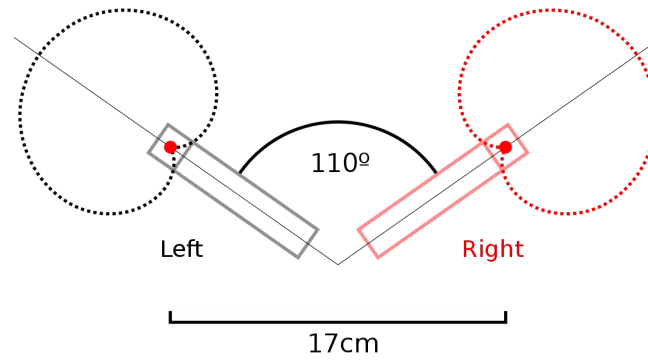


Fig. 3.6 The near-coincident ORTF microphone configuration

however, with the use of two crossed cardioid pattern microphones instead of bi-directional ones. The XY coincident microphone setup can be seen in Figure 3.4. This means that less of the reflected acoustic energy, or ambience is recorded in the signal. The second popular adaptation of the Blumlein technique is the Mid-Side pair configuration, which places a cardioid microphone (Mid or M) facing forward and a bi-directional microphone (Side or S) directed towards the sides. This setup can be seen in Figure 3.5. This means that the area of rejection is directly behind the microphone array. The stereophonic image of an M-S pair is then created using a simple decoding matrix:

$$\begin{aligned} L &= M + gS \\ R &= M - gS \end{aligned} \tag{3.1}$$

where L and R are the loudspeaker signals, M is the Mid microphone input, S is the Side microphone input and g is a positive scalar gain. Here the Left channel signal is a summation of the signal from both microphones while the right channel signal is summation of the Mid microphone and the phase-inverted Side microphone. The summation of the M and S microphones creates a cardioid like pattern similar to that observed in the XY microphone technique. The gain value g can be adjusted to balance the amount of Side mic content observed in the MS output. As the Mid microphone is generally facing the source, the Side mic captures the reverberant signal from the space around the source.

In addition to the spaced and coincident microphone techniques already discussed are near-coincident techniques. The most popular near-coincident mic technique is the ORTF (Office de Radiodiffusion Television Francaise) method which can be visualized in Figure 3.6. Here, two cardioid microphones are placed approximately 17cm apart and are separated at an angle of 110° . This is to simulate the typical distance between left and right ears and the typical angle of each ear relative to the other. This in general creates a wider stereophonic image than in the case of XY recordings.



Fig. 3.7 The KEMAR dummy head microphone from GRAS [66]

The dummy head microphone is a popular way to record binaural audio. This type of microphone aims to simulate the human ear as realistically as possible in creating stereo recordings. In general, dummy head microphones contain two omnidirectional microphones mounted inside the ears of a realistic model of a human head. The KEMAR mic from GRAS as seen in Figure 3.7, goes one step further to include an accurate model of the neck and torso. This allows for measurements of reflections from the body to be made when making recordings resulting in a very realistic auralisation. These type of microphones are often used when testing new technology in the field of telecommunications.

3.1.2 Multichannel Stereophony

The development of multichannel surround formats has been spearheaded by companies such as *Dolby* and *DTS* who have brought immersive audio experiences to a wide variety of audiences. Multichannel loudspeaker formats can be broadly classified as "surround sound" and can be most easily identified in 5.1 home speaker setups. *ITU 5.1* specifies a loudspeaker configuration typically observed in home environments consisting of 5 surround loudspeakers and a sub-woofer carrying low-frequency effects (LFE) [171]. The 5.1 loudspeaker setup can be seen in Figure 3.8 where the listener is positioned directly in front of a central speaker with Front-Left and Front-Right speakers at 30° from the centre and Back-Left and Back-Right speakers at $110^\circ \pm 10^\circ$ from the centre and behind the listener.

In addition to the 5.1 system, other more complex multi-channel stereophonic configurations are used in cinemas, theatres, installation spaces and at home. Traditional cinema

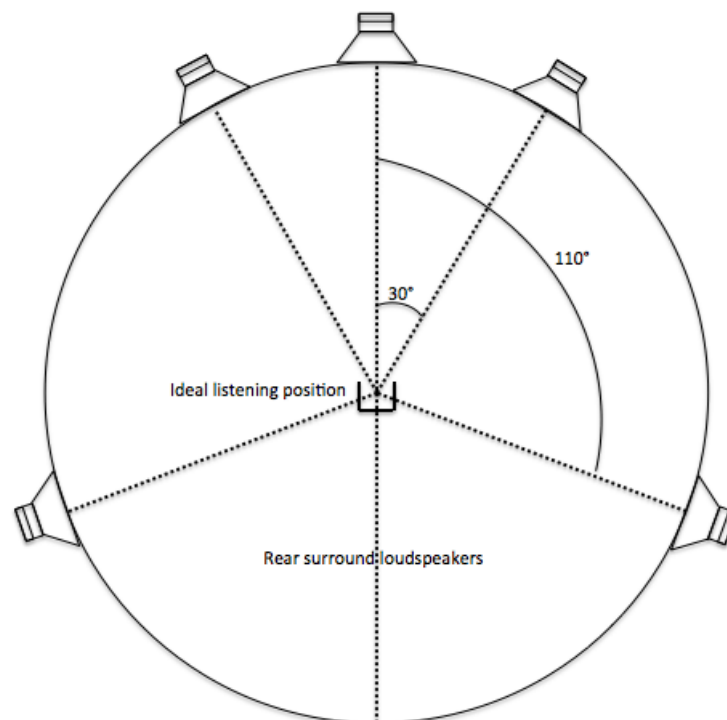


Fig. 3.8 Recommended 5-channel configuration for the *ITU 5.1* surround sound reproduction

theaters make use of 7.1 systems where multiple speakers can be carrying signals from the same channel. Dolby Atmos expanded on traditional surround sound technology by adding height channels. A similar system was also developed in Aura-3D. These systems greatly increase the sense of immersion experienced by encompassing a listener with elevated sound sources. These systems use different encoding and decoding methods to support different layouts and to provide high-quality playback and fidelity. They are also hindered by the same complications as other phantom source based systems in that an ideal listener position or sweet-spot can only be attained at a particular point. Therefore movement or a listener position far off the sweet-spot gives rise to the precedence effect which causes the stereo image to collapse. This is why in some applications such as cinema, there is an increase in the number of same-channel loudspeakers in order to enlarge the area of the sweet-spot to include more listeners. In general, however, when mixing audio for multi-channel loudspeaker systems, important information such as dialogue is kept in mono and only played from the centre speaker so as to ensure it's audible for all listeners. Phantom sources in multichannel stereo are typically panned in a pairwise manner just like in two-channel stereo systems. This method is known as Pairwise Constant Power Panning (PCPP) [38].

3.1.3 Vector Base Amplitude Panning

Vector Base Amplitude Panning (VBAP) is a vector formulation of the pairwise panning law [154]. The pairwise panning law states that any signal of equal amplitude and phase that is played in both channels of a stereo system will increase in loudness by up to 6.02dB, provided that there is a perfect response in the loudspeaker system and perfect acoustics in the environment. VBAP allows for the inclusion of another spatial dimension (elevation) by extending stereophony to include loudspeaker arrangements which add height to the configuration via loudspeaker triplets. In 2-D VBAP, the position of the listener relative to the loudspeakers constitutes two unit length vectors, I_1 and I_2 . For 3-D VBAP, an additional unit length vector, I_3 is present. A phantom source is then positioned by multiplying the unit vectors by an appropriate set of gain factors, g_i where i is the number of the loudspeaker. A linear combination of the resulting scaled vectors creates another vector, $\mathbf{p}$ pointing in the desired direction of the phantom source. This system is shown in Figure 3.9.

So, in a generalised 3-D case, the phantom source vector $\mathbf{p}$ can be expressed as:

$$\mathbf{p} = g_1 I_1 + g_2 I_2 + g_3 I_3 = \mathbf{g} \mathbf{L} \quad (3.2)$$

where $\mathbf{g} = [g_1 \ g_2 \ g_3]$ and $\mathbf{L} = [I_1 \ I_2 \ I_3]^T$. To obtain gain coefficients $\mathbf{g}$ for a particular phantom source $\mathbf{p}$ one solves:

$$\mathbf{g} = \mathbf{p} \mathbf{L}^{-1} \quad (3.3)$$

VBAP can be reformulated for any number of loudspeakers in almost any configuration. When an arbitrary loudspeaker configuration is used resulting in different varying distances from the speakers to the sweet-spot, the system should be calibrated so that the initial time delays and level differences from the loudspeakers correlate to the desired sweet-spot. As with other stereophonic systems, localization blur can occur with VBAP depending on the phantom source position and the loudspeaker separation. Difficulties also arise with tonal coloration when changing the number of loudspeakers.

3.2 Digital Sound Projection

While multichannel surround formats have enjoyed some prevalence in home settings, the practicalities of installing a 5.1 surround sound system in ones sitting room are often challenging. Correctly positioning 6 loudspeakers can prove problematic if not impossible, not to mention unsightly. A recent development in multichannel systems is that of soundbars. Soundbars consist of an array of closely packed loudspeakers capable of performing

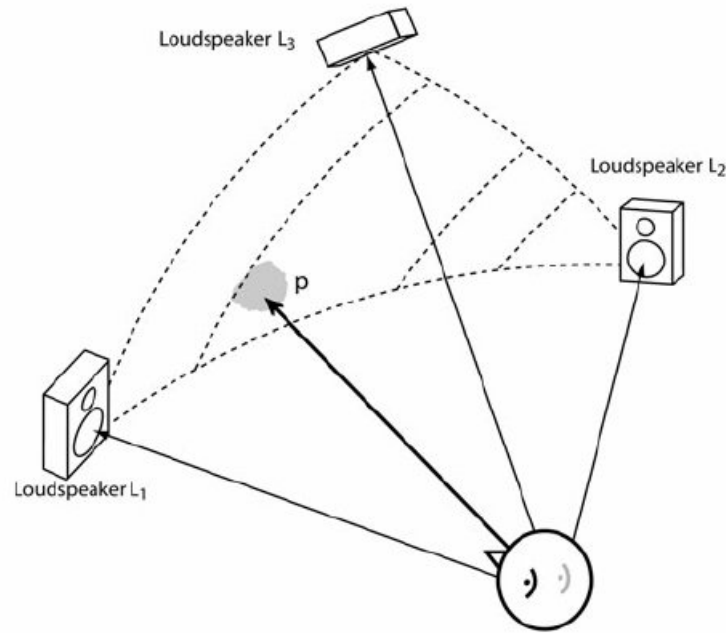


Fig. 3.9 3-channel stereophonic configuration reformulated with vectors, i.e. VBAP

digital sound projection through first and second order reflections. Projection is achieved by focusing beams of sound energy along existing reflection paths in a scene, between the source and the receiver, to provide an impression of surround sound from a front-only array of loudspeakers.

Soundbar speaker arrays are linear and horizontally aligned, this allows for multiple beams of focused sound energy to be guided to arbitrary directions along the horizontal half-plane in front of the array [81]. The direction, intensity and width of each sound beam can be controlled by manipulating the delay and gain terms of each loudspeaker. Virtual loudspeakers can be created on the surface of a reflector where beamforming delivers a focused sound energy. This artificially gives the impression of loudspeakers distributed around the space much like in the case of multichannel stereophonic systems. The process of beamforming to obtain a 5.1 virtual signal through a digital sound projector can be visualized in Figure 3.10.

Beamforming is a signal processing technique which has traditionally been used in radio telescopes and communications while also having applications in under-water military endeavours such as in towed-arrays. For the purpose of this thesis, we will be concerned only in terms of its use in audio signal processing. It is used to process spatio-temporal data from multiple sensors and can be used to identify the location of sound sources and moving objects as well as room characterisation and geometry estimation [86]. The theory can be applied to loudspeaker arrays for the synthesis of acoustic plane waves, and, both

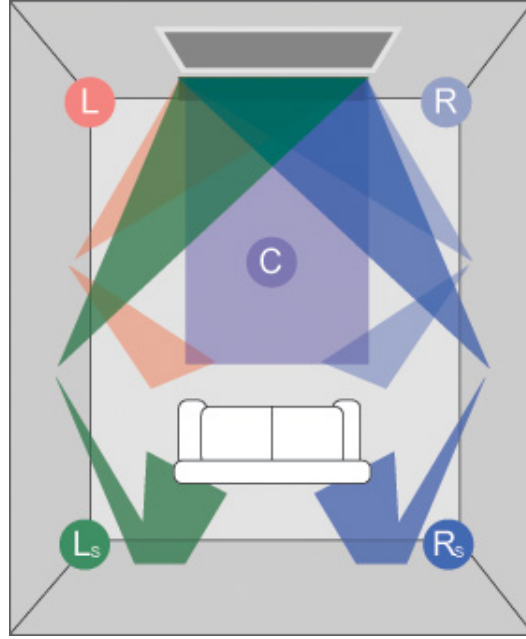


Fig. 3.10 Beamforming from array of loudspeakers creates virtual loudspeaker positions to emulate a 5.1 surround sound setup

convergent and divergent beams. In soundbar technology, delay and sum beamforming is applied whereby signals are sent to the loudspeaker array with linearly increasing time delays. This linearity causes approximate planar waves to be emitted into the surrounding environment at an angle specified by a constant lag between two neighbouring loudspeakers. For the synthesis of a convergent beam, the time delay applied to successive array elements is not linearly increasing, but instead specified by the travel time (i.e. the distance) from the array element to the desired point of convergence.

Consider the loudspeaker array shown in Figure 3.11. The desired point of convergence, P , is parameterised by its polar coordinates, (r, θ) , relative to the centre of the array, S . To acquire the delay and gain terms to be applied to each loudspeaker, the distances, $r_1, \dots, r_{N_s}$, from each of the speaker elements to P are required. The N_s loudspeakers shown are uniformly distributed along the line of the array, with a spacing of Δ_S units between adjacent loudspeaker elements, the position of which is generally given by:

$$S_n = S + \Delta_S \left(n - \frac{N_s - 1}{2} \right) \vec{X}_S \quad (3.4)$$

where $\vec{X}_S$ is a unit vector in the direction of the speaker placement in the linear array, and $n = 1, \dots, N_s$. By choosing a reference frame with its origin at the centre of the array, the position of the n^{th} loudspeaker is given by:

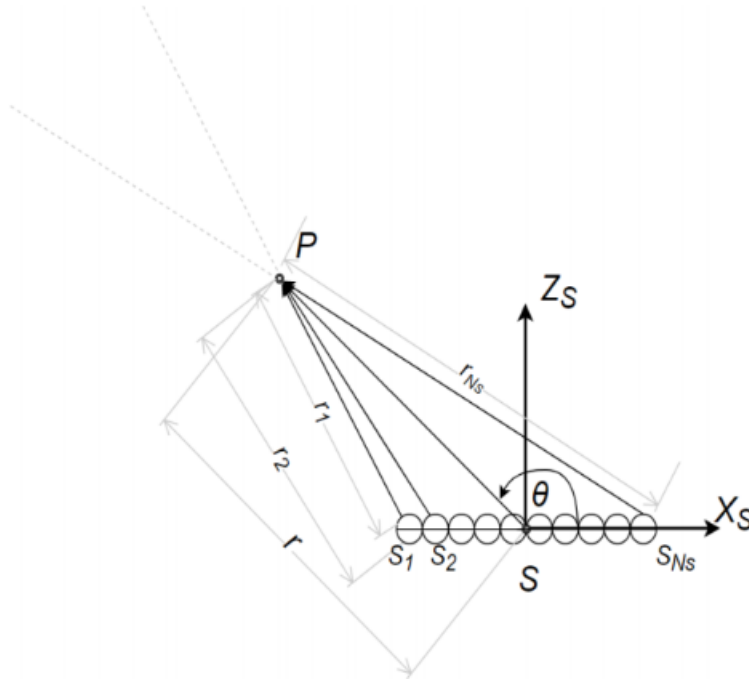


Fig. 3.11 Convergent Audio Beamforming with a line array

$$S_n = \left[\Delta_S \left(n - \frac{N_s - 1}{2} \right) \quad 0 \quad 0 \right]^T \quad (3.5)$$

The focal point is given by:

$$P = \left[r \cos(\theta) \quad 0 \quad r \sin(\theta) \right]^T \quad (3.6)$$

where r is the length from the center of the array to the focal point. Thus, the relevant distances are given by:

$$r_n = |P - S_n| \quad (3.7)$$

The delays, τ_n , applied to each loudspeaker are derived from the distances stored in the vector, $\underline{r}$, where $\underline{r} = [r_1 \dots r_{N_s}]^T$. To preserve causality and to ensure minimum buffer size, the delays are defined as:

$$\tau_n = \frac{\max(\underline{r}) - r_n}{c} \quad (3.8)$$

where c is the speed of sound.

3.2.1 The Image Source Method (ISM)

The Image Source Method (ISM) is a method commonly employed in soundbar technology to simulate realistic audio for rectangular rooms [6]. The ISM as shown in Figure 3.12 was developed in 1979 by Allen et al. It replaces reflections from a single source in a room with new sources which are placed in corresponding locations in a set of "mirrored rooms". This mirroring process is performed symmetrically about each boundary of the room and can be repeated to generate higher order reflections. This allows for the location of the "reflected sources" to be quickly and easily calculated. By manipulating the delay and gain terms for each loudspeaker in a digital sound projector, the direction, intensity and width of a sound beam can be controlled. Many beams can be generated simultaneously and a spatial sound image can be presented, by generating virtual loudspeakers on the surface of reflectors, to artificially create the impression of loudspeakers distributed around the room as per ISM.

Fundamental to ISM is that any scene object, be it the source, receiver, or boundary can be reflected or mirrored about a planar surface to give its image, which geometrically encodes the timing and directionality of the reflected signal. Higher order reflections may be reflected many times before arriving at the receiver. These can be computed recursively by successfully computing images of the result from the previous stage. The efficiency of the method was further advanced in 1988 by Lee et al. to compute higher-order reflections without the need to store previously computed reflections [104]. ISM is also a useful method for predicting reverberation time [108] and room impulse responses [107]. ISM based RIR estimation is most suited to polyhedral shaped spaces as mirroring about boundaries in these environments is simpler. It has been shown that ISM can be extended to incorporate diffraction in order to model RIRs in irregular places [183]. More recently, more work has been done to extend ISM to process spherical harmonics [164]. As shown in Figure 3.12, ISM can be used to compute relevant information about a reflected wave such as its direction of arrival and travel time.

For a scene with N_W reflectors, the maximum number of m^{th} - order reflections possible is given by

$$N_m = N_W(N_W - 1)^m - 1, m \geq 1 \quad (3.9)$$

and the maximum total number of possible transmission paths, including the direct path, using up to M^{th} -order reflections is given by:

$$N_M = 1 + \sum_{m=1}^M N_m = 1 + N_W \sum_{m=1}^M (N_W - 1)^m - 1 \quad (3.10)$$

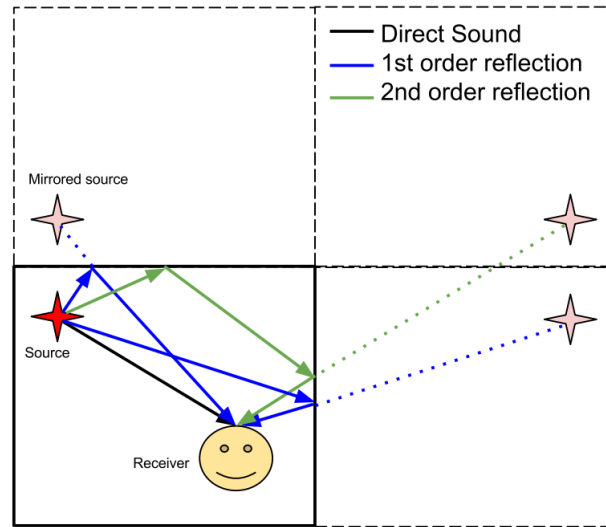


Fig. 3.12 Lattice with mirror-image sources used for calculating early reflections in a simple RIR model. The dashed lines represent the natural path of the reflections while the solid arrow tipped lines represent the actual paths between source and receiver. Here blue represents 1st order reflections and green represents 2nd order reflections

N_m is a theoretical limit, but typically above $m = 1$ there are some reflection paths that are computable but do not manifest real reflections due to obstructive barriers and other geometric constraints.

3.2.2 The PRobabilistic Image Source Method (PRISM)

A 2019 paper by Ferguson and co-written by the author presents a novel method for the calibration of a digital sound projector to an uncertain scene description [55]. This is referred to as the PRobabilistic Image Source (PRISM) Method. It is an adaptation of the ISM which can be used where knowledge of the scene is expressed with probability. It is applicable in any instance where the reflectors are parallel or orthogonal to one another. With the calibration of a digital sound projector in mind, the PRISM method can be used to reflect the image of the receiver position.

PRISM is applicable in applications where there is uncertainty in the perpendicular distance between the source or receiver and the reflectors in the scene as seen in Figure 3.13. Distributions for the location of virtual receivers are implicit in the given uncertain scene geometry. The scene is analyzed to give a range of likely virtual receiver positions along with their associated probabilities. Instead of a point estimate for the virtual receiver location, PRISM gives a distribution, which can be interpreted to give probability distributions for the beamforming parameters of a digital sound projector. As ISM already has a

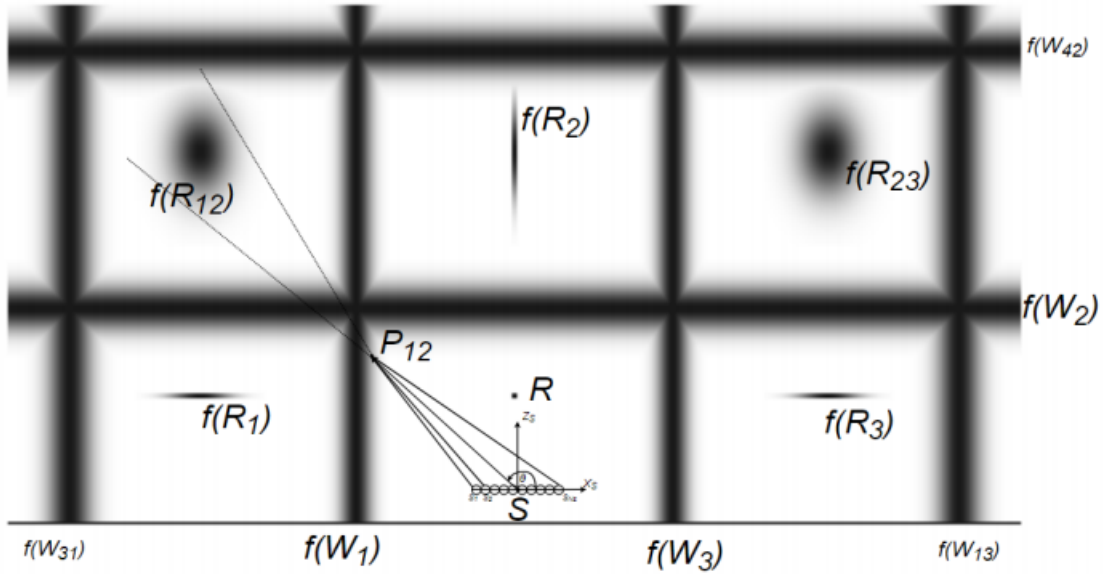


Fig. 3.13 The PRobabilistic Image Source Method (PRISM). The uncertain reflectors, W1, W2 and W3, are represented with blurry lines. The uncertainty of first order, R1, R2 and R3, and second order virtual receivers, R12 and R23, are also shown. The plane behind the source array is not considered for reflections. The reflectors and the virtual receiver positions become more spread out as the number of reflections increases

wide area of usage, PRISM may also be applied in other situations, where uncertainty in the scene geometry is present, to associate probabilities with quantities of interest.

The full derivation of this method can be found in [55]. The extraction of beamforming parameter distributions from the cumulative distribution function (CDF) of the distance parameter of the image model for first order reflections is detailed. It is validated by computer simulations and maximum likelihood estimation.

3.3 Ambisonics

Ambisonics is a full-sphere surround sound format capable of representing audio signals in both horizontal and vertical planes. With the prevalence of Virtual Reality (VR) and 360° video in recent years, ambisonics has become the go-to audio format for spatial audio and is the only audio format supported by YouTube 360-degree and VR videos. However, despite being developed in the 1970s, its use has only become widespread in recent years.

The British Ambisonics recording and reproduction system grew from the success of the recording industry when technical advancements allowed for new experimental recording techniques such as quadruphonics. The brain child of British academics including Peter Fellgett and the mathematician Michael Gerzon [39, 54], Ambisonics is a recording

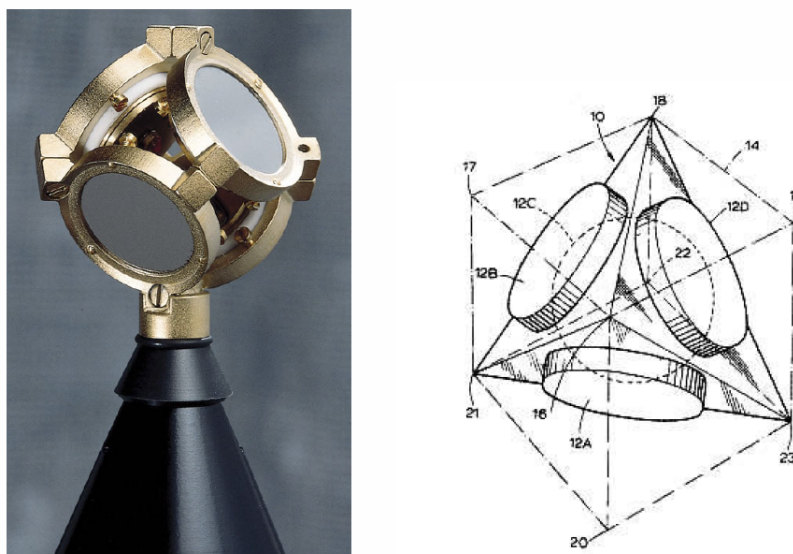


Fig. 3.14 The practical tetrahedral microphone (left) and design from patent document (right), reproduced after Craven [39].

and reproduction technique which builds on Blumlein's original work on stereophony [25]. Specifically it can be regarded as a spatial extension of the coincident microphone techniques presented in the previous section on stereophony, particularly the combination of a Blumlein pair and the Mid-Side technique.

Unlike traditional audio signals, Ambisonic signals are speaker-independent and can be transmitted and decoded for a variety of speaker arrangements. These Ambisonic representations of a sound field are referred to as *B-format* signals. B-format signals in First Order ambisonics can be broken into four channels labelled W, X, Y, and Z. The W-channel is a measure of omni-directional sound pressure and is akin to the Mid channel in a Mid-Side pair. The X channels represents front-minus back sound pressure, Z represents up-minus down while Y represents left-minus-right (the S channel of a Mid-Side pair).

Transmitted signals in ambisonics are not loudspeaker signals but are encoded directional sound within a set of signals. Three signals are required for horizontal only ambisonics while a fourth is needed to include the sensation of height and create a "surround sound" type effect. These loudspeaker independent signals can be used to drive a variety of loudspeaker arrangements made up of at least 4 loudspeakers. Depending on the locations within the speaker setup, the sound field can be 'steered' to a suitable arrangement.

3.3.1 Encoding Ambisonics

Fundamental to the recording of ambisonics and the B-format file format is the tetrahedral microphone. This type of microphone is made up of four hyper-cardioid microphone capsules spaced evenly on the four faces of a tetrahedron. This arrangement can be seen in Figure 3.14. This specific type of microphone array is also known as the soundfield microphone and was invented by Ambisonic forefathers, Peter Craven and Michael Gerzon [39]. A soundfield microphone produces four signals given as: left-back down, right-back up, left-front up, & right-front down. These signals are known as Ambisonic A-format.

A-format signals are not however used in signal processing, transmission or recording in the Ambisonics system. They are matrixed to B-format in the microphone base station using the following linear equations:

$$\begin{aligned}
 W &= 1/2(Lb + Lf + Rf + Rb) \\
 X &= 1/2(-Lb + Lf + Rf - Rb) \\
 Y &= 1/2(Lb + Lf - Rf - Rb) \\
 Z &= 1/2(-Lb + Lf - Rf + Rb)
 \end{aligned} \tag{3.11}$$

where Lf , Lb , Rf and Rb refer to left-front up, left-back down, right-front down and right-back up sub-cardioid capsules of the tetrahedral array. Therefore, using a tetrahedral microphone it is possible to generate all of the signals required to formulate a B-format signal from Equation 3.11. This formulation is achieved using a piece of circuitry known as the AB module which. It converts the A-format signals from the four hyper-cardioid microphones into the W (omni-directional signal) and the X, Y & Z (directional signals) components of the B-format signal. Following conversion, the B-format signals are encoded to represent bi-directional microphone signals facing forward, leftward and upward respectively as shown in Figure 3.15.

It is also worth noting that B-format spherical characteristics can be described using simple analytical functions:

$$\begin{aligned}
 W(\theta, \phi) &= 1/\sqrt{2} \\
 X(\theta, \phi) &= \cos(\theta) \cos(\phi) \\
 Y(\theta, \phi) &= \sin(\theta) \cos(\phi) \\
 Z(\theta, \phi) &= \sin(\phi)
 \end{aligned} \tag{3.12}$$

where θ , $[0 \leq \theta < 2\pi]$ is the azimuth angle (counted from the positive x direction in the anticlockwise manner) and ϕ , $[-\frac{\pi}{2} \leq \phi < \frac{\pi}{2}]$ is the elevation angle (counted from the equator toward positive z direction - upward, and the negative z direction - downward). The conversion between spherical and cartesian coordinates is then defined as follows:

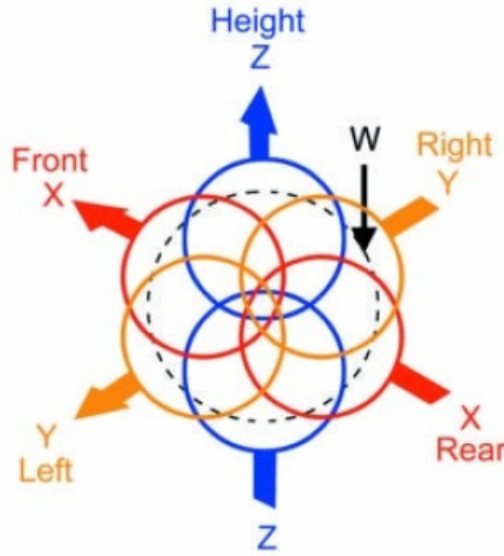


Fig. 3.15 B-format signal configurations, reproduced after Craven [39].

$$\begin{aligned}
 x &= ||\vec{r}|| \cos(\theta) \cos(\phi) \\
 y &= ||\vec{r}|| \sin(\theta) \cos(\phi) \\
 z &= ||\vec{r}|| \sin(\phi)
 \end{aligned} \tag{3.13}$$

Recordings of sound sources made in mono, s can be encoded into B-format signals using arbitrary assigned positions, (θ_S, ϕ_S) on the unit sphere. This is a fairly straightforward process which allows close-microphone sources to be encoded into Ambisonics after being recorded using the following formulation:

$$\begin{aligned}
 w(t) &= s(t) \times 1/\sqrt{2} = s(t) \times W(\theta_S, \phi_S) \\
 x(t) &= s(t) \times \cos(\theta_S) \cos(\phi_S) = s(t) \times X(\theta_S, \phi_S) \\
 y(t) &= s(t) \times \sin(\theta_S) \cos(\phi_S) = s(t) \times Y(\theta_S, \phi_S) \\
 z(t) &= s(t) \times \sin(\phi_S) = s(t) \times Z(\theta_S, \phi_S)
 \end{aligned} \tag{3.14}$$

where $w(t)$, $x(t)$, $y(t)$, and $z(t)$ are the B-format signal vectors, $s(t)$ is the pressure signal of the sound source to be encoded in a particular direction given by θ_S and ϕ_S .

3.3.2 Decoding Ambisonics

As already discussed, Ambisonics encodes the spatial information within B-format channels (W, X, Y and Z) with no particular prescribed loudspeaker layout for reproduction. The only specification is that the number of loudspeakers in an array configuration must exceed the number of Ambisonic channels. This in turn transfers the system complexity onto

the decoder which must prescribe signals from the B-format channels to each loudspeaker. While Gerzon never specified any one particular method for decoding Ambisonics, there have been many studies discussing the solutions to these simultaneous, non-linear, differential equations. The simplest case for decoding Ambisonic signals is for a two-dimensional layout with no height information. If the speakers are spaced equally apart in a circle, we can derive the loudspeaker signals, P_n from the B-format W,X and Y channels as follows:

$$P_n = W + X \cos(\theta_n) + Y \sin(\theta_n) \quad (3.15)$$

where θ_n is the direction of the loudspeaker along the unit circle. However, when we want to account for the additional Z channel in an Ambisonic decode we must look to other methods which allow us to incorporate this height information. Some of these methods are described in the followings sections.

Projection Method

When decoding an Ambisonic signal, an optimum performance is obtained when the loudspeaker array is regular in its layout, i.e. that there is equal spacing between each loudspeaker. This regularity in loudspeaker location allows for decoding through the method of projection. In this method, the feed for a particular loudspeaker L_n at position (θ_n, ϕ_n) on the unit sphere can be decoded as follows:

$$L_n = \frac{1}{\sqrt{2}} \cdot W + (\cos(\theta_n) \cos(\phi_n)) \cdot X + (\sin(\theta_n) \sin(\phi_n)) \cdot Y + (\sin(\theta_n)) \cdot Z \quad (3.16)$$

There are however only five platonic solids in three dimensions which satisfy the requirements for the projection method which can be seen in Figure 3.16. These being the tetrahedron (4 vertices), octahedron (6 vertices), cube (8 vertices), icosahedron (12 vertices), and dodecahedron (20 vertices). However, only two of these shapes is capable of offering near circular horizontal rings of loudspeakers, octahedron and icosahedron. This is an obvious limitation of the projection method, particularly for higher orders of Ambisonics.

Inverse/Pseudo-inverse Matrix Method

A common method of performing Ambisonic decoding is by inverse or pseudo-inverse matrix operation [75]. This operation is performed by first defining the vector $\mathbf{B}$ as a column vector made up of the concatenated Ambisonic channels:

$$\mathbf{B} = [W, X, Y, Z]^T \quad (3.17)$$

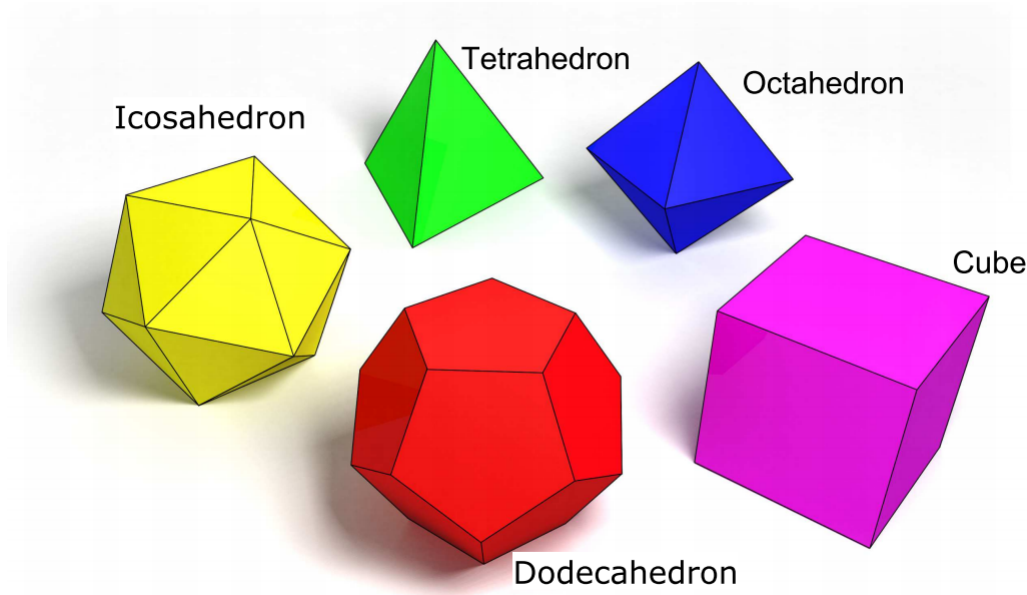


Fig. 3.16 Shown here are the five platonic solids. Placing loudspeakers at their vertices offers the only means of uniformly distributing loudspeakers in 3D space

We can also define $\mathbf{I}$ as a column vector of loudspeaker feeds. We can define $\mathbf{B} = \mathbf{C}\mathbf{I}$ where $\mathbf{C}$ is regarded as a re-coding matrix such that:

$$\mathbf{I} = \mathbf{C}^{-1}\mathbf{B} \quad (3.18)$$

where $\mathbf{C}^{-1}$ is the decoding matrix. This equation only applies if the number of loudspeakers matches the number of Ambisonic channels. If the number of loudspeakers is greater as is generally the case then $\mathbf{I}$ is formed by taking the right pseudo-inverse of $\mathbf{C}$:

$$\mathbf{I} = \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{B} \quad (3.19)$$

which yields a least squares solution. Clearly this technique requires that the matrix $\mathbf{C}$ be well conditioned for the solution to be accurate. This places further restrictions on loudspeaker placement.

Sound-field Rotation in Ambisonics

One of the major advantages of Ambisonics is the ease in which sound-field rotation can be performed via simple matrix operations as described in [123]. About the x-axis, sound field rotation is referred to as a *tilt* and is performed as follows:

$$\begin{aligned}
W^* &= W \\
X^* &= X \\
Y^* &= Y \cdot \cos(\phi) - Z \cdot \sin(\phi) \\
Z^* &= Y \cdot \sin(\phi) + Z \cdot \cos(\phi)
\end{aligned} \tag{3.20}$$

About the y-axis, sound field rotation is referred to as a *tumble* and is performed as follows:

$$\begin{aligned}
W^* &= W \\
X^* &= X \cdot \cos(\phi) - Z \cdot \sin(\phi) \\
Y^* &= Y \\
Z^* &= X \cdot \sin(\phi) + Z \cdot \cos(\phi)
\end{aligned} \tag{3.21}$$

About the vertical z-axis, rotation is performed as follows:

$$\begin{aligned}
W^* &= W \\
X^* &= X \cdot \cos(\theta) - Y \cdot \sin(\theta) \\
Y^* &= -X \cdot \sin(\theta) + Y \cdot \cos(\theta) \\
Z^* &= Z
\end{aligned} \tag{3.22}$$

Rotation about multiple axes can be performed by multiplying matrices together

$$\begin{aligned}
\begin{bmatrix} X^* \\ Y^* \\ Z^* \end{bmatrix} &= R_x R_y R_z \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \\
\begin{bmatrix} X^* \\ Y^* \\ Z^* \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \phi_x & -\sin \phi_x \\ 0 & \sin \phi_x & \cos \phi_x \end{bmatrix} \begin{bmatrix} \cos \phi_y & 0 & \sin \phi_y \\ 0 & 1 & 0 \\ -\sin \phi_y & 0 & \cos \phi_y \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}
\end{aligned} \tag{3.23}$$

3.3.3 Higher Order Ambisonics

B-format Ambisonic signals only capture first order spherical harmonics. These first order spherical harmonics are shown in the first two rows of Figure 3.17. A greater spatial resolution can be attained by using Higher Order Ambisonics (HOA). However, capturing HOA signals is difficult due to the complex microphone geometries required. These spherical harmonics, up as far as fourth order are also shown in Figure 3.17. One such microphone which makes use of HOA is the Eigenmike by MH Acoustics [132]. This

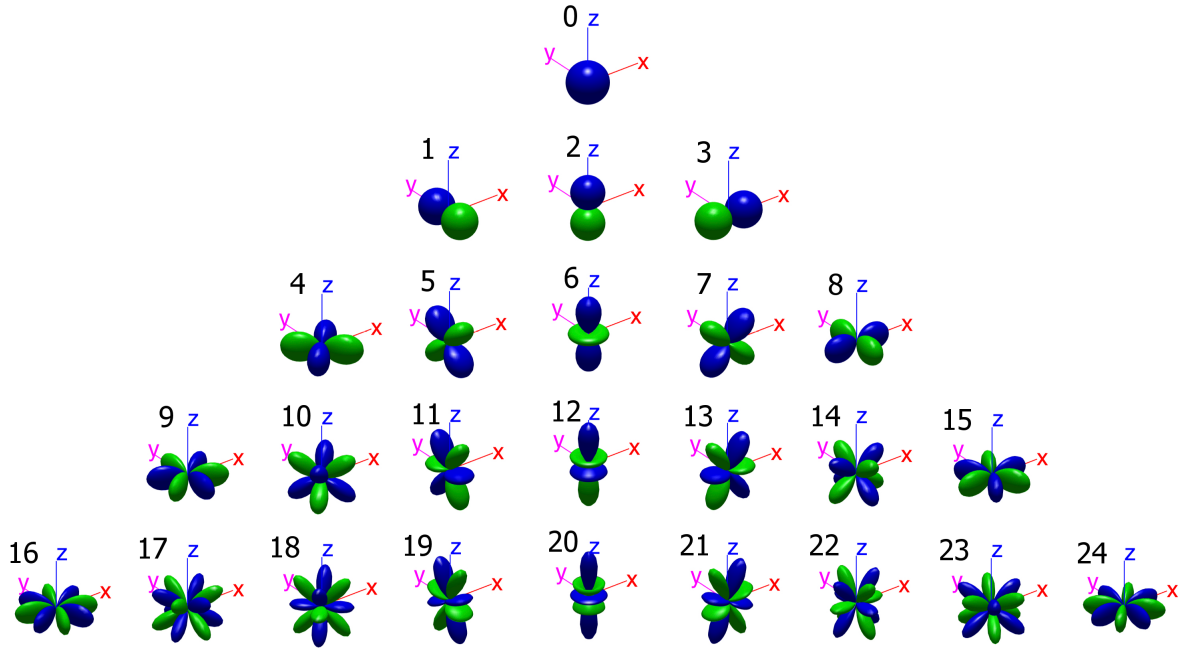


Fig. 3.17 Spherical Harmonics up to fourth order

is a 32-channel microphone capable of capturing Fourth Order Ambisonic signals based on designs from [17] and [136]. Increasing orders of ambisonics require an increase in the number of loudspeakers in the array used for reproduction. The minimum number of loudspeakers, N is defined by the Ambisonic order, M as follows:

$$\begin{aligned} N &= (M + 1)^2 & \text{3D array} \\ N &= 2M + 1 & \text{2D array} \end{aligned} \quad (3.24)$$

In Figure 3.17, we can see spherical harmonics up to forth order whereby there's an increasing sharpness to each node as the order increases. This contributes to a higher fidelity in terms of localization accuracy.

As there are no platonic solids with a great enough number of vertices to offer uniform loudspeaker distribution, the projection method of Ambisonic decoding is unsuitable for HOA and thus the matrix inverse method must be used. The complexity of the required rotation matrices also increases for HOA.

3.4 The TCD Audiosphere

A Third Order Ambisonics (TOA) audio sphere was constructed by the author for the purpose of capturing and recreating three-dimensional auditory environments. This 16-channel loudspeaker array was dubbed the Trinity College Dublin (TCD) Audiosphere and

Table 3.1 The distances and locations of each loudspeaker relative to the centre of the TCD Audiosphere

Speaker Number	Azimuth (Degrees)	Elevation (Degrees)	Radius
1	0	-13.6	1.41m
2	45	13.6	1.43m
3	90	-13.6	1.43m
4	135	13.6	1.43m
5	180	-13.6	1.40m
6	225	13.6	1.41m
7	270	-13.6	1.41m
8	315	13.6	1.41m
9	0	51.5	1.39m
10	90	51.5	1.41m
11	180	51.5	1.40m
12	270	51.5	1.40m
13	45	-51.5	1.42m
14	135	-51.5	1.42m
15	225	-51.5	1.40m
16	315	-51.5	1.41m

was used to reproduce and record signals for the research detailed in chapters 4 and 5 of this thesis. The order of the reproductive system was limited to third order due to space and cost constraints as creating a fourth order audio sphere would require 25 speakers. The sphere was constructed at the centre of a room with dimensions 420cm x 350cm x 270cm (W x L x H). It consisted of four layers of loudspeakers with the most elevated speaker layer attached to a metal frame which was dampened with insulating foam. The middle two speaker layers were placed on insulated speaker mounts while the bottom layer of speakers were placed on wooden frames wrapped in foam. The array consisted of Equator D5 loudspeakers based on a coaxially designed transducer with a response of 53Hz to 20kHz. All the speakers were angled towards the centre of the sphere with an average distance of 1.41m between the speaker cones and the centre of the sphere which was located 1.47m from the floor directly above the centre point of the array. The angles of each loudspeaker relative to the centre of the sphere can be seen in Table 3.1.

After initial construction, the room response of the audiosphere was captured using a DPA 4006 omnidirectional microphone placed at the centre of the array. A logarithmic sine sweep of 10 seconds was used to stimulate the room and this signal was played through each loudspeaker individually at a playback level of 20dBFS and recorded by the microphone. These recordings were then deconvolved with the original excitation signal to obtain multiple RIRs. Using the MATLAB function IR_stats it was found that

Table 3.2 Reverberation time per octave band after treatment of room.

Octave Band, Hz	Reverberation Time, R_{t60}
250	0.152s
500	0.132s
1000	0.1281s
2000	0.1079s
4000	0.0911s
8000	0.0891s

the average RT_{60} of the room was 0.22 seconds. While this value signifies that the room is sufficiently dry, it could be further improved upon and thus, additional dampening materials were applied to the room. The walls were treated with acoustic foam while drapes were hung behind the speakers themselves. The treatment of the room resulted in a RT_{60} of 0.13 seconds with a background noise level of 27.2dBA. After this treatment, it was determined that further dampening was not required and that the TCD Audiosphere could be used to perform experimental research safe in the knowledge that the room was properly acoustically treated.

3.5 Comparing Ambisonic Microphones

In a series of studies, the TCD Audiosphere was used to test the directional encoding capabilities of Ambisonic microphones. These studies contained both a subjective and objective evaluation of Ambisonic Microphones based on their timbral quality and directional encoding respectively. This work was collaborative for which the author conducted the tests determining the directional encoding of each microphone under test. The results of these studies were published in two separate academic papers at consecutive international conventions for the Audio Engineering Society (AES) [11, 13]. While the timbral quality was evaluated using subjective listening tests, the directional encoding of each B-format signal was examined using Directional Audio Coding (DirAC) methods [155]. More information on these studies can be found in Appendix 1 of this thesis.

3.5.1 Review of Directional Analysis Methods

Methods for estimating the direction of a source in a B-format recording include both analysis of the decoded signal [136], and energetic analysis. Energetic analysis may be applied directly to a B-format signal, therefore this method is potentially more useful when evaluating the fundamental directional accuracy of different microphones. Methods of energetic analysis include A-ELB [45] and intensity vector analysis. For the purposes

of this study an intensity vector based method was used. The intensity vector at a point in a soundfield is computed from the pressure and particle velocity at that point. The measurement of particle velocity is commonly performed indirectly using gradient microphone techniques [52]. Both the sound pressure and particle velocity may be measured using an Ambisonic microphone [130]. Previous directional analysis studies based upon intensity vector analysis include those of Merimaa [130] and Tervo [185]. Merimaa applied directional analysis to B-format impulse responses in the context of Spatial Impulse Response Rendering (SIRR). Tervo applied directional analysis for the purpose of direction estimation in a reverberant environment using a microphone array. Further intensity vector based studies include [122, 129]. Intensity based vector analysis for the purposes of the assessment of the performance of an Ambisonic microphone has been performed in [122] and [189].

For the purposes of this study intensity vector based analysis is used for the comparison of a range of Ambisonic microphones, using a modified form of DirAC analysis [190, 155]. DirAC analysis seeks to decompose a soundfield into perceptually relevant spatial qualities for each band of the Equivalent Rectangular Bandwidth (ERB) scale. The method applied in this test is intended specifically for the purpose of comparing Ambisonic microphones with respect to directional accuracy. The method used here differs from the works of Merimaa and Tervo as computations are performed in the time domain allowing for the computation of the instantaneous intensity vector, with the intensity vector calculation performed across the full bandwidth of the audio spectrum. In contrast, Merimaa analyses the spectrum up to 10 kHz, while Tervo analyses the range from 100 Hz to 5 kHz. The method introduced here is applied in all cases to recordings of a continuous, full bandwidth stimulus, as opposed to the B-format impulse responses studied by Merimaa, or the full bandwidth and instrumental recordings studied by Tervo. The analysed recordings are also captured in a controlled acoustic environment, as opposed to reverberant halls. This set of experiments represent the first use of intensity vector based directional analysis for the purpose of comparing the localisation accuracy of Ambisonic microphones.

3.5.2 Method

The five microphones under test were placed in turn at the centre of the TCD Audiosphere.

Two laser pointers were mounted on the frame to mark the exact centre point of the array and ensure the same position and orientation for each microphone. The source stimuli consisted of 5 second samples of pink noise at 48 kHz/24bit. Recordings were made using either the microphone's proprietary interface where supplied, or a MOTU 8M audio interface. Each recording was converted to B-format as before using the Furse-Malham weighting and channel order. Monophonic recordings from a DPA 4006 omnidirectional

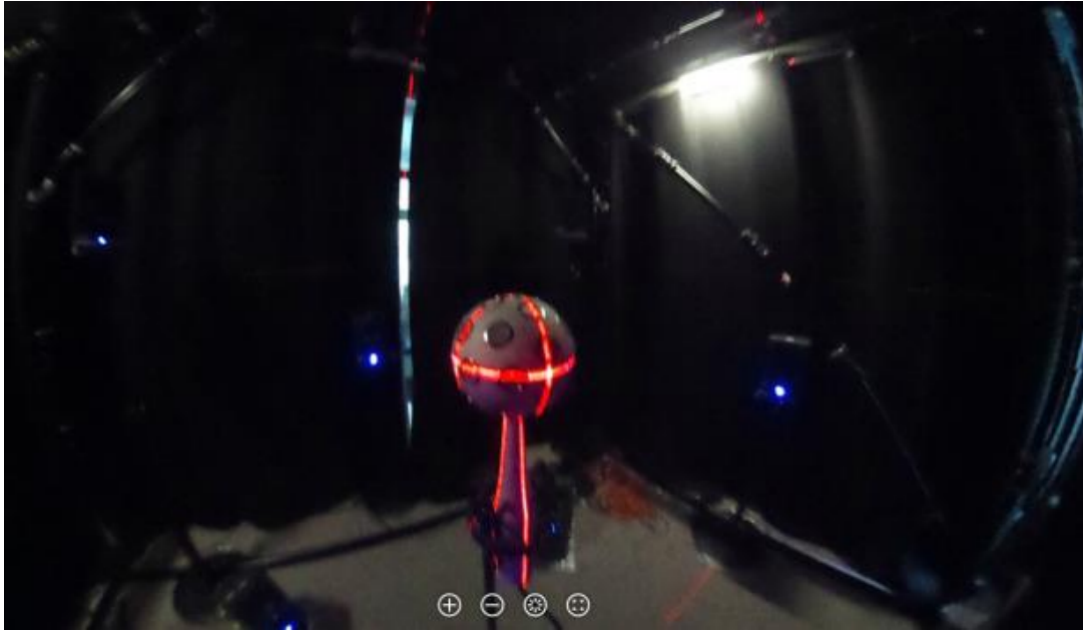


Fig. 3.18 Positioning system for Experiment B. The MH acoustics Eigenmike is placed at the centre of the TCD Audiosphere [11]

microphone were synthetically encoded into B-format using the Ambi-X encoder [17] and the azimuth and elevation angles of the loudspeakers in the array and this served as a reference for the directional analysis.

Drawing upon the principles of DirAC [190, 155] the directional analysis method used in this study extracts a direction estimate from a B-format recording utilising the instantaneous intensity vector:

$$I(t) = p(t)u(t) \quad (3.25)$$

Where $p(t)$ is the sound pressure, and $u(t)$ is the particle velocity. In the presence of a phase shift between $p(t)$ and $u(t)$, which may occur in the presence of acoustic interference, the instantaneous intensity vector will contain an active and a reactive contribution. The sound pressure $p(t)$ and an estimate of particle velocity $u(t)$ are calculated from the B-format components $[w(t), x(t), y(t), z(t)]$ as outlined in [130] yielding:

$$I(t) = \left(\frac{1}{\sqrt{2Z_0}} (w(t)x(t)e_x + w(t)y(t)e_y + w(t)z(t)e_z) \right) \quad (3.26)$$

Or, when written in terms of its directional components:

$$I(t) = I_x(t)e_x + I_y(t)e_y + I_z(t)e_z \quad (3.27)$$

where Z_0 is the characteristic acoustic impedance of air and e_x , e_y , and e_z are unit vectors in the directions of the corresponding Cartesian coordinate axes.

In advance of intensity vector estimation, the signal is band pass filtered according to the 25 bands of the Bark scale. Bark band filtering emphasises the contribution of low frequency content, which for B-format recordings are more accurate in terms of source localisation [58]. An intensity vector is then calculated for each band and each sample. The letter i is used to denote the band, the letter j is used to denote the sample. The source angles $\theta_{i,j}$ (azimuth) and $\phi_{i,j}$ (elevation) may be computed as follows.

$$\theta_{i,j} = \begin{cases} \arctan \frac{I_{y,i,j}}{I_{x,i,j}}, & I_{x,i,j} \geq 0 \\ \arctan \frac{I_{y,i,j}}{I_{x,i,j}} - 180, & I_{x,i,j} < 0 \end{cases} \quad (3.28)$$

$$\phi_{i,j} = \arctan \frac{I_{z,i,j}}{\text{mag}_{a,i,j}} \quad (3.29)$$

$$\text{where } \text{mag}_{a,i,j} = \sqrt{I_{x,i,j}^2 + I_{y,i,j}^2} \quad (3.30)$$

As the intensity vector is computed on a sample by sample basis, errors in the signal representation of the B-format recording must be considered. As the B-format components of a 24-bit recording approach zero, errors may result in a directional component erroneously crossing the axis in advance of the omnidirectional component. This results in a change of sign of the intensity component, which for the purposes of azimuth angle calculation results in a $\pm 90^\circ$ shift in the angle estimate. If both intensity components change sign, a 180° shift in the azimuth angle estimate is observed. Such numerical errors do not arise in synthetically encoded, 32-bit signals, but will occur in the finite resolution of 24-bit recorded B-format signals. Therefore, in advance of angle computation a thresholding algorithm was applied to $I_{y,i,j}$, $I_{x,i,j}$, $I_{z,i,j}$ and $\text{mag}_{a,i,j}$.

For each band and each sample, the mode of angle estimates was taken as the estimated angle to the source. As can be seen in the histogram of azimuth angle data for a loudspeaker at 45° azimuth and 0° elevation (shown in Figure 3.19), the angle estimates are spread across the range -180° to 180° but the mode correlates strongly with the true source angle. A second peak of lower amplitude can also be observed in the histogram. This is a feature of B-format recordings in a non-anechoic environment, and is not confined to a particular set of frequency bands. Modal angle estimation ensures that such artefacts introduced to the distribution of angle estimates do not influence the estimated angle to the source.

3.5.3 Results & Analysis

Using the method described above, an estimate for the azimuth and elevation angles of each of the 16 loudspeakers was obtained for recordings from each of the five microphones in this study, and the Reference. This analysis was also performed on the Ambeo microphone recording without the optional “Ambisonics Correction Filter”. In addition, the analysis was also applied to Zoom H2n recordings using both the native Ambisonics recording mode, and a freeware conversion plugin. All of these estimates were compared to the actual loudspeaker locations and an absolute offset error determined. These results are summarized in Table 3.3 and also visualized in Figure 3.20. Note that as the Zoom H2n does not include a vertical component, no elevation results are included.

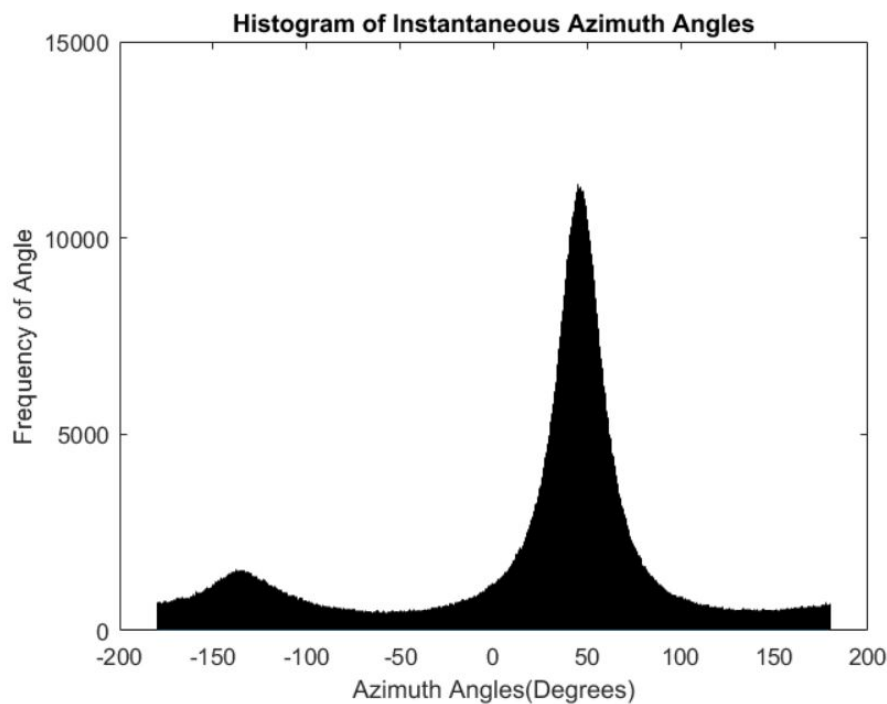


Fig. 3.19 Histogram of azimuth angles across 25 frequency bands [13]

3.5.4 Discussion

The minimal azimuth and elevation angle offsets show that the method described is sufficient in determining the source angle from a B-format signal (noting that these B-format signals were obtained for sound fields with a single, distinct, point-like sound source). Because of the unequal variances as well as half-normal distribution of the sample data, the differences between each pair of microphones were formally statistically evaluated using a paired t-test, as opposed to one-way ANOVA. The results revealed that

Table 3.3 Mean Azimuth (AZ) offset and Elevation (EL) offset error $\pm 95\%$ Confidence Interval for 16 loudspeaker locations

Microphone	AZ error (°)	EL error (°)
Reference	0.09 ± 0.07	0.23 ± 0.10
SF MKV	2.27 ± 0.84	5.69 ± 2.71
Tetra-Mic	4.46 ± 2.76	3.35 ± 1.39
Ambeo - No Filter	2.25 ± 1.24	1.62 ± 0.57
Ambeo - With Filter	2.24 ± 1.31	1.67 ± 0.57
Eigenmike	1.28 ± 0.50	1.56 ± 0.58
H2n - native	8.58 ± 2.47	N/A
H2n - plugin	11.25 ± 3.8	N/A

there are significant differences between microphones. Figures 3.21 & 3.22 show the average azimuth and elevation errors and variance for each microphone, as within 95% confidence intervals.

In terms of azimuth, the Reference performed better than all other microphones, as expected. The Eigenmike was on par with the Ambeo (both versions) and performed better than the Soundfield MKV, TetraMic, and H2n (both versions). Both versions of the Ambeo were on par with the Eigenmike and Soundfield MKV and performed significantly better than the TetraMic and H2n (both versions). No significant differences were found in terms of azimuth localization accuracy for the Ambeo with and without the correction filter. The Soundfield MKV performed on par with both Ambeos and TetraMic, significantly better than both H2ns, but worse than the Eigenmike. The TetraMic was on par with the Soundfield MKV and H2n-native, performed better than the H2n-plugin, and worse than the Ambeo (both versions) and Eigenmike. Finally, no statistically significant differences were found for the H2n in native Ambisonics recording mode when compared to the conversion plugin, with both producing worse results than all other microphones, with the exception of the TetraMic which was on par with the H2n-native. It should be noted however that as the H2n does not include a vertical component and as all sources were offset from the horizontal plane to some degree, this increased the overall localization error for this microphone.

In terms of elevation, the Reference again performed better than all other microphones. No significant difference was found between the Eigenmike and Ambeo (both versions), with both of these microphones performing better than the Soundfield MKV and TetraMic. No significant differences were found between the Soundfield MKV and TetraMic, with both performing worse than all other microphones in terms of elevation.

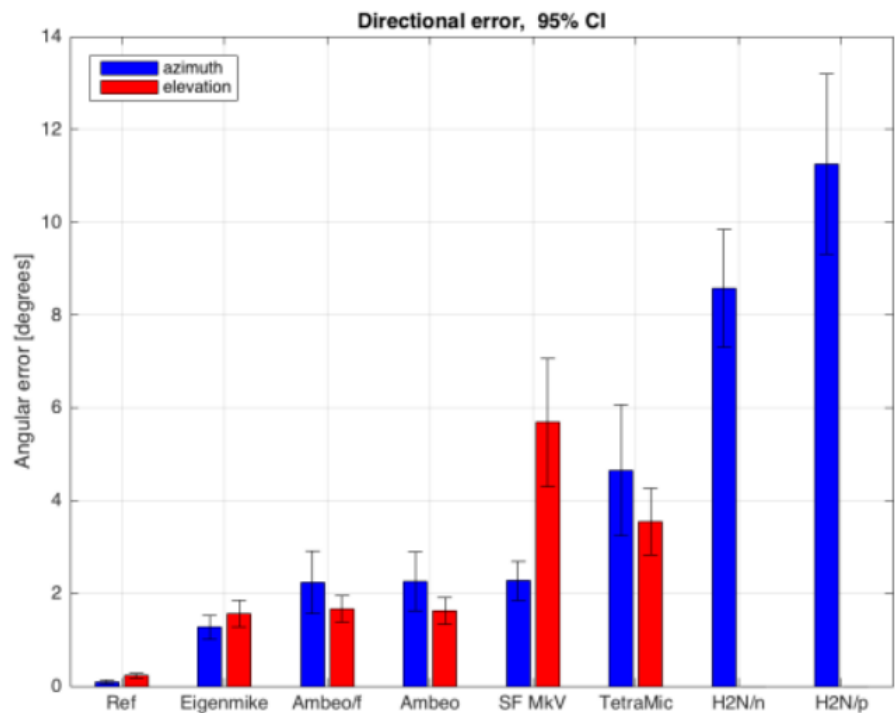


Fig. 3.20 Mean azimuth and elevation error for 16 sources [13]

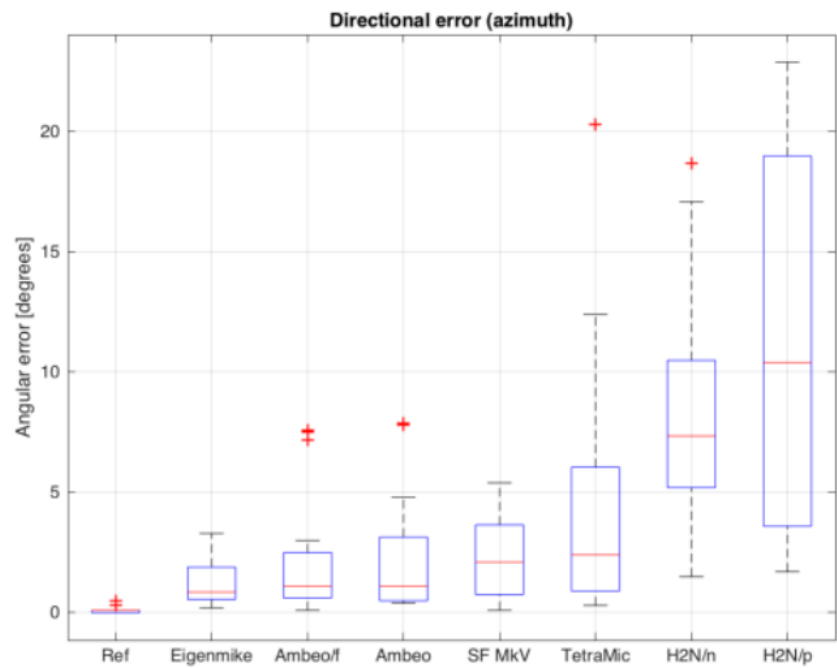


Fig. 3.21 Azimuth average error and variance [13]

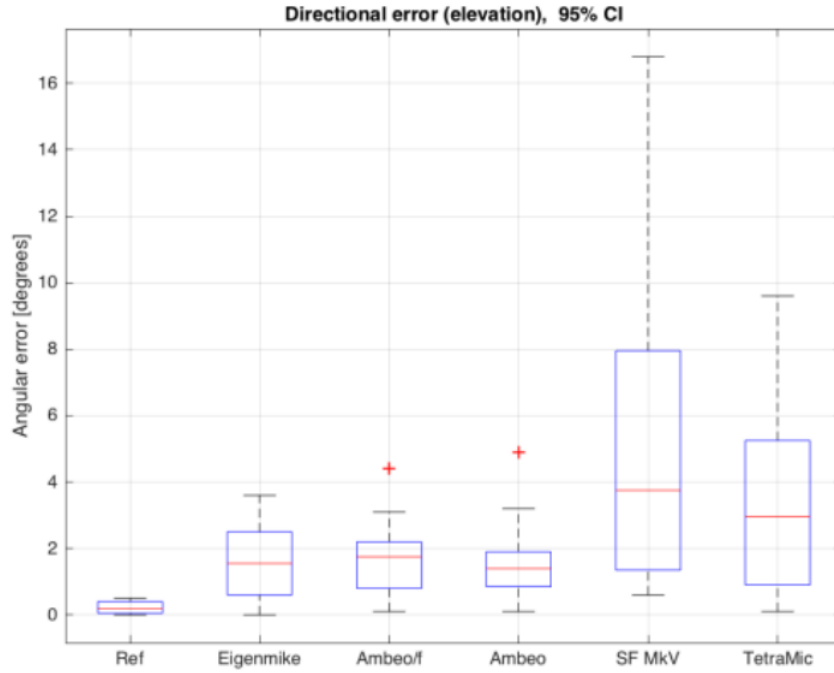


Fig. 3.22 Elevation average error and variance [13]

3.6 Binaural Reproduction

Thus far this chapter has discussed a variety of loudspeaker-based reproduction techniques of spatial sound. However, reproduction of audio over headphones is more common and often preferred for a number of reasons. Headphones offer users a privacy which is unattainable through conventional loudspeaker setups. In addition to this, headphones will always place the listener at the sweet-spot of an auditory environment negating any prevalence effects typically associated with loudspeaker arrays. Many listeners also prefer headphones to loudspeaker arrays due to their portability and ease of use.

It's important at this point to distinguish between "binaural" audio and "biphonic" audio. Where "biphonic" audio refers to regular audio signals such as music delivered over headphones, "binaural" audio aims to create an apparent sound source located outside the listener's head [124].

In order to auralise a monophonic sound event using headphones, the source audio needs to be convolved with a pair of left and right Binaural Room Impulse Responses (BRIRs), $b_L^{\Theta_n}$ and $b_R^{\Theta_n}$ corresponding to the source direction Θ_n . The resultant audio then delivered to the listener is the summation of all the sound sources convolved with their relevant BRIR pair (see Figure 3.23). Of course, the correct BRIR relating to the desired source position is required for each sound event in a scene. This requires a large database of BRIR sets to be available for a large number of source locations. Moreover, when the

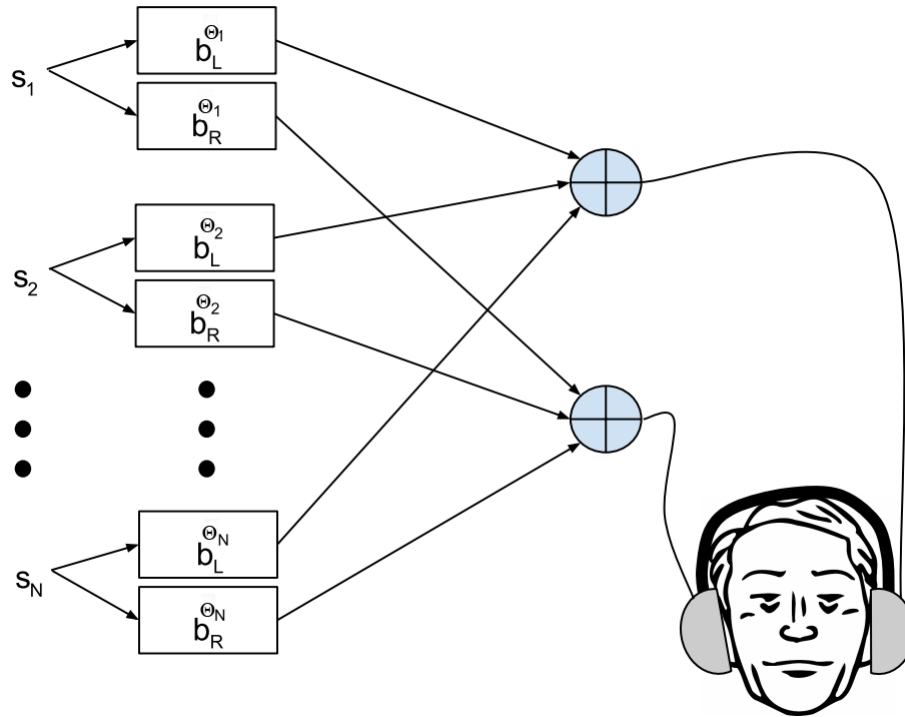


Fig. 3.23 Shown here is the setup for binaurally rendering sound, with each sound source processed separately

source-listener position changes in the scene, the sound field needs to be re-rendered for the new spatial configuration. Simply switching from one BRIR set to another in this instance can cause auditory artefacts caused by wave discontinuity in the convolved binaural signals [147]. Therefore, another approach is required to dynamically change a sound field and present it over headphones.

3.6.1 Virtual Loudspeaker Approach

One such method which addresses the problem of sound field rendering over headphones is presented by McKeag and McGrath in [127]. Here, a method for decoding Ambisonic B-format representations of sound fields to form a suitable headphone playback is presented which allows continuous, smooth head-tracking at all angles. Consider a listener sitting in the middle of an array of loudspeakers over which a B-format signal is being presented. The objective is then to recreate the signals that the listener would observe at each ear from the loudspeaker array. Seeing as the signals fed to the loudspeakers are linearly related to the B-Format signals, the signals received at the ears are in turn linearly related to the speaker signals. Therefore there must be a linear relationship between the four B-Format channels and the two signals received at the ears. So using this method, instead of feeding

the signals generated to physical loudspeakers, each loudspeaker feed is instead convolved with the BRIR pair corresponding to that loudspeaker location. Thus the left and right headphone signals h_L , h_R are formed as:

$$\begin{aligned} h_L &= \sum_{n=1}^N b_L^{\Theta_n * I_n} \\ h_R &= \sum_{n=1}^N b_R^{\Theta_n * I_n} \end{aligned} \quad (3.31)$$

where I_n is the n^{th} loudspeaker feed and ϕ_n denotes the location of that loudspeaker in terms of angle to the listener.

When considering head-tracking, it's necessary to compensate for motion in some way. Conveniently with B-format signals it is simpler to rotate the acoustic environment than it is to rotate the listener within that environment. This can be done by applying a three dimensional rotation matrix to the X, Y and Z signals on a per sample basis as discussed in Section 3.2.2.

3.6.2 Optimised Virtual Loudspeaker Approach

Real-time virtual loudspeaker formation can be computationally expensive due to the length of HRIR signals. To reduce this, HRIRs (h^ϕ) can be broken into a direction independent subsystem (f) which is common to the entire HRIR set and a direction dependent residual (g^ϕ). This factorization process is proposed in [92] by Kearney et al. where it is shown how a 500 sample HRIR filter can be approximated as a convolution of a 451-sample long common (direction independent) filter and a 50-sample long direction dependent residual.

3.6.3 Headphone Design

The design of headphones can have a significant impact on their ability to reproduce spatial audio. As headphones become more widely used, the variety in their shape, size & fit becomes a variable of increasing magnitude. Generally speaking, headphones can be categorised as in-ear, over-ear, or on-ear headphones. Examples of each can be seen in Figure 3.24. In-ear headphones fit snugly inside each ear canal with their light-weight design making them popular for use during exercise. While some in-ear headphones are excellent for passive noise isolation, they are generally far inferior to both on-ear and over-ear headphones in terms of quality. Over-ear headphones are generally considered to be the best in terms of audio quality [70]. They are characterised by their large circumference which surrounds the entire ear. Most over-ear headphones perform well in isolating the user from outside sound using what's known as "passive noise reduction". Over-ear



Fig. 3.24 From left to right - in-ear, over-ear, and on-ear headphones

headphones come in two distinct styles, "closed-back" and "open-back". Closed-back headphones have a sealed back surrounding the transducer preventing any audio signals from escaping. Binaural sound reproduction assumes that to recreate natural sounding audio at a listener's ears using headphones, the same pressures experienced in a natural environment are required [15]. As such, the pressure created at the ears by closed back headphones make it difficult to perceive binaural audio as natural sounding. Open-back headphones on the other hand allow sound from outside the headphone to pass through the transducer, reducing the pressure at the ear. As a result, they are generally preferred for the playback of binaural audio in VAEs. Open-back headphones are problematic however as they produce a large amount of audio spill. Lastly, on-ear headphones are similar to over-ear headphones but instead of encompassing the entire ear they are smaller and sit on top of the outer ear. They generally produce good audio quality but with less bass response than their over-ear counterparts. There exist several other styles of headphones such as bone-conduction headphones. However, these are not examined as part of this thesis.

Satongar et al. found that headphones have a significant influence on the localization of audio in VAEs [168]. This effect was intensified with multiple re-positionings of the headphones which was examined using the spectral error in HRTFs. Closed-Back headphones introduced the largest distortions in HRTFs while Electrostatic headphones maintained the most stable HRTFs.

3.7 Conclusion

This chapter has provided an insight into the variety of ways in which spatial audio can be produced. Beginning with 2-channel stereophony, we saw how phantom sources can be created between pairs of loudspeakers by adjusting the gain or the time of arrival of the sources from each speaker. Also introduced in this section were some of the most

common stereo microphone techniques which are used to capture an auditory scene. Using just two microphones a stereo recording can be made which can then be reproduced over a 2-channel loudspeaker system. Multiple loudspeaker stereophonic systems were then discussed with regards to the ubiquitous 5.1 system seen in many home entertainment systems. Next we examined how height can be introduced into a stereophonic loudspeaker system with vector base amplitude panning. In the same way that phantom sources can be created along the horizontal using 2-channel stereophony, phantom sources can be placed anywhere in a 360° sphere around the listener when additional channels of height are added. The following section of this chapter contained a discussion on digital sound projectors and how a line-array of loudspeakers can be used to beamform a signal in a particular direction, causing reflections of the signal to occur when they coincide with a wall or perimeter of a room. These so-called soundbars are thus able to create the same sensation of a 5.1 speaker setup using just one linear array of loudspeakers. The traditional image source method of soundbars is discussed next as part of a larger discussion on the probabilistic image source method, a novel method of calibrating soundbars in rooms with uncertain geometries. This method was in part developed by the author. Perhaps the most significant part of this chapter lies in the section on Ambisonics, the method of producing spatial audio which has now been incorporated into everyday technology such as YouTube's 360° video format. Ambisonic B-format signals can be encoded and decoded for systems consisting of many different arrangements of loudspeakers and as such is a valuable and versatile format which is examined in greater detail in Chapter 4. As part of the discussion on Ambisonics, we reference the TCD Audiosphere, a third-order Ambisonic loudspeaker array built by the author to run experiments on Ambisonic arrays and microphones. Finally, headphone technology and binaural reproduction is discussed bringing the focus of the chapter back to 2-channel reproductive audio systems.

Chapter 4

Virtual Headphone Testing for Spatial Audio

The delivery of spatial audio with a high degree of sound quality and accurate localization is essential for creating a sense of immersion in a virtual environment. Typically, spatial audio in virtual environments is delivered via headphones. Other means include using speaker arrays although these methods can be problematic due to a user's natural tendencies to move and adjust their position. Thus, it is important to develop an accurate method of comparing the sound quality and localization accuracy of spatial audio across a range of headphones. By doing this, an attempt can be made to quantify what properties are most significant in creating a sense of immersion in a virtual environment. Virtually immersive games and visual experiences are becoming increasingly prevalent as leading tech companies such as Facebook, HTC, Sony and Google are launching high-end virtual content and the technology to present it. Examples of this including the Oculus Rift, HTC Vive, Playstation VR, and Youtube's new 360° video presentation.

With the standard of the visual component of virtual experiences increasing so rapidly it is important that the standard of the accompanying spatial audio improves at a similar rate. Establishing a suitable method of audio playback is essential to this and determining the most accurate method of headphone testing is fundamental in this regard as headphones are most often the preferred choice for audio playback in virtual environments. In this chapter a study comparing two virtual headphone testing methodologies is discussed. This work was published in 2016 at the International Audio Systems Symposium in York, UK [145].

4.1 Virtual Headphone Testing

Conventional headphone testing usually consists of a swap-and-compare strategy where listeners are presented with the same audio stimuli over an array of different headphones and are asked to compare the quality of the audio over each. This methodology is difficult to execute owing to the challenges in controlling nuisance variables and their inherent biases. For instance, a listener's preference can be heavily influenced by sighted and tactile biases such as the fit, design, and branding of the headphones [78]. Other inconsistencies arise due to leakage effects, a variation in fit across listeners, and auditory memory loss caused by long periods of time between swapping headphones. This is particularly an issue when comparing in-ear headphones where a good seal is required for good bass performance below 500 Hz [198].

Alternative solutions which alleviate these difficulties have been proposed in recent years. These include methods which employ a virtual rendering of each headphone set via binaural recordings (*recorded* methodology) or by applying the frequency response of the headphones to the test stimuli (*filtered* methodology). In virtual headphone testing, the subject uses just one set of headphones throughout the testing procedure [78, 145, 70]. These replicator headphones are of a high quality and are compensated to have a neutral frequency response; it is assumed that any coloration caused by these headphones will not have a significant effect on the listener's preference of the audio. By doing this, listeners can make instantaneous comparisons between different headphones in a controlled, double-blind way.

In the first virtual method, listeners are presented with a series of test samples that have been produced by another set of headphones and recorded onto a binaural head microphone. This is the *recorded* test strategy as presented by Hirvonen et al. [78]. In this study, 6 headphones including four circumaural and two intra-concha models were evaluated. Based on the preference ratings of 21 subjects, the study showed that results attained for the 'Recorded' method and the conventional swap-and-compare method were similar. However, statistical analysis showed significant differences between the two methods meaning the virtual method could not be validated based on these results. These statistical differences were the result of a single intra-concha model headphone performing significantly worse in the virtual test methodology than in the standard swap-and-compare method. This disparity between the methods was attributed to tactile biases as errors related to headphone fit in the recording stage. That study was also limited to speech signals and was conducted at a sampling frequency of 10kHz, meaning that any difference in headphone quality above 5kHz was not accounted for.

A different method of headphone virtualization was employed in a 1992 study by Brielle and Voinier [29]. This work used a filtered approach whereby a high quality

replicator headphone was equalized to match the frequency response of the headphones under test. A similar method was employed by Olive et al. in 2013 [145]. In this study, the Headphone-Related Transfer Functions (HpTF) of several headphones were measured using an ear and cheek simulator microphone. Using deconvolution methods, an impulse response is generated which is representative of the headphones frequency response. Although it has been shown that there is a wide degree of variability in the characterization of the headphone transfer function [100], multiple reseatings of the headphones and an averaging process were set in place to account for this. These impulse responses were then convolved with a set of stimuli to simulate the sound of the headphones. Listeners were then asked to rate their preference of virtualized headphone sounds as well as rating the same headphones using the conventional swap-and-compare method. A substantial set of listening tests examining preference and spectral balance showed a good correlation between the two methods when examining overall headphone preference (correlation coefficient $r = 0.85$). Following on from this study, further research has been conducted to reduce the error in estimating these HpTFs [26]. The Princeton Headphone Open Archive (PHOnA), a database of HpTFs, has since been established and shared with the public in an attempt to optimize equalization algorithms and provide more universal solutions for perceptually transparent headphone reproduction [27].

This equalization approach to virtualizing headphones is advantageous as it eliminates the need for tediously recording each stimuli with each set of headphones. Instead, the magnitude and phase response of the headphones are measured just once and can then be applied in real-time to any audio source using digital filters. A digital signal processing framework for performing this filtration including simulating the headphones isolation capabilities in noisy environments is outlined in [157]. While an impulse related method of virtualizing different headphones is advantageous in terms of its speed and efficiency, it does not capture nonlinear distortions which are accounted for when directly recording the headphones under test. However, a recent study investigating circumaural headphones found that distortion levels have little bearing on preference rating unless they are very high [184].

A study published in 2016 by Gutierrez et al. investigated the influence that the perceived quality of consumer headphones has on the spatial impression of the audio [70]. Subjects were presented with a virtual headphone comparison method similar to the one explored in [145] and were asked questions based on both the general timbral quality and the spatial impression of the headphones. The study concluded that there is a strong correlation in the perceived quality and the spatial impression of the headphones under test. The virtual testing conducted in this study was presented to subjects using a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) type test [85]. This required subjects to rate the headphones under test in relation to a high quality reference

headphone as well as a low quality ‘anchor’ rendering of the high quality reference [85]. For questions regarding perceived quality, a low-passed render of the reference was used as the anchor whereas a monaural render of the reference was used for questions regarding spatial impression.

While the filtered approach to virtual headphone testing is undoubtedly more efficient than the recorded method, it does not accurately account for nonlinear distortions captured during recording. Digital filtering can also introduce artifacts during convolution which are not representative of the true sound of the headphones. Previous studies addressing the differences between headphone comparison methods have only addressed one virtual testing method with relation to physical headphone testing [78, 145]. However, this work aims to compare two virtual methods of headphone testing to determine whether the results they provide are significantly similar. In addition to comparing these two methods, the correlation between perceived audio quality and spatial impression is examined and how this may differ for the two methods. As it was shown in [70] that these two measures are highly correlated when using a ‘Filtered’ virtual test, this study sought to observe whether this may be the case while using the ‘Recorded’ virtual method.

4.2 Methodology

In this study, two MUSHRA style listening tests were conducted, each featuring stimuli captured using either binaural recordings of different headphones as in [78] or by applying equalization to the source stimuli to simulate the effect of these headphones as per [145]. Subjects were asked questions relating to both the general quality and spatial impression of the audio in both tests. These MUSHRA tests employed the use of a high-quality reference signal for which the other signals should be compared to. While the employment of a reference signal in listening tests has been shown to influence comparisons of 3D audio signals [179], it was included in this study as a means of ensuring participants were capable of distinguishing a high quality signal and thus validating their subsequent results. The objectives of the study were to examine the similarities in results attained using these methods and to verify that the correlation between the preference in general quality and spatial impression as found in [70] was also to be found when the source stimuli consisted of binaural recordings. As previous attempts to validate the testing method described in [78] have been unsuccessful, it would be beneficial to determine whether the results yielded from this type of testing can be comparable to those attained from a method which has been validated. The experimental design and methods employed in this study are described in detail in the remainder of this section.

4.2.1 Headphones

In this study, the general quality and spatial impression of four circumaural headphones is examined via two separate MUSHRA tests. Beyerdynamic DT-770 Pro headphones were used in this instance as high-quality replicator headphones. These headphones can be seen in Figure 4.1. The frequency response of these headphones was captured using a Neumann KU 100 dummy head microphone. A swept sinusoidal tone was played through the headphones and recorded using the KU 100. This recording was deconvolved with the original signal to attain an impulse response representative of the HpTF. The HpTF was then found by finding the FFT of the impulse response. The resultant HpTF was sampled at 31 points in 1/3rd octave bands from 20Hz to 20kHz. The inverse filter was taken as the negative of the gain values at these frequencies and applied to the test signals using a graphic equaliser in Adobe Audition. Using this, an inverse filter was designed and applied to each stimuli for the listening test to ensure that any coloration caused by these headphones was accounted for. The measured Headphone Transfer Function (HpTF) for the DT-770 headphones and the resulting inverse filter can be seen in Figure 4.2.

Details of the four headphones compared via the listening tests are featured in Table 4.1 including the headphones' manufacturer, model, price, and style. These headphones can also be seen in Figure 4.3, modelled on a Neumann KU-100 dummy head microphone. While the prices of these headphones are in the low to mid-range, there is one notable exception, the Sennheiser HD-650. These open-back headphones were chosen to act as the high-quality reference signal in the MUSHRA tests as they do not block the ear canal. Compared to the other headphones under test these headphones also had the flattest frequency response making them a good candidate for a reference signal.



Fig. 4.1 Beyerdynamic DT-770 headphones. The replicator headphones used in this study for their flat frequency response and high fidelity [148]

Open-back headphones allow the user to hear the ambient sounds around them without extra attenuation making them ideal for spatial audio and virtual auditory environments

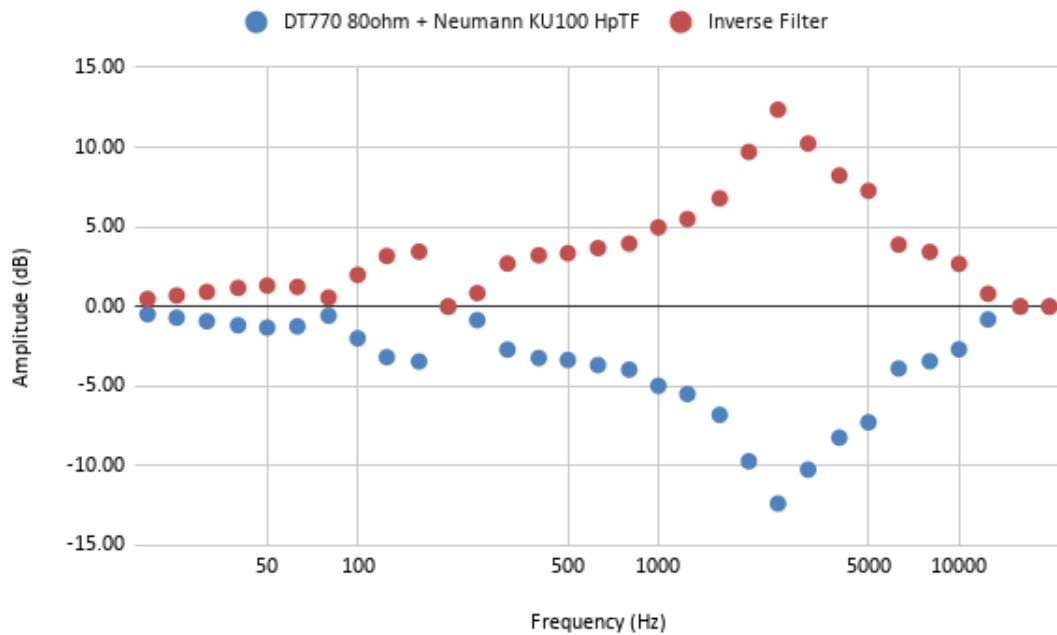


Fig. 4.2 The frequency response of the Beyerdynamic DT-770 headphones measured on the Neumann KU-100 Dummy Head Microphone and its inverse filter. A total of 31 measurements were taken in 1/3rd octave bands from 20Hz to 20kHz.

[187]. Hearing both the synthetic signal from the headphones and the ambient sounds from the natural environment allows for *Audio-augmented reality* which features hear-through audio [79]. Bone-conduction headphones also allow for hear-through audio as they don't block the ear canal [109]. Headphones B,C, and D are all on-ear headphones. However, headphones B are open-back while C and D are closed back. In theory this should give them some advantage over the others when it comes to the preference of spatial audio.

Table 4.1 The headphones examined in this study:

Label	Manufacturer	Model	Price, €	Style
A	Sennheiser	HD 650	315*	Over-Ear, Open, Reference
B	Sennheiser	PX-100-II	53**	On-Ear, Open
C	Koss	KSC 75	21**	On-Ear
D	Philips	SH 5200	23.50**	On-Ear

*Price from Thomann.de as of September 2016

**Price from Amazon.co.uk as of September 2016



Fig. 4.3 A variety of headphones mounted on a Neumann KU-100 dummy head microphone. Top-left: Sennheiser HD 650 (Circumaural, Open-Back), Top-Right: Sennheiser PX-100-II (Circumaural, Open-Back), Bottom-Left: Koss KSC 75 (Circumaural, Closed-Back), Bottom-Right: Philips SH5200 (Circumaural, Closed-Back) [148]

4.2.2 Stimuli

Four stimuli were selected for use in this study which represented a wide range of binaural audio. These stimuli differed in length with the shortest being 10 seconds long and the longest being approximately 30 seconds long. A description of each of these Stimuli can be found in Table 4.2. Each stimuli examines different categories of audio signals. Stimuli A & B are both broad band signals while C & D are narrow band. Stimulus A features stationary musical sources set around a wide sound stage while the other 3 stimuli all feature moving sound sources. This movement places emphasis on the spatial effects of the audio.

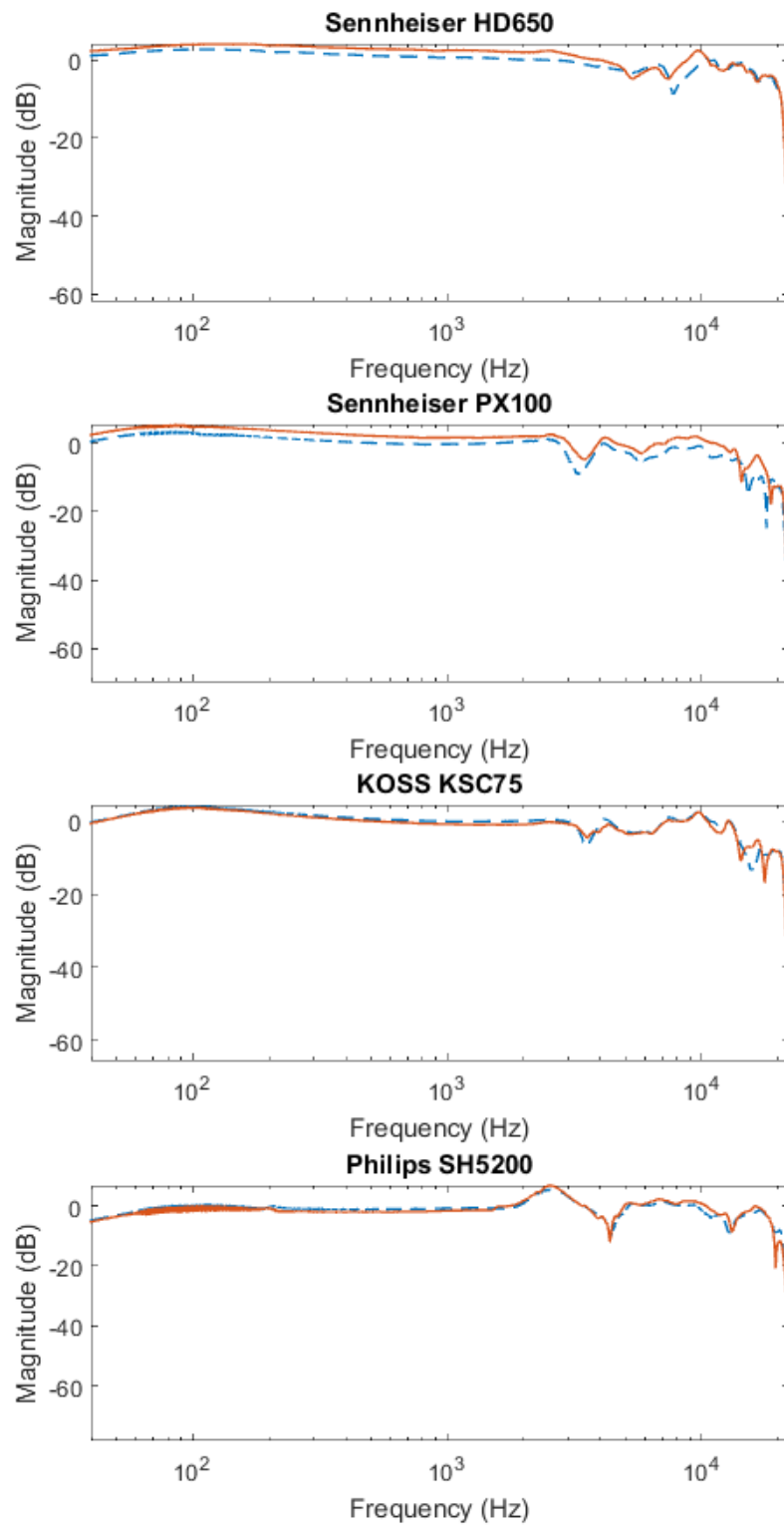


Fig. 4.4 Frequency Response of the headphones used in the study. Solid curves, left channel; dashed curve, right channel [148]

Table 4.2 The stimuli used in this study:

Label	Stimulus	Description	Length, seconds
A	Mark Ronson ft. Bruno Mars - Uptown Funk	Up-tempo pop song with high quality production, a wide sound stage, and responsive bass	26
B	Nature Scene	A binaural recording of a thunder storm with moving sound sources	30
C	Binaural Shaker Recording	A recording of a shaker moving about a binaural head microphone	20
D	Surround Sound Speaker Test	A binaural rendering of a 5.1 speaker surround test with 5 separate source locations (Front Right, Front Left, Centre, Back Left, & Back Right)	10

4.2.3 Stimuli Preparation - Binaural Recording and Headphone Frequency Response Measurement

Both MUSHRA listening tests conducted in this experiment required separate preparation of stimuli. In the first method, known as the *recorded* method, each of the four headphones under test were placed in turn at an optimal seating on to a Neumann KU-100 dummy head. The four test stimuli were then played at an appropriate level through the headphones and a recording of each was made using the dummy head microphone. Each stimulus had a sampling rate of 44.1kHz and a bit rate of 32 bits floating point. These recordings were made in the TCD Audiosphere, a treated room with an RT_{60} of 0.13 seconds at 1kHz and a background noise level of 27.2dBA. During the playback of the *recorded* stimuli during testing, an inverse filter was applied to balance the Beyerdynamic headphones. The maximum loudness of the test stimuli was set to -3dBFS according to the EBU recommendation on loudness normalization [47] to ensure a consistency in volume across all stimuli.

In the second method, henceforth known as the *filtered* method, the magnitude and phase response of each headphone was measured using the Neumann KU-100 dummy head microphone. This was achieved by playing a swept sinusoid at 44.1kHz/24bit through each set of headphones mounted on the dummy headphone. These measurements were made in the same acoustically treated room as described in the previous paragraph. The swept sinusoidal tone used to capture the headphone's frequency response was generated using Voxengo's Deconvolver [191]. Impulse responses representative of each headphone's frequency response were obtained by deconvolving the recorded swept sinusoid with the original signal. These impulse responses were of length 512. These were transformed to the frequency domain using a 512-point Fast Fourier Transform (FFT) with a hamming window

to return the resulting frequency response. These measurements were repeated 5 times and averaged to account for the variability in the characterization of the headphone transfer function due to variations in the coupling with the ear as per Kulkarni's recommendation in [100]. The resulting HpTFs can be seen in Figure 4.4. These impulse responses were then convolved with the test stimuli to simulate the sounds of the stimuli through each headphone. As per the *recorded* method, the inverse filter was applied to all stimuli to account for the coloration of the Beyerdynamic headphones. the stimuli were normalized to -3dB as per EBU recommendation [47].

In Figures 4.5- 4.8, spectrograms of the original raw recording of the stimuli can be seen, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100. It can be noted that while both the recorded and filtered methods produced reliable replications of the original signal, there are discrepancies between the two methods for frequencies above approximately 7kHz. While the decreased energy at these frequencies can be attributed to notches found in the Frequency Response of the HD650 headphones above 7kHz, it can be seen that the filtered signal retains more of this high-frequency information than the recorded method. Overall the filtered method produces a spectrogram which is far more similar to the original recording than the recorded method. This is most visible for the 'Surround' and 'Shaker' stimuli.

4.3 Testing

Two separate virtual headphone tests were conducted in this study using the stimuli described in Section 4.2.3. A total of 10 subjects took part in these MUSHRA tests with none being disqualified for results that did not satisfy the criteria of the ITU-R recommendation such as failing to identify the hidden reference and anchor signals. Subjects were gathered from the department of Electronic Engineering in Trinity College, all of which were post-graduate students working in the area of audio engineering and were familiar with MUSHRA testing. There were a total of 9 male subjects and 1 female subject while the average age of participants was 26. No subjects reported any issues in relation to their hearing. Prior to testing, each participant was guided through a demo MUSHRA test which required the participant to successfully identify both the reference signal and anchor signal before proceeding onto the main test. This training procedure was repeated for both tests and both general and spatial quality test conditions.

In the first test, listeners were presented with headphone signals virtualized using the *Filtered* method and subsequently in the second test with signals virtualized using the *Recorded* method. These tests examined the perceived general quality and spatial

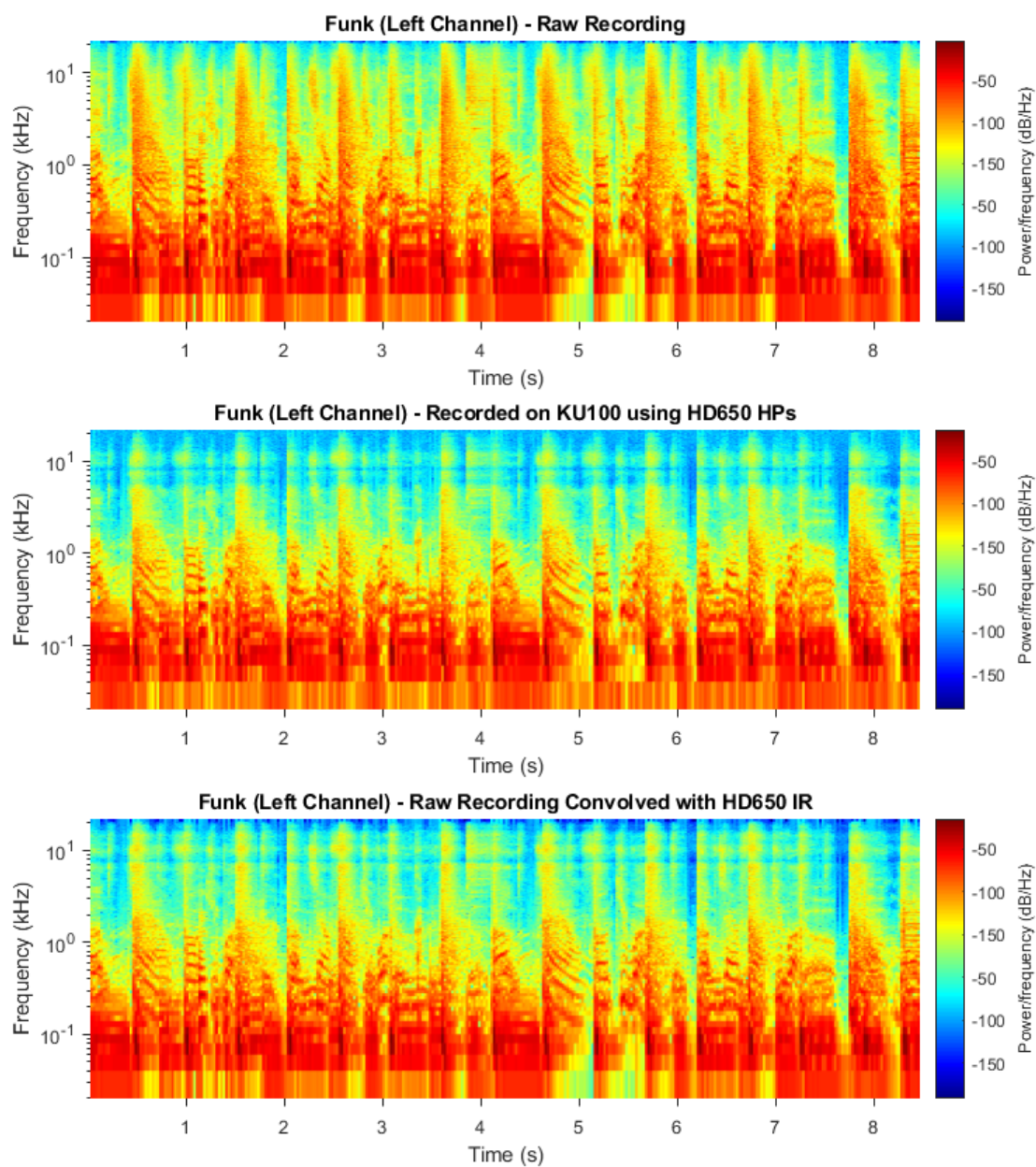


Fig. 4.5 Spectrograms of the original raw recording of the 'Funk' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.

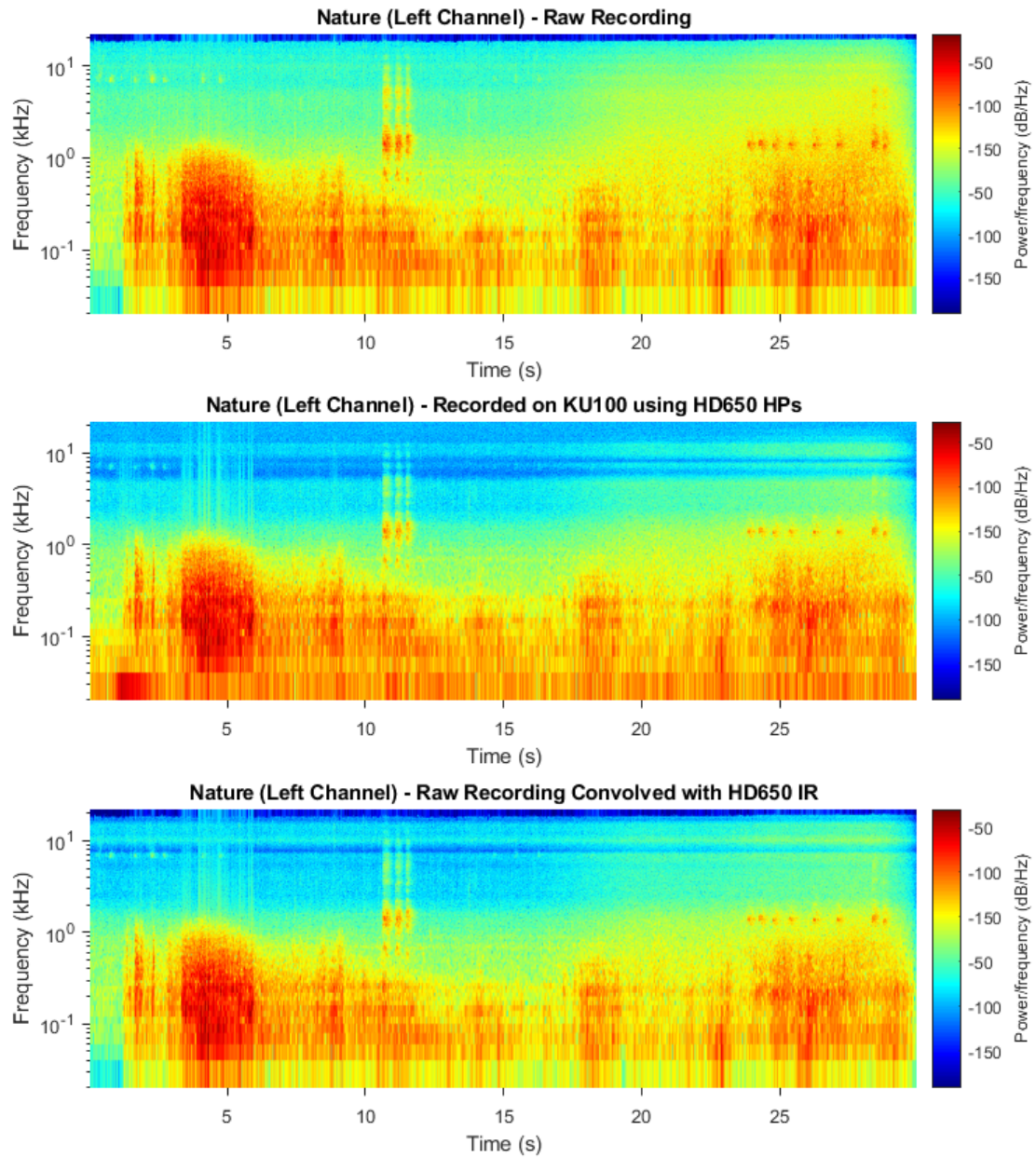


Fig. 4.6 Spectrograms of the original raw recording of the 'Nature' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.

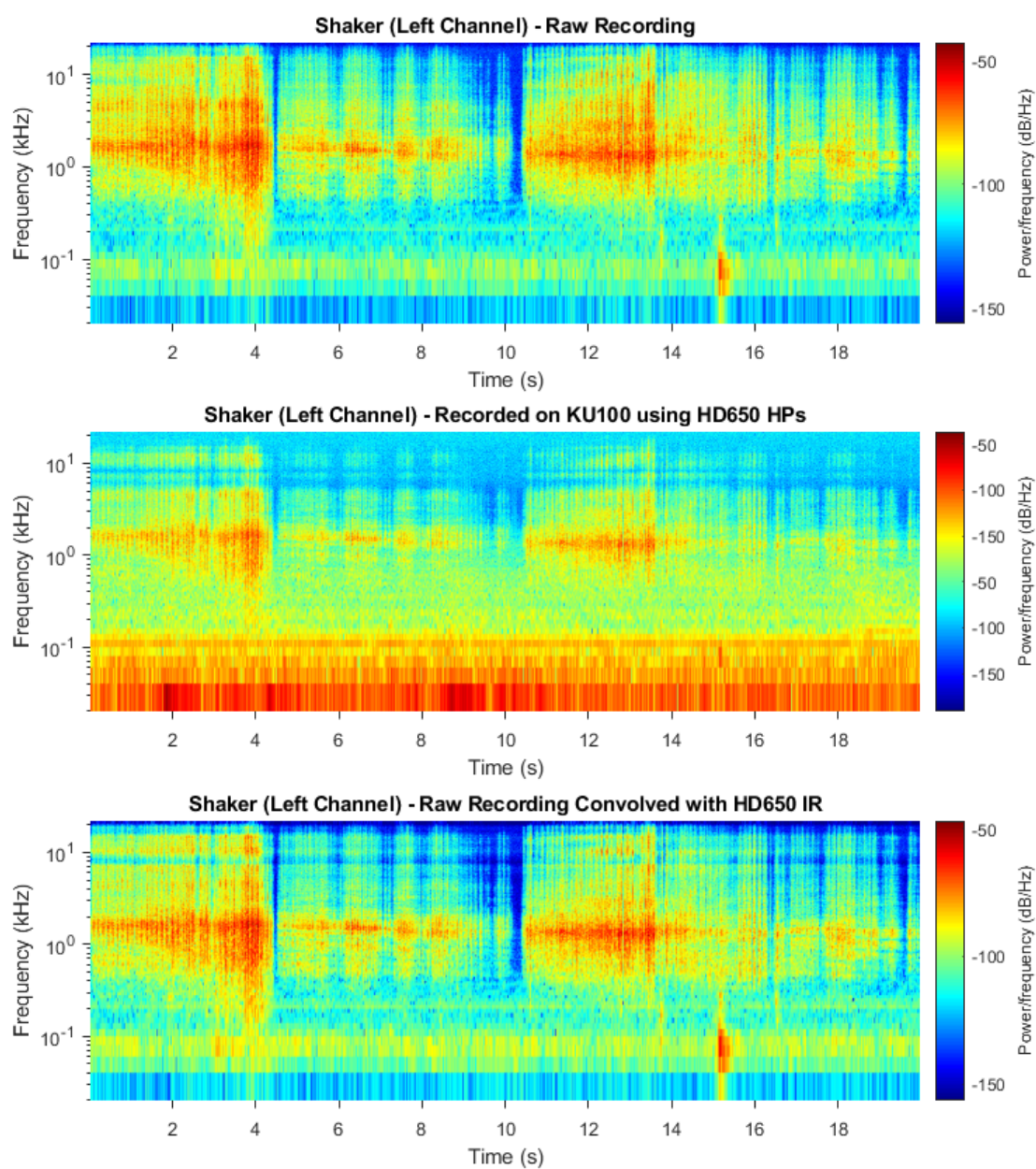


Fig. 4.7 Spectrograms of the original raw recording of the 'Shaker' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.

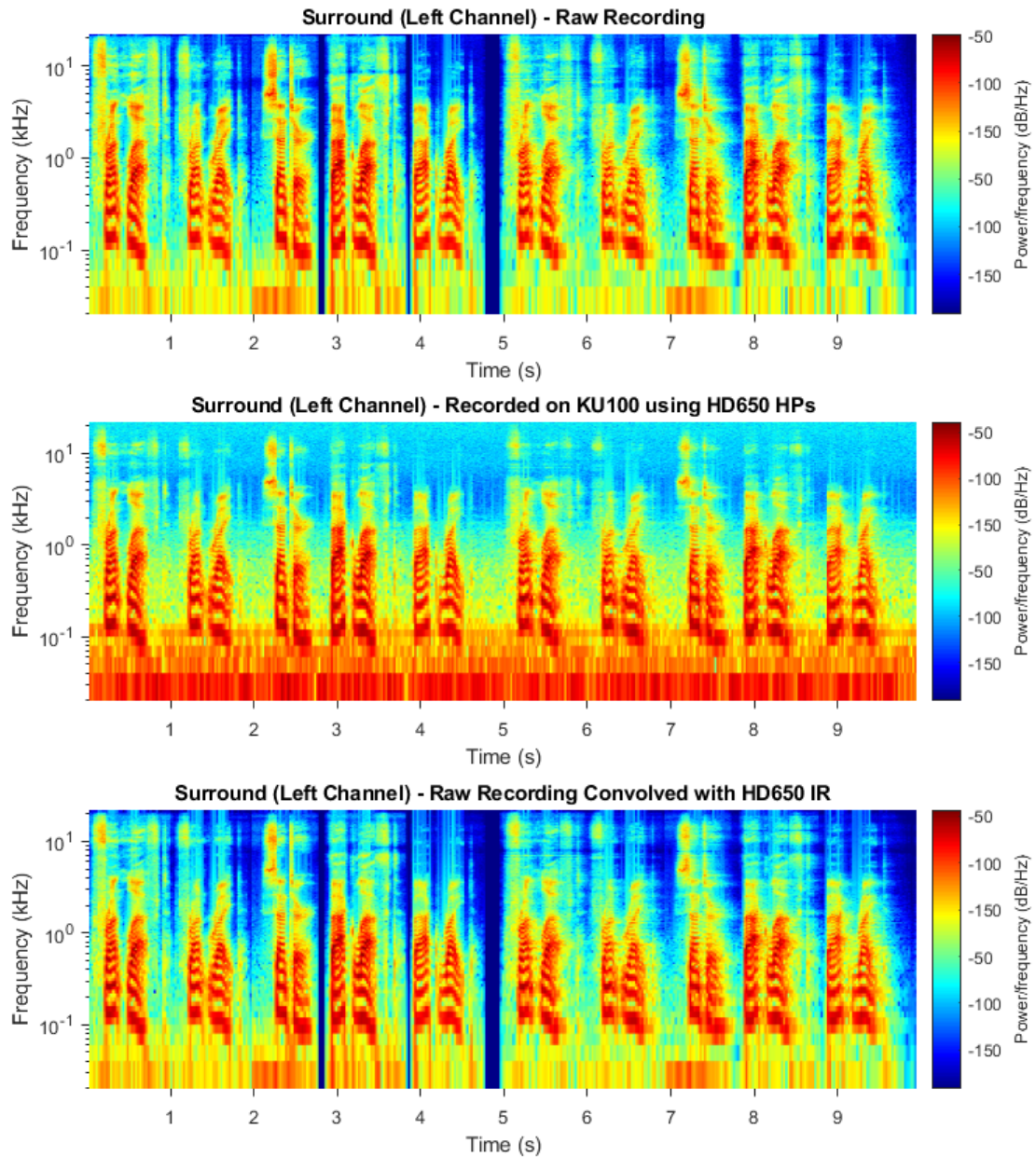


Fig. 4.8 Spectrograms of the original raw recording of the 'Surround' stimulus, the equivalent sample recorded onto the KU100 microphone using the reference HD650 headphones, and the raw recording convolved with the HD650 IR as measured on the KU100.

impression of four pairs of headphones using a MUSHRA style testing framework. In this type of listening test, subjects are first presented with a high quality reference signal. In this instance, this was the virtual renderings of the Sennheiser HD650 headphones as described in Section 4.1. After observing the reference signal, listeners were then presented with 5 additional signals - a hidden replica of the reference signal, 3 signals representative of the other headphones under test, and an anchor signal which is a low quality version of the reference signal. A low-passed rendering of the reference headphone signal to 3.5kHz was used as the anchor for questions regarding general quality and a mono rendering of the reference signal was used for questions regarding spatial impression. The anchor signals are included as per the MUSHRA test guidelines to ensure there is both a high quality and low quality reference for subjects to compare the test signals to [85].

Participants listened to these simulations using a single set of replicator headphones - Beyerdynamic DT-770 PRO, a closed back headphone. Both sets of virtual test samples were equalized to remove the effect of these headphones and were normalized to -3dB to ensure the same volume was heard across the four headphone simulations.

These perceptual tests were designed according to the International Telecommunications Union, Radio communication sector (ITU-R) recommendation BS.1534-1 [85]. This recommendation describes the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) perceptual test which is intended for spatial audio psychoacoustic tests. This test uses a reference sound to comparatively evaluate the quality of several other lower quality sounds. Using a continuous scale from 0 to 100, samples are graded by participants on their relative quality compared to the reference sample. In both tests, subjects were asked eight individual questions, each time evaluating either the general or spatial quality of the four test passages in comparison to the reference. Subjects were instructed that spatial audio quality pertained to the sense of immersion of the stimuli and their ability to localize the individual sound sources in each. General audio quality on the other hand was simply a measure of the quality of the stimuli in terms of tonal balance and preference. Subjects were asked to identify and rate the hidden reference signals with a score of 100 and rate the remaining signals accordingly. Scores were then normalized with the lowest score set to 0 and the remaining 3 scores positioned accordingly. In total, subjects attended to 80 stimuli (two tests x eight questions x five stimuli). On average testing took approximately 40 minutes to complete.

4.4 Experimental Results

A total of 10 subjects completed this study consisting of two separate MUSHRA tests each containing 8 questions. The average results and standard errors for each question have been included in Figures 4.9 and 4.10.

The correlation between all of the scores was calculated using Spearman correlation and a high correlation between all four test results have been identified as illustrated in Tables 4.3 & 4.4. Prior to determining the correlation between testing conditions, data was normalized setting the highest scoring and lowest scoring answers under test conditions to 1 and 0 respectively. All other answers were scaled in relation to this. In this instance, our null hypothesis is that the results obtained by these two virtual tests and two test criteria are statistically different and our alternative hypothesis is that the results are statistically similar. As the p-values in each instance are <0.001 a conclusion can be made from the data that there is a high correlation between both test methods and questions relating to general quality and spatial impression. This shows that the ranking of the headphones is consistent under all test conditions.

Table 4.3 Correlation coefficients (R) [148]

Test	Filtered-general	Filtered-spatial	Recorded-general	Recorded-spatial
Filtered-general	1	0.902	0.958	0.906
Filtered-spatial	0.902	1	0.926	0.931
Recorded-general	0.958	0.926	1	0.947
Recorded-spatial	0.906	0.931	0.947	1

Table 4.4 Correlation *p-values* [148]

Test	Filtered-general	Filtered-spatial	Recorded-general	Recorded-spatial
Filtered-general	-	<0.001	<0.001	<0.001
Filtered-spatial	<0.001	-	<0.001	<0.001
Recorded-general	<0.001	<0.001	-	<0.001
Recorded-spatial	<0.001	<0.001	<0.001	-

To further investigate the possible differences in the results, 1-way ANOVA testing was performed for each method and for both questions regarding general and spatial quality preference. Results across all the test samples were included in the comparison. Pairwise comparisons of all the headphone pairs were conducted with p-values being computed using Tukey's Honestly Significant Difference (HSD) test in order to determine whether

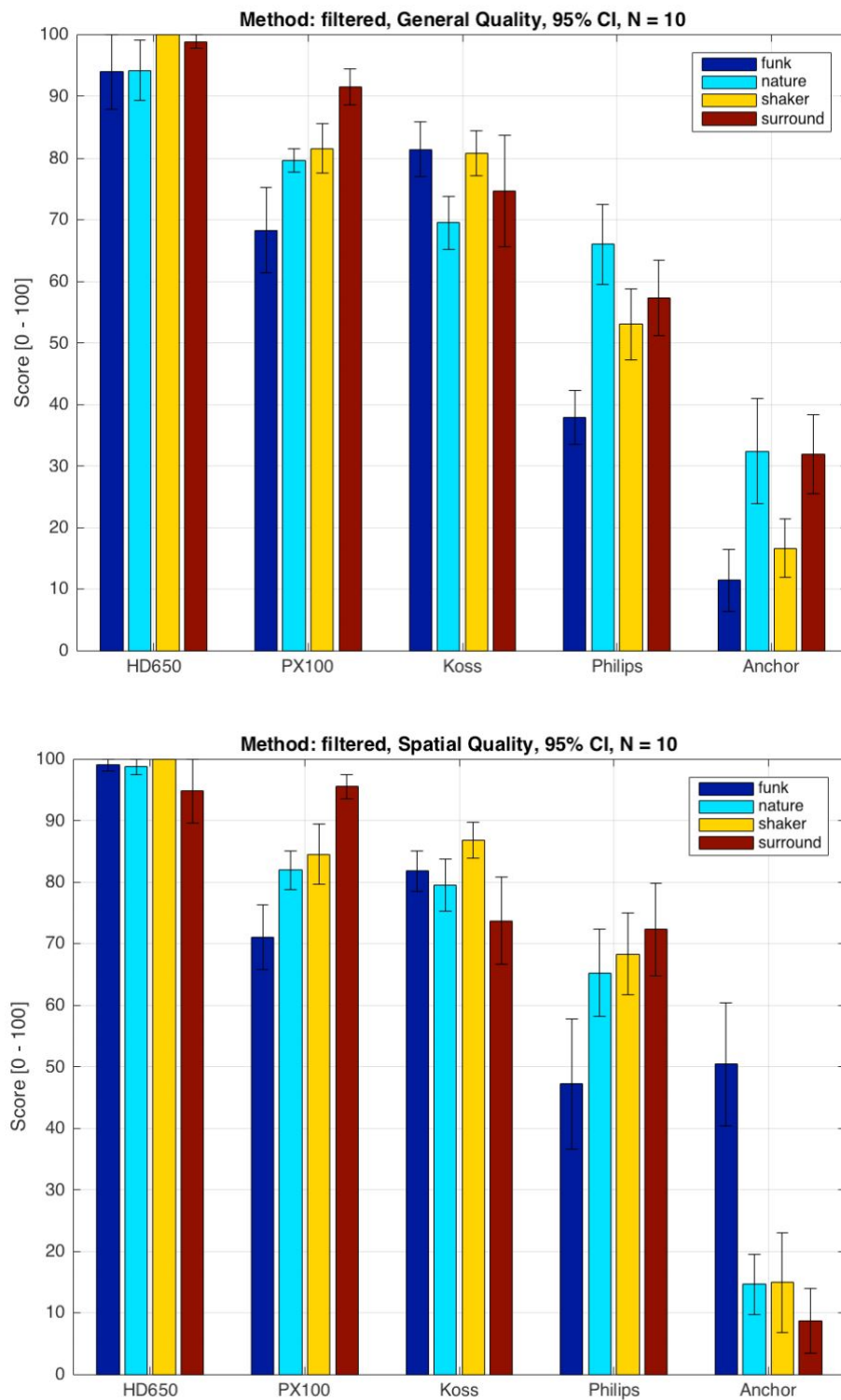


Fig. 4.9 The experimental results for the first MUSHRA test where subjects ranked the general and spatial quality of the audio signals obtained using the 'filtered' method [148]

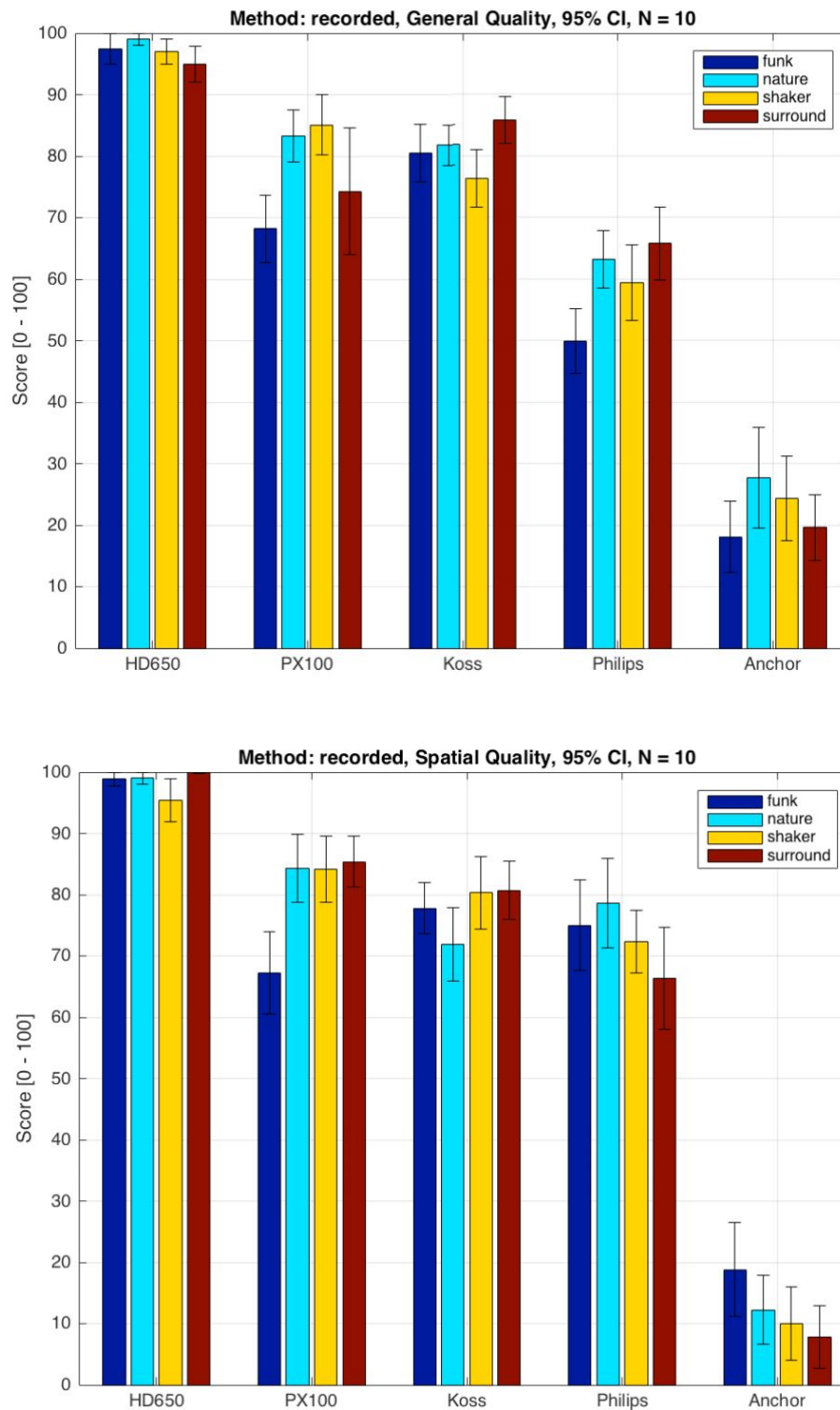


Fig. 4.10 The experimental results for the first MUSHRA test where subjects ranked the general and spatial quality of the audio signals obtained using the 'recorded' method [148]

the differences are statistically significant. Tables 4.5 to 4.8 below visualize pairwise comparisons of all the headphone pairs and use p-values computed using Tukey's HSD method in order to determine whether the differences are statistically significant, p-values larger than 0.05 (in bold) indicate that the difference is not significant.

Table 4.5 General quality preference. Test method - *filtered*

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	<0.001	<0.001	<0.001	<0.001
PX100	<0.001	-	0.89	<0.001	<0.001
Koss	<0.001	0.89	-	<0.001	<0.001
Philips	<0.001	<0.001	<0.001	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-

Table 4.6 Spatial quality preference. Test method - *filtered*

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	0.008	<0.001	<0.001	<0.001
PX100	0.008	-	0.97	<0.001	<0.001
Koss	<0.001	0.97	-	0.001	<0.001
Philips	<0.001	<0.001	0.001	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-

4.5 Discussion

As can be seen in Figures 4.9 & 4.10, there is a similar ranking of the four headphones for both the *filtered* and *recorded* methods as well as for the questions relating to perceived quality and spatial impression. Although not exactly the same across all conditions, a consistent trend between headphone preference can be seen. Across all conditions, the HD650 headphones were correctly identified as the reference signal while the anchor signal was correctly identified in a majority of cases with one noticeable exception. When rating the spatial quality of the monaural anchor for stimulus A in the *filtered* test, a relatively high score was achieved compared to the other monaural anchors and compared to the monaural anchor in the recorded stimuli. This result is most likely due to the high quality

Table 4.7 General quality preference. Test method - *recorded*

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	<0.001	<0.001	<0.001	<0.001
PX100	<0.001	-	0.9	<0.001	<0.001
Koss	<0.001	0.9	-	<0.001	<0.001
Philips	<0.001	<0.001	<0.001	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-

Table 4.8 Spatial quality preference. Test method - *recorded*

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	<0.001	<0.001	<0.001	<0.001
PX100	<0.001	-	0.96	0.34	<0.001
Koss	<0.001	0.96	-	0.76	<0.001
Philips	<0.001	0.34	0.76	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-

of this particular music sample and how it was most likely mixed to still provide a balanced, high quality listening experience even when played in mono.

As was shown in [70], the results shown for the *filtered* test methodology show a clear correlation between perceived general audio quality and spatial quality. The p-values found using Tukey's HSD test and presented in Tables 4.5 and 4.6 show a significant difference in the scores associated with each pairwise comparison ($p < 0.001$) with the exception of the PX100 and KOSS headphones ($p = 0.89$ for general quality preference and $p = 0.97$ for spatial quality preference). This indicates that the testing under both conditions fails to differentiate a preference rating between these two headphone signals. While p-values between the HD650 and PX100 headphones ($p = 0.008$) and between the KOSS and Philips headphones ($p = 0.001$) for spatial quality preference differ from those found for general quality preference, these values are still well below 0.05 indicating there is no significant difference in the results found using this statistical analysis.

Insignificant differences between the rating of the KOSS headphones and the Sennheiser PX-100 headphones are consistent for both general and spatial quality for the *filtered* stimuli and for the questions relating to general quality for the *recorded* stimuli. The only exception is that the Spatial Quality test in which the *recorded* method fails to find

significant differences between these headphones and the Philips headphones. In Table 4.8, we can see that when using the *recorded* test methodology, questions regarding spatial audio quality revealed statistically similar scores for all three of the Koss, PX100, and Philips headphones. While the p-values are smaller for comparisons between the Philips headphones and the other two pairs ($p = 0.34$ for the PX100 headphones and $p = 0.76$ for the KOSS headphones), this result adds ambiguity to the rating of the headphones and an inconsistency between both test conditions in the *recorded* test methodology. Similar to the results presented in [78], this result shows an unreliability in the *recorded* virtual test methodology. If a conclusion can be drawn from the results found using the *filtered* methodology that general quality preference and spatial quality preference are consistent, then the results shown in Tables 4.7 and 4.8 show that the *recorded* test methodology is inconsistent and unreliable.

The ranking of the headphones under test across multiple testing conditions is in general agreement with the Philips headphones consistently performing worse than both the PX100 and KOSS headphones. A lack of distinction between the PX100 and KOSS headphones is also consistent across all conditions. With the results found in this study it is not possible to rank these two headphones in relation to each other and further testing with a larger cohort of subjects would be required if such a distinction were to be made. Increasing the number of headphones under test could result in a distinction being made however an increase in test stimuli for MUSHRA testing would require the introduction of a blocked design of the experiment [85] to avoid listener fatigue.

Spectrograms of the stimuli produced using both *recorded* and *filtered* virtual headphone testing methodologies show that stimuli created using impulse response convolution are more similar to the original signal than those recorded onto a binaural headphone. The recorded method results in a build up of low frequency content and the presence of background noise across the entire spectrum. While the use of on-ear headphones does increase the build up of low frequencies it seems that this may account for the lack of distinction between headphones in terms of spatial audio quality.

4.6 Conclusion

This study aimed to demonstrate whether two different methods of virtual headphone testing produced statistically similar results and to verify whether a correlation exists between perceived general quality and spatial impression when using a *recorded* style virtual headphone testing methodology. Previous studies have concluded that using HpTFs to generate stimuli for virtual headphone testing produces similar results to that of physical headphone testing [145, 70]. Other studies suggest that recording stimuli on to binaural

microphones does not produce results that are significantly consistent with physical headphone testing [78]. It has also been shown that for questions regarding perceived audio quality and questions regarding spatial impression, results are significantly similar in a MUSHRA type test [70] using a *filtered* style virtual test.

This work has found that *recorded* and *filtered* virtual headphone testing methodologies are significantly different. While there is a high correlation between both test methodologies for preference in terms of both general audio quality and spatial audio quality found using Spearman's correlation. 1-way ANOVA testing with p-values computed using Tukey's HSD test for pairwise comparison show that in terms of spatial quality preference, the *recorded* methodology fails to distinguish a difference in quality for a single pair of On-Ear headphones, a distinction easily made using the *filtered* methodology. In agreement with results presented in [70], this study found that when using a *filtered* virtual headphone test methodology, there is an agreement in preference of headphone in terms of general audio quality and spatial audio quality.

Since this study was first presented in 2016, more work has been done in the field of virtual headphone testing. Of particular note is a study published by Welti et al. in 2016 [199]. Here a listening test was employed comparing In-Ear (IE) headphones using two testing methods, one which simulated the frequency response of each headphone and another which played recordings of each headphone producing the same stimuli. The Pearson correlation coefficient between the actual and virtualized headphone ratings was $r = 0.98$ indicating agreement although these results were found for In-Ear (IE) headphones and not circumaural headphones as were used in this study. The recording of in-ear headphones using a binaural head microphone is subject to far less error than when recording on-ear headphones as was performed in this work. For instance, improper seating of the headphones during recording can affect the perception of their quality. When measuring a HpTF as was done in the *filtered* methodology in this study, multiple re-seatings of the headphone are performed and an average response is taken to counteract this issue. With this in mind, a conclusion can be drawn that for the comparison of on-ear headphones, a *filtered* virtual headphone test methodology is preferred.

Chapter 5

Machine Learning and its Applications in Audio Engineering

The fields of Acoustics and Audio Engineering are sciences with many quantifiable properties forming the basis of our understanding in both. Recent advances in machine learning (ML) have allowed us to gain a better insight into these topics through deep learning and other methods. ML consists of a wide variety of different techniques which can automatically detect and utilize patterns in data. ML techniques are capable of discovering complex relationships between acoustic features as well as developing models to describe complex acoustic phenomena such as human speech and reverberation. In this chapter we will give an overview of ML and its fundamental concepts as well as describing its use in audio engineering. We will discuss in detail the performance of Supervised Learning and why it is the ideal branch of ML for solving the sound source localization problems discussed in later chapters of this thesis.

5.1 Machine Learning

Machine Learning is a term used to refer to computer algorithms that can automatically detect and utilize patterns in data [18, 137]. These algorithms are typically trained using data from past events to make accurate predictions for future events. The concept of machine learning is inspired by learning performed by humans whereby an action can be observed and repeated. However, algorithms require many more examples of an action than a human would before being able to reliably replicate a behaviour. The experience that is available to a learning algorithm typically is in the form of electronic data collected and made available for analysis. This data can take many forms including human labelled data sets, or other information types which may or may not be labelled. Crucial to the

success of a machine learning algorithm is the quality and size of the training data available to it.

An ideal machine learning algorithm is both efficient and accurate in its prediction. The time it takes to train an algorithm and its complexity are both critical measures of its quality. In machine learning we must also consider the notion of sample complexity and evaluate the sample size required for the algorithm to learn a particular concept. The field of machine learning is intrinsically linked to data science as the success of an algorithm is dependent on the data used. Therefore, we must also be familiar with ideas from statistics, probability and optimization to fully understand machine learning. With advances in computational power and the availability of large datasets, machine learning has proven to be an effective tool in many fields including but in no way limited to Natural Language Processing (NLP), Speech Processing, Computer Vision, and Document Classification.

Some standard machine learning tasks can be listed as follows:

- **Classification:** this is the problem of assigning a category for a given set of input data. For instance, determining the sex of a speaker based on a sample of speech.
- **Regression:** In this problem, a real numerical value must be predicted for a set of input data. For instance, predicting the price of a house based on its size.
- **Ranking:** this is the problem of learning to order items according to some criterion. For instance, returning web pages relevant to a search query.
- **Clustering:** This is the problem of partitioning a set of items into homogeneous subsets. For example, within social networks, clustering algorithms attempt to identify natural communities within large groups of people.
- **Dimensionality reduction or manifold learning:** This problem consists of transforming an initial representation of items into a lower-dimensional representation while preserving some properties of the initial representation. A common example involves pre-processing digital images in computer vision tasks.

In this thesis, both classification and regression are used to complete SSL tasks. A clustering algorithm in the form of Affinity Propagation is also used in the work presented in Chapter 8.

5.1.1 Machine Learning Terminology

Machine Learning algorithms are typically generated using three unique sets of data - training data, validation data, and testing data. The algorithm uses training data to make its initial observations about the problem it is trying to solve. The choices it then makes

are referenced with validation data to ensure its assumptions are correct. Finally the fully trained algorithm is tested using testing data, a dataset of completely unseen examples. The performance of a machine learning algorithm is typically measured as the percentage of test data examples it correctly labels. Each Machine Learning algorithm is unique and thus has its own distinct properties. The following is a list of terms and vocabulary that is commonly observed when describing a machine learning algorithm:

- **Examples:** Items or instances of data used for learning or evaluation.
- **Features:** A set of attributes, often represented as a vector, that constitute an example. These are also referred to as inputs.
- **Labels:** Values or categories assigned to examples. These are also referred to as outputs.
- **Hyperparameters:** Free parameters that are not determined by the learning algorithm, but rather are specified as inputs to the learning algorithm.
- **Training Sample:** Examples used to train an algorithm.
- **Validation Sample:** Examples used to tune the parameters of a learning algorithm when using labelled data.
- **Test Sample:** Examples used to evaluate the performance of a learning algorithm.
- **Loss Function:** A function that measures the difference, or loss, between a predicted label and a true label.
- **Hypothesis Set:** A set of functions mapping features (feature vectors) to the set of labels, y .

Following the successful training of an algorithm, its performance is evaluated based on its error on the set of test samples.

Machine learning scenarios can differ based on the types of training data available for training, the order and method by which training data is received and the test data used to evaluate the learning algorithm. The following are a list of some of the most common machine learning scenarios:

- **Supervised learning:** For supervised learning to be used, the learner must receive labelled training and validation data to make predictions for unseen data vectors. This is the most common scenario associated with classification, regression, and ranking problems.

- **Unsupervised learning:** In these scenarios, the learner only receives unlabelled training data and makes predictions for all unseen points. Since there are no labelled examples to compare against for results, it can be difficult to evaluate the performance of unsupervised learning algorithms. Clustering and dimensionality reduction are examples of unsupervised learning problems. One such application for unsupervised learning is recommendation algorithms on streaming services such as YouTube, Netflix, and Spotify.
- **Semi-supervised learning:** The learner receives a set of training samples consisting of both labelled and unlabelled data, and makes predictions for unseen points. This distribution of unlabelled data accessible to the learner can help achieve a better performance than in a supervised setting. This can be useful for classification, regression, or ranking tasks.
- **Reinforcement learning:** In reinforcement learning, the learner actively interacts with an environment and receives a reward for each action. The objective for the learner is to maximise its reward over a course of actions and iterations with the environment. Reinforcement learning can be used to teach a learner how to successfully play a game. A famous case of this was when Google and DeepMinds AlphaGo program successful beat the worlds best Go player in 2016 [178].

5.1.2 Generalization

Fundamental to Machine Learning is the concept of generalization. This is the ability of learned models to perform well on unobserved data. Machine learning algorithms are formulated by selecting a function out of a hypothesis set, that is a subset of the family of all functions. This function is subsequently used to label all instances, including unseen examples. Complexity is used to describe the level of sophistication of the data relationships or ML task and the ability of an ML model to approximate relationships of a certain complexity is the capacity. Both complexity and capacity are not strictly defined but there has been some effort to mathematically define them. For instance, the Vapnik-Chervonenkis (VC) dimension provides a means of quantifying model capacity in the case on binary classifiers [24]. A complex hypothesis set may be selected allowing the learner to choose a function or predictor that can successfully categorise all of the training samples. Conversely, a function chosen from a less complex hypothesis may incur some errors on the training samples. However, the best predictor on the training data may not be the best predictor overall as a predictor from a complex family can essentially memorize the training data meaning it may not be robust to any data that differs from the training data. Generalization is affected here by the trade-off between the sample size and complexity.

With a small sample size, choosing too complex a family may lead to poor generalization, otherwise known as *overfitting*. When too simple a family is used, it may not be possible to achieve a sufficient accuracy, which is known as *underfitting*. Both underfitting and overfitting degrade the performance of an ML model and can be quantified using the bias and variance of the ML model. The bias is the difference between the mean of our estimate targets and the true mean, and the variance is the expected squared deviation of the estimated targets around the estimated mean value [74].

The performance of an ML algorithm is estimated based on its performance on unseen observations, which in turn assesses its generalization. To do this, generally a set of test data is removed from the training data before the algorithm is trained. The performance of the algorithm on the test data is used to hone the algorithm by adjusting hyperparameters such as the number of layers and nodes. As often, there is a lack of sufficient data to rely on a single set of test data, a technique known as K-fold cross-validation is commonly employed [97]. Here, the model is trained and tested k times using a different split in the training and testing data. The cross-validation score is then taken as the average performance on test data over the k algorithms.

5.2 Machine Learning and Audio Engineering

Research in acoustics and audio engineering has traditionally focused on the development of high-level physical models to infer properties of an environment and objects within that environment. For example, the calculation of reverberation time in a room based on its size and the materials used for its construction. ML can build upon these physical models using large amounts of data to improve on our interpretation of how certain features may relate to certain properties. That is, ML can find some representation of a set of features that is optimal for performing a particular task.

In recent years, ML and in particular Deep Learning (DL) have yielded promising advances in acoustics. With access to large datasets, DL approaches have provided results comparable to highly complex signal processing methods. While often, there is insufficient data to apply DL to tasks, the ability to synthesise training data and the use of data augmentation has allowed for large training sets of audio signals to be created. Examples of audio related tasks where DL has been applied include speech modelling, source separation, and enhancement. These include a deep clustering algorithm which separates two speakers of the same gender which uses a DNN to find embedding features for each time-frequency region of a spectrogram [76]. A dereverberation algorithm which uses DNNs to remove reverberation from speech recordings using a single microphone is shown in [49].

Two tasks where DL algorithms have excelled in acoustics and audio signal processing are sound event detection and sound source localization. In both problems, DL methods have replaced physics based acoustic propagation models to achieve state of the art results. In [33], Cakir et. al showed that convolutional neural networks (CNNs) yielded optimal results in the 2017 Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [131]. CNNs were also shown to be proficient in sound source localization tasks as shown by Chakrabaty and Habets in [34]. Using synthetically generated noise, they used only the phase component of the STFT to predict the direction of arrival (DOA) in the horizontal plane. This work showed results comparable to the physics based steered response power phase transform (SRP-PHAT) beamforming method [28], and even generalized well to speech signals. Recent studies have shown the effectiveness of using DL methods to perform sound source localization using binaural signals. In [119] & [120], Ma et al. present a machine learning framework to determine the azimuthal position of a speech source in both multi-speaker and reverberant environments. A similar network is employed in a study by Lovedee-Turner and Murphy [115] which estimates the DOA from the raw HRIR signal. In [207], Xu et al. used machine learning to estimate source location using two 100-channel cochlear models to analyse binaural signals. Each channel in the left cochlea is compared with each channel in the right cochlea in parallel to generate a 2-D instantaneous correlation matrix (correlogram). These correlograms are then used as the input to a CNN which performs sound source localization with an average absolute error of 12.8° in the -90° to 90° range. This model performed SSL in the horizontal only and was expanded in [208] where a root mean square localization error (RMSE) of 3.68° was reported.

Sound source localization and in particular, speaker localization is directly applicable to many other audio related tasks such as automated camera steering, teleconferencing systems, and robot audition. Such applications must have robust solutions in the presence of noise, reverberation, and other adverse conditions. These conditions remain difficult to overcome and have inspired a recent challenge on acoustic source localization and tracking (LOCATA), endorsed by the IEEE Audio and Acoustic Signal Processing technical committee who have established a database to encourage researchers to test their algorithms [113]. This database allows for the assessment of source localization algorithms in real-life scenarios

5.3 Supervised Learning

The problem examined in this thesis is that of sound source localisation using binaural audio. The feature sets in these instances consist of monaural and interaural features

extracted from the left and right audio channels. As the signals have been synthesised by convolving HRIRs with monaural signals, we know the location of these binaural signals, i.e. we know both the azimuth and elevation corresponding to each feature set. Therefore we have a labelled set of features to train the learner on. Because of this, it is evident that this particular problem is one that would be best classified as a supervised learning scenario.

Supervised learning aims to learn a mapping from a set of inputs to desired outputs given labelled input and output pairs.

Based on the type of output $\mathbf{y}$, supervised learning can be split into subcategories: regression and classification. When $\mathbf{y}$ is a real or complex numerical value, the task is regression. When $\mathbf{y}$ is categorical, the task is classification.

With all ML methods, we aim to find the function f that describes the relationship between input x and output y . In most instances we use the tools of probability to find f , i.e. we can describe the task of supervised ML as maximising the conditional distribution $p(\mathbf{y}|\mathbf{x})$. One such example is the *maximum a posteriori* (MAP) estimator

$$\hat{\mathbf{y}} = f(\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}} p(\mathbf{y}|\mathbf{x}), \quad (5.1)$$

which gives the most probable value of $\mathbf{y}$, based on the mode of the distribution conditioned on the observed evidence $p(\mathbf{y}|\mathbf{x})$.

5.3.1 Multi-Layer Perceptrons: Neural Networks

Deep Neural Networks (DNNs) are a type of machine learning algorithm that are trained to model a set of input variables to another set of output variables. An artificial neural network consists of 3 components: an input layer, hidden (computational) layers, and an output layer. Learning takes part in two separate stages called Forward-Propagation and Back-Propagation. During Forward-Propagation, the algorithm makes a guess about the output of the network. Back-Propagation then refines the network to minimise the error between the actual answer and the guessed answer. During training, a set of input data along with its corresponding output data is given to the network and using this data it applies appropriate values to the weights and biases of the Hidden Layers to best model the data. This training process finds the closest mathematical manipulation that equates the input data to the output data. This relationship is either classified as linear or non-linear.

The simplest deep learning algorithms are classification algorithms which seek to assign a set of input variables to a particular output class. This could be a task such as identifying a letter or number in an image. This is seen in the ubiquitous MNIST database which presents the task of correctly identifying handwritten numbers [44].

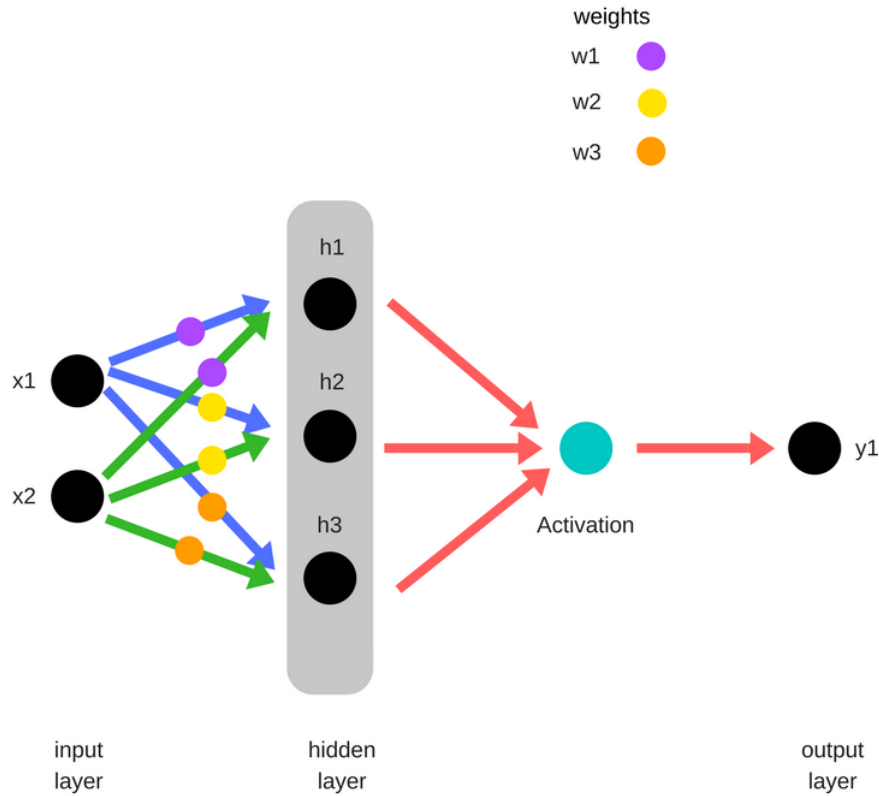


Fig. 5.1 A neural network with an input layer, hidden layer and an output layer

The relationship between input features and outputs in a linear classification problem can be defined as follows:

$$WX + b = Y \quad (5.2)$$

where X and Y are the inputs and outputs respectively, W represents the weights attached to the input vector and b is a bias. The number of nodes in the input, output and hidden layers determines the complexity of the algorithm. For the network shown in Figure 5.1, there are 2 input nodes and 3 hidden nodes; the values of these hidden nodes are given as:

$$\begin{aligned} h_1 &= (x_1 * w_1) + (x_2 * w_1) \\ h_2 &= (x_1 * w_2) + (x_2 * w_2) \\ h_3 &= (x_1 * w_3) + (x_2 * w_3) \end{aligned} \quad (5.3)$$

where the weights have random initial values. The output of the algorithm, y is given as a function of these hidden node values. This function is referred to as an *activation function*.

Over the course of many iterations, these weights and bias values are trained to attain the closest possible relationship between the input and output variables. For classification problems, each possible class in y is given a score or 'logit' based on the likelihood that it is the correct class. These scores can be converted to probability values, p which sum to 1 using the 'Softmax' activation function defined as:

$$S(y_i) = \frac{err_{y_i}}{\sum_j err_{y_j}} \quad (5.4)$$

where i is the current class of output and j is the total number of output classes. Following Softmax encoding, probabilities for each class are usually assigned a value of 1 or 0 using what's called 'One-Hot' encoding. Here, each label is represented by a vector which is as long as the number of classes where the highest probability value is assigned a new value of 1 and all other classes are given a value of 0. Cross-entropy defines the distance between 'logit' probability scores and their one-hot encoded counterparts. It is defined as:

$$D(S, L) = - \sum_i L_i \log(S_i) \quad (5.5)$$

where S represents the probability scores and L is its 'One-Hot' encoded equivalent. This entire process is called Multinomial Logistic Classification and is defined as:

$$D(S(wx + b), L) \quad (5.6)$$

The next step in designing any Deep Learning algorithm is to define a loss function. In this particular example, the loss function is defined as the average cross-entropy as follows:

$$\mathcal{L} = \frac{1}{N} \sum_i D(S(wx_i + b), L_i) \quad (5.7)$$

where $\mathcal{L}$ is the loss and N is the total number of estimations made in that particular iteration. The goal of Deep Learning is to minimise the value of this loss function with each iteration by adjusting the values of W and b . This is done using an *optimizer*. One common way to do this is to use Gradient Descent where weight values are adjusted by some fixed amount in either the positive or negative direction in order to reach a local or global minimum value.

During training, expected values of the output are compared to values attained from forward propagation by passing them through a *loss function*. This loss function computes

the error between the correct output and the estimated one. While the loss function can be a simple subtraction, it is more commonly a function such as 'Mean Square Error' (MSE) which compares the results of multiple estimations. Partial derivatives of error are then calculated with respect to each weight which measure the contribution of each weight to the total loss. These partial differentials are then multiplied by a learning rate constant, η . The learning rate is a hyperparameter that controls how much change the model undergoes in response to estimated error each time the model weights are updated. The result is then subtracted from the respective weights and the updated weights are given as:

$$\begin{aligned}w1 &= w1 - (\eta * \frac{\delta(err)}{\delta(w1)}) \\w2 &= w2 - (\eta * \frac{\delta(err)}{\delta(w2)}) \\w3 &= w3 - (\eta * \frac{\delta(err)}{\delta(w3)})\end{aligned}\tag{5.8}$$

While the role of hidden layers is to map the shape of the underlying function in the data, the role of bias is to laterally shift the learned function so it overlaps with the original function. The learning rate is important to configure at this stage in developing a neural network. As deep neural networks use stochastic gradient descent results may vary each time the algorithm is trained. Using a learning that is too small may result in a long training process that gets stuck at a relative minimum value. If the learning rate is too large, weights will be updated rapidly and may result in an unstable training process.

After a certain condition is met, the algorithm will cease training and be formed with the values for the weights and biases it currently has. This condition may be a set number of iterations or perhaps once the loss value reaches a certain value. Two important terms to note here are *epoch* and *batch*. The term epoch is used to define the number of times the learning algorithm will work through the entire training dataset while the batch size defines the number of samples the model will work through before updating the internal model parameters. Batch size can be anywhere between 1 to the size of the entire training set. In most, DNNs mini-batch gradient descent is employed where the batch size is some fraction of the entire training set. This means the model weights and biases will updated several time within each epoch.

Returning for a moment to activation functions. If the problem the algorithm is trying to solve is not a classification problem, an alternate activation function at the output layer can be used such as a 'linear' one. In this instance, instead of mapping the output to a particular class, it may pick a value for which there is no example in the training data. This is done using interpolation which allows the algorithm to select numerical values for the output which may be linear or non linear based on the outputs seen in the training data.

One way in which performance of a Deep Learning algorithm can be performed is by preprocessing the training data. This usually means ensuring that it has a mean equal to 0 and unit variance. This conditioning makes it much easier for the optimiser to reach a good solution to the problem. When performing optimization it's also important to select good initial values for the weights and biases. The simplest method is to draw these weights randomly from a gaussian distribution with mean, $\mu = 0$ and standard deviation, σ .

5.3.2 Decision Trees and Random Forests

Decision Trees are another type of supervised learning algorithm that are mostly used in classification problems but can also be used to perform regression. They are capable of using both categorical and continuous input and output variables. Decision trees aim to split a population or set of samples into two or more homogeneous sets based on the most significant splitting variable from the input features. A type of ML model consisting of multiple decision trees called a Random Forest was used by Wu and Talagala [204] to perform SSL in both Horizontal and Vertical Planes in the presence of noise producing competitive results.

A decision tree begins with the root node which contains the entire population as shown in Fig 5.2. The population within the root node is then divided into two or more sub-nodes based on a single input variable. When a sub-node is split into further sub-nodes based on another input variable, these are referred to as decision nodes. Nodes that do not split further are called leaf or terminal nodes. A node that is divided into sub-nodes is called the parent of these sub-nodes whereas the sub-nodes are the children of the parent node. A branch refers to a sub-section of the entire tree. Pruning is the process of removing sub-nodes from a decision tree, this process could be considered the opposite of splitting.

Decision trees are useful in determining the most significant variables in estimating an output variable and are often used in data exploration. Decision Trees are considered to be non-parametric meaning they have no assumptions about the space distribution and the classifier structure. They are prone to over fitting but this can be addressed using model constraints and pruning.

The feature importance of each variable can be calculated using the Gini Impurity Index or Information Gain. Gini is used in Classification and Regression Trees (CART) to create binary splits in a population. This is done for each subnode by finding the sum of the square of probability for success and failure, i.e. $p^2 + q^2$ where p is the probability that population will be split into one category and q is the probability it will be split into the other. The variable which yields the highest Gini value is chosen as the splitting variable. Gini Impurity is found by subtracting the Gini value from 1.

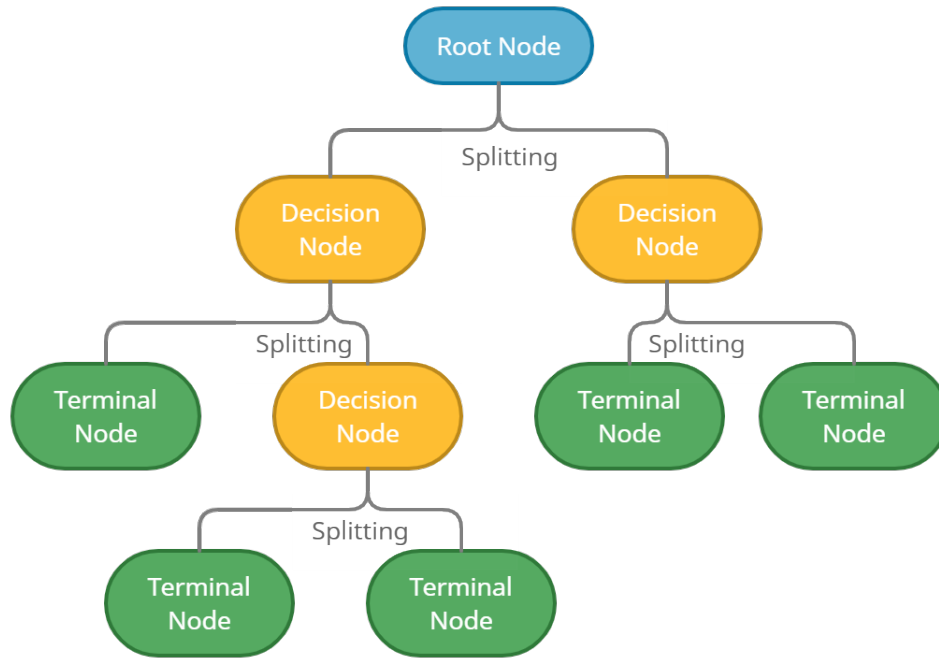


Fig. 5.2 A Decision Tree

Information gain is another method for determining feature importance. Information gain uses entropy to define the purity of a node, i.e. if the population within a node is homogeneous under a variable, then the entropy of that population is 0 where entropy is defined as

$$E = -p \log_2 p - q \log_2 q. \quad (5.9)$$

The entropy for a split is calculated by finding the entropy of a parent node and each individual node in a split and then finding the weighted average of all sub nodes available in a split. Information gain is calculated as $1 - \text{Entropy}$ where the variable that yields the highest value is chosen as the splitting variable.

When designing a decision tree model, various constraints can be specified to avoid over fitting. These include a minimum number of samples required for a node split, a minimum number of samples required for a node to become a terminal node, the maximum depth of the tree, a maximum number of terminal nodes, and the maximum number of features considered to split a node.

Decision trees like all other learning based algorithms have both bias and variance. While in other ML methods, these can be decreased by increasing the complexity of the model, in decision trees this often leads to over-fitting. This problem can be addressed by using ensemble learning. Bagging is one such example of ensemble learning where

a population is split into multiple subsets and a group of classifiers are trained on each separate subset. The prediction of each classifier is combined using a mean, mode, or median value as the overall prediction. A Random Forest (RF) is another type of ensemble learning considered to be a versatile solution for data science problems. In Random Forest models, multiple decision trees are grown, each developing their own classification based on a subset of the input variables. The output of the RF is then taken to be the mode of the outputs from each tree in the case of classification or the mean in the case of regression.

5.3.3 Mutual Information

Information gain can also be referred to as *mutual information* and it is an effective way of determining the most effective features in training an ML algorithm. For instance, in [204], mutual information was used to examine which frequencies along the magnitude and phase spectrum for binaural audio are best for estimating the elevation of a sound source. These estimations were performed using a sampling rate of 8kHz and have been performed again in this thesis at a sampling rate of 44.1kHz allowing mutual information to be calculated for frequencies up to 20kHz. To perform this, speech signals from the TSP speech database [90] were convolved with a HRIRs from subject 3 of the CIPIC HRIR database [5]. The frequency response of these signals was then calculated using a 2048-point FFT with a window size of 50ms and a 25ms overlap. ILD and IPD values were then taken measured across all frequencies from 0-20kHz for each HRTF in the dataset ranging from -80° to 80° along the horizontal plane and -45° to 45° along the vertical plane. Mutual information for each cue at each frequency for estimating elevation was found using open-source mutual information for max-dependency code made available by Peng et al. [150].

Mutual Information is widely used to determine feature selection that is maximally relevant for a particular outcome. The MI of two discrete variables, x and y , is defined as

$$MI(x, y) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)}, \quad (5.10)$$

where $P(x)$ and $P(y)$ are the joint probability and marginal probabilities of variable x and y . Similarly, the MI between an interaural feature μ_f and a source angle (θ, ϕ) is given by:

$$MI_\theta(\mu_f; \theta) = \sum_{\mu_f} \sum_{\theta} P(\mu_f, \theta) \log \frac{P(\mu_f, \theta)}{P(\mu_f)P(\theta)} \quad (5.11)$$

$$MI_\phi(\mu_f; \phi) = \sum_{\mu_f} \sum_{\phi} P(\mu_f, \phi) \log \frac{P(\mu_f, \phi)}{P(\mu_f)P(\phi)} \quad (5.12)$$

With regards to elevation localization, given a training data set denoted as $\mu^{(n)}, \phi^{(n)}_{n=1}^N$ where $n = 1, \dots, N$; probability components, $P(\mu_f, \phi), P(\mu_f)$ and $P(\phi)$ can be calculated using histogram based probability estimation as:

$$P(\mu_f; \phi) = \frac{1}{N} \sum_{n=1}^N z(\mu_f^{(n)} = \mu_f, \phi) \quad (5.13)$$

$$P(\mu_f) = \frac{1}{N} \sum_{n=1}^N z(\mu_f^{(n)} = \mu_f) \quad (5.14)$$

$$P(\phi) = \frac{1}{N} \sum_{n=1}^N z(\phi^{(n)} = \phi) \quad (5.15)$$

where the indicator function $z(\mathbf{Z})$ is 1 if $\mathbf{Z}$ is true and 0 otherwise.

Visualisations of the variation of mutual information along the frequency spectrum can be seen for both ILD and IPD at 4 different signal to noise ratios (SNRs) in Figures 5.3 and 5.4. As can be seen at an infinite SNR, the most important frequencies for determining source elevation begin around 7kHz for both ILD and IPD. However, when additional noise is added, these higher frequency cues are increasingly deteriorated as can be seen at SNRs of 30dB, 20dB, and 10dB. At which point, the best localization cues are below 8kHz for both ILD and IPD cues.

5.4 Conclusion

In this chapter, the broad topic of machine learning has been outlined in terms of its applications to audio engineering. The fundamentals of machine learning are described to provide relevant background literature for the experiments described in Chapters 6 and 7. This review gives an overview to the different types of machine learning algorithm and the necessary terminology required to understand them. A section on machine learning and its applications to audio engineering is included to provide insight to how this new technology is being applied to traditional audio engineering tasks. Most importantly, a detailed account of supervised learning techniques including neural networks and random forest algorithms is given to provide a necessary background for the experiment results presented in subsequent chapters. Fundamental to the Random Forest is the concept of information gain or mutual information which is described in terms of its ability to determine optimal features for learning algorithms. A visualisation is provided showing how mutual information for determining sound source elevation using interaural level difference and interaural phase difference is highest at frequencies above 7kHz. The addition of noise to the signal decreases the importance of these higher frequencies however.

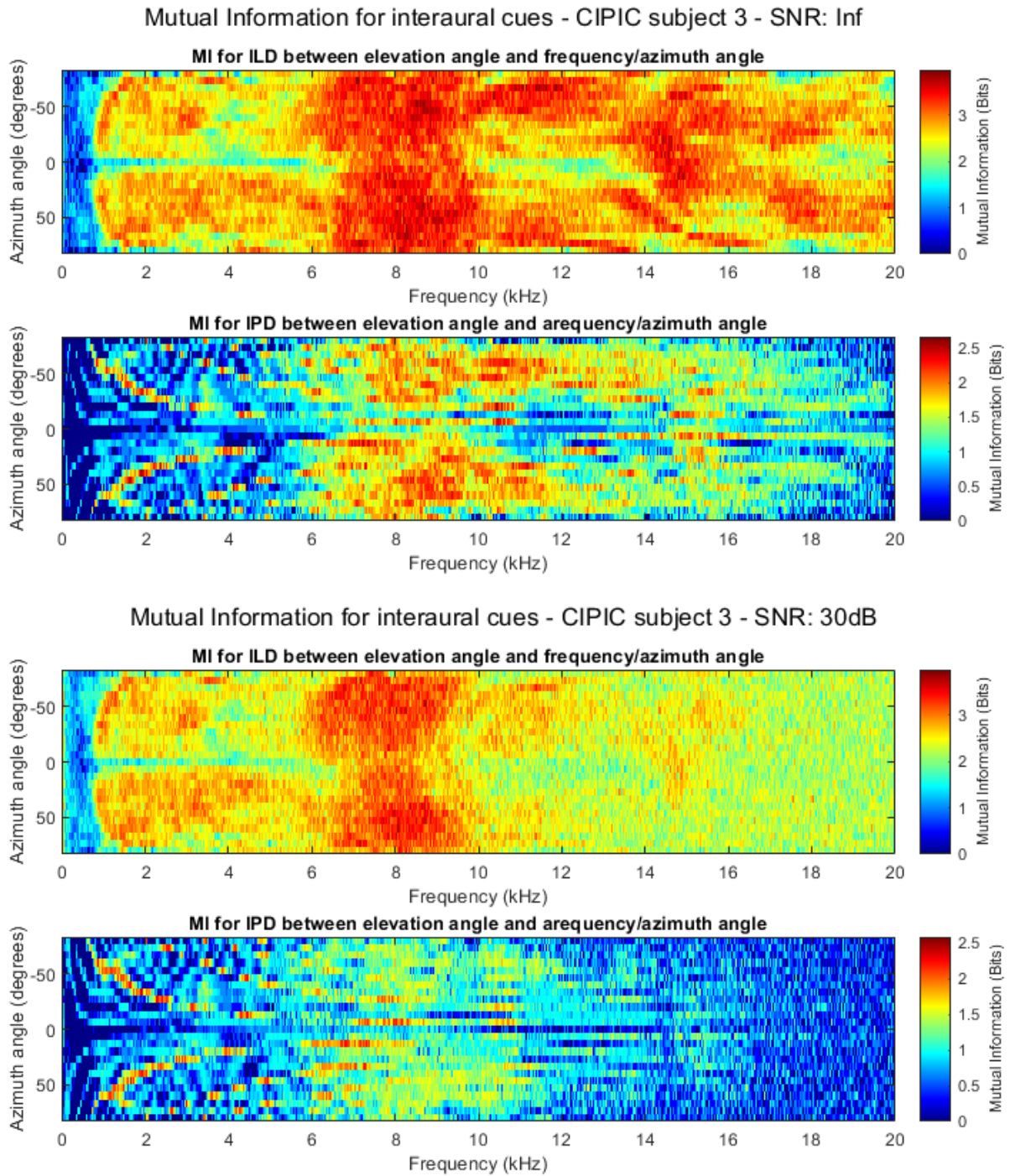


Fig. 5.3 The variation in mutual information for ILD and IPD up to 20kHz for binaural speech signals with SNR: Inf and SNR: 30dB.

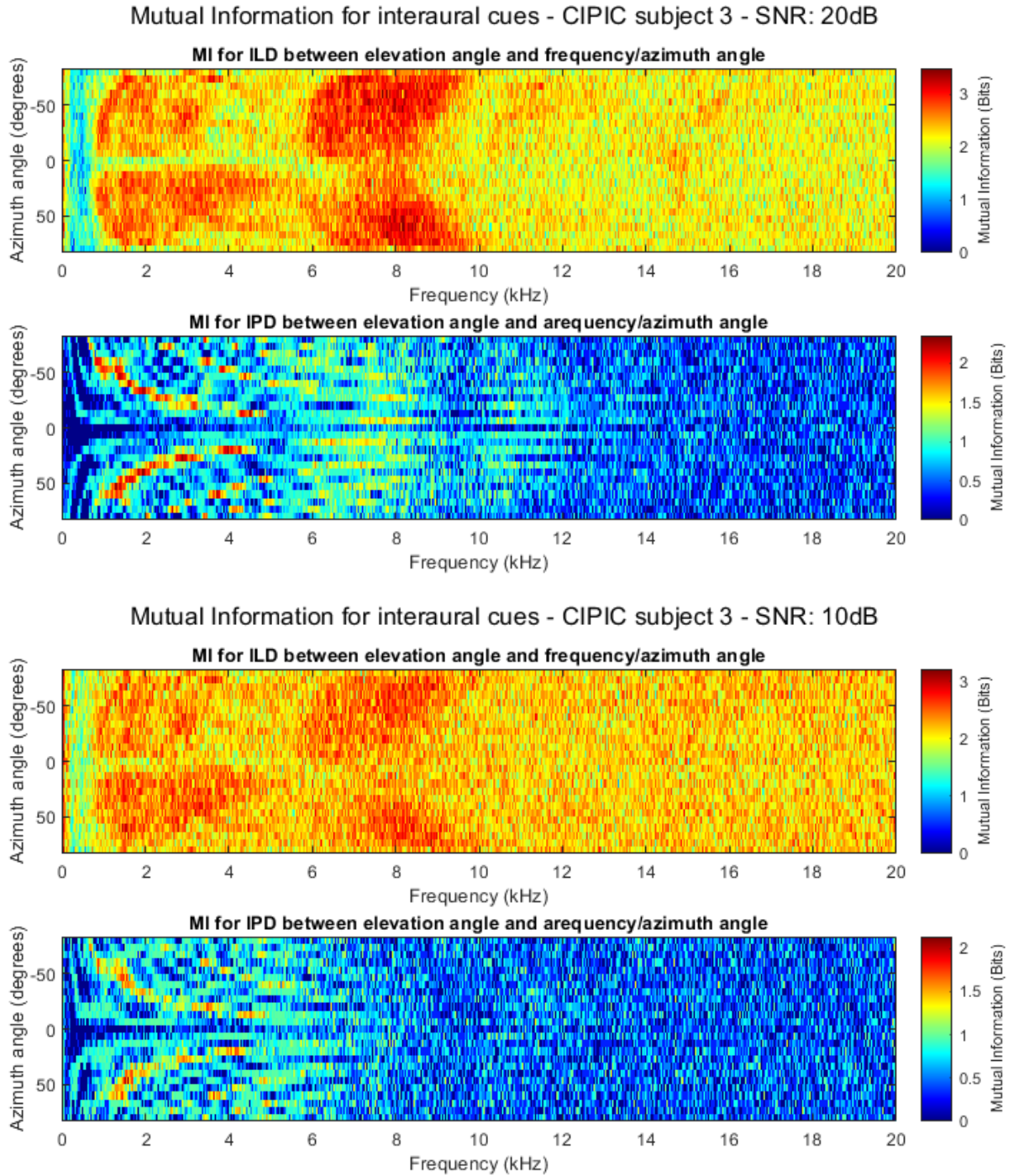


Fig. 5.4 The variation in mutual information for ILD and IPD up to 20kHz for binaural speech signals with SNR: 20dB and SNR: 10dB.

Chapter 6

Binaural Feature Selection for Sound Source Localization

Overview: This chapter presents a study of binaural features and how they can be used to learn the temporal and spectral cues of the HRTF using machine learning algorithms. Firstly, a variety of binaural signals are created by convolving binaural HRIR pairs with speech signals. Then, features of interest are extracted from the binaural signals and used to train deep neural networks and random forest algorithms to determine the source position of the HRIR used to create the binaural signal. Testing is performed to evaluate the performance of algorithms in both the matched and mismatched HRTF condition to determine the generalization of networks trained on each feature. While interaural cues such as ITD and ILD are strong predictors of source location along the horizontal plane, spectral cues are shown to perform better for source localization along the vertical plane. Using deep neural networks it can be seen that gammatone cepstral coefficients (GTCCs) outperform other spectral features including mel-frequency cepstral coefficients (MFCCs) and an estimation of the relative transfer function using cross-power spectral density. Recent studies in binaural localization have found that combinations of cues increase the performance of machine learning algorithms. A combination of Cross-Correlation Function (CCF), ILD, and GTCCs are presented in this chapter as a novel cue configuration for performing sound source localization of binaural sources. This configuration is shown to outperform two state of the art algorithms in the mismatched HRTF configuration for localization along vertical planes by up to 5%.

6.1 Introduction

In recent years, there has been a growing number of studies which use supervised-learning for audio source localization using neural networks. Many of these studies rely on multiple

microphone arrays and employ CNNs. While these networks demonstrate a high degree of success – there are many challenges which remain e.g. robustness to changes in microphone arrays and the acoustic environment, and for multiple sound sources. Sound source localization is a complex task which involves extracting information from the signals arriving at each ear. While this process has been studied widely, we do not fully understand how binaural cues are mapped to source location by the mammalian brain, particularly in relation to the vertical plane. We do know however that the cues which allow us to perform sound source localization are contained within the Head-Related Transfer Function (HRTF). However, we do not know how densely a HRTF must be sampled in terms of locations to sufficiently capture and represent an individual's listening. A sound source localization algorithm which uses just two receivers in the same way humans do would be of interest in the area of machine perception. For example it could be used for event detection and localization. Such an algorithm could also be used to assist the hearing impaired or within an augmented reality system [72, 87]. Other uses may include humanoid robotic systems and target tracking systems [94, 138, 212].

6.2 Binaural System Model

In order to train an ML algorithm, it is necessary to have a large amount of data to train, validate, and test the algorithm. Binaural signals can be easily created by convolving a two-channel HRIR or BRIR with a source signal allowing the synthesis of large databases of binaural audio to develop an algorithm. Synthesised binaural signals are modelled on a convolution of a source signal $s(t)$ with a pair of left and right HRIRs at a location $\Theta = (\theta, \phi)$. The HRIRs are thus notated as $h_l(t, \theta, \phi)$ and $h_r(t, \theta, \phi)$, and the signals received at the binaural receivers can be represented as:

$$\begin{aligned} x_l(t) &= s(t) * h_l(t, \theta, \phi) + n_l(t) \\ x_r(t) &= s(t) * h_r(t, \theta, \phi) + n_r(t) \end{aligned} \quad (6.1)$$

where n represents the additive noise term and t represents time. The received signal can then be interpreted in the frequency domain as:

$$\begin{aligned} X_{l,k}(f) &= H_l(f, \theta, \phi) \cdot S_k + N_{l,k}(f) \\ X_{r,k}(f) &= H_r(f, \theta, \phi) \cdot S_k + N_{r,k}(f) \end{aligned} \quad (6.2)$$

where k is equivalent to the frame number and f corresponds to the frequency. The binaural speech signals are separated into K frames, such that the frame length is less than the full duration of the signal. Typically in speech and audio signal processing a frame

size of between 10 and 50ms is used with an overlap of half of the frame length. In this study a frame length of 50 ms is used with an overlap of 25 ms. It has been suggested previously that a frame size of 50 ms be used for the extraction of interaural cues based on the temporal resolution of the auditory system [125]. When applying an FFT to a signal, the number of points must be specified which correspond to the number of frequencies at which the FFT is performed. These frequencies are evenly spaced from 0Hz up to the sampling rate of the audio which in this instance is 44.1kHz. Therefore approximately half of FFT points lie within the range of human hearing. With this in mind, the received signal spectra are represented by spectra vectors as:

$$\begin{aligned}\mathbf{x}_{l,k} &= [X_{l,k}(0), \dots, X_{l,k}(f_s/2)] \\ \mathbf{x}_{r,k} &= [X_{r,k}(0), \dots, X_{r,k}(f_s/2)]\end{aligned}\tag{6.3}$$

Using an FFT size of 2048, there are 1024 points in the signal spectra corresponding to different frequencies allowing us to easily examine the magnitude and phase values of the signals at each frequency.

6.2.1 Synthesising Binaural Signals

In this study, HRIRs from the CIPIC database [5] were convolved with speech signals from the TSP database [90]. The CIPIC database consists of HRTF measurements from 45 subjects including 2 HRTF sets measured on a KEMAR dummy manikin, one with large ears representative of the average male ear size and the other with small ears representing the average female ear. Each HRTF set in the CIPIC database contains measurements at 25 azimuthal positions ranging from -80° to $+80^\circ$ positions, i.e. from left to right. At each azimuthal position there are 50 HRTF measurements at different elevations ranging from -45° to 231° . These can be referred to as sagittal planes or the median plane in the instance where $\theta = 0$. Each HRTF measurement is represented as a 200 tap impulse response saved as a '.wav' file with a sampling rate of 44.1kHz.

The telecommunications and signal processing (TSP) laboratory at McGill University created the TSP speech database in 2002 [90]. It consists of 1400 utterances spoken by 24 speakers, half of which are male and the other half female. The database was recorded at a sampling rate of 48kHz in an anechoic room with the microphone placed 15cm and slightly off-axis from the speaker. This placement was intended to provide as high a signal-to-noise ratio as possible without excessive "pops" due to breath noises. Recording gain is consistent across all recordings. The reason why the TSP database was chosen in this study over the more popular TIMIT speech database [61] is because of its higher sampling rate in recording. This should insure that any high frequency content in the

speech recordings above 8kHz is maintained. As has been demonstrated earlier in this thesis, notch frequencies above 7kHz are important for vertical localization.

When synthesizing binaural speech signals, speech samples from the TSP database are first downsampled to 44.1kHz to match the sampling rate of the HRIRs from the CIPIC database. The samples are then trimmed to one second in length centered around the middle of the segment. Each sample is normalized to ensure a consistent volume. Finally, the binaural speech signals are created by convolving the HRIRs with the speech samples.

6.2.2 Binaural Cues and Their Extraction

Following the synthesis of binaural speech signals as described in the previous section, binaural features and cues can be extracted and used to train the DNN in predicting Sound Source Localization. In this section we describe these cues and how they can be extracted from binaural signals in the time and frequency domains.

Interaural Level Difference

Probably the most commonly used feature in related studies on binaural localization is the ILD, which reflects the shadowing of the head. ILD is a measurement of the difference in signal level between the left and right ears of a binaural signal and can be measured as a singular value over the entire frequency range or as an array of values measured at different frequencies. ILD is a result of high frequency attenuation caused by the head as sound waves travel around it, as a result, ILD values below 1.5kHz are relatively minimal.

In this study, ILD is calculated by first transforming both left and right signals into the frequency domain using a Fast-Fourier Transform (FFT) of length 2205. The resultant magnitude values are defined as E_l and E_r and ILD is calculated as follows:

$$ILD(\Theta, f) = 20 \log_{10} \left| \frac{E_l(\Theta, f)}{E_r(\Theta, f)} \right| \quad (6.4)$$

In this study, a single value ILD is used taken as the average ILD value for frequencies between 1.5kHz and 20kHz.

Interaural Time Difference

ITD is a measure of the difference in time of arrival (TOA) of signals arriving at each ear. As previously mentioned in Chapter 2, ITD can be defined using the Woodworth equation which treats the head as if it were a perfect sphere:

$$ITD(\theta) = \frac{r}{c} (\cos \theta + \theta) \quad (6.5)$$

where $0 \leq \theta \leq \frac{\pi}{2}$, c is the speed of sound (340.3 m/s) and r is the radius of the human head, approximated here as 8.75cm. This formula provides a peak ITD at 90° giving a value of approximately $661\mu\text{s}$. ITD is therefore only an important cue for localization below approximately 1500Hz as above this frequency, the wavelength of sound is less than the diameter of the human head. ITD can also be calculated as the lag time at which the peak of the cross-correlation function occurs. As can be seen in Figure 6.1, the peak of the CCF differs in lag time depending on a binaural signals position along the horizontal plane. At -80° , the peak is found at lag time of approximately $-590\mu\text{s}$ while at $+80^\circ$, the peak is found at lag time of approximately $+580\mu\text{s}$. These values are equivalent to the ITD and can be extracted easily by first finding the index position of the max value of the CCF and then finding the equivalent lag value at that index. This is the method used for ITD calculation in this study.

Cross-Correlation Function

While ITD is a reliable representation of the difference in the time of arrival of a sound at both ears, it has been shown by Ma et al. in [120] that the normalized cross-correlation function (CCF) is a superior cue as it contains more information related to the position of a sound source. The method used by Ma uses a vector consisting of both ILD and CCF to estimate the horizontal location of multiple speakers in reverberant conditions to a high degree of accuracy.

In this study, features are extracted by first dividing both left and right signals into frames of length 50ms with a 25ms overlap. This is done using a hamming window. The normalized CCF for each frame is then calculated as follows:

$$CCF(k, \tau) = \frac{\sum_m (x_{l,k}(m) - \bar{x}_{l,k})(x_{r,k}(m - \tau) - \bar{x}_{r,k})}{\sqrt{\sum_m (x_{l,k}(m) - \bar{x}_{l,k})^2} \sqrt{\sum_m (x_{r,k}(m - \tau) - \bar{x}_{r,k})^2}}, \quad (6.6)$$

where $x_{l,k}$ and $x_{r,k}$ refer to the left and right signals; k refers to the time frame index; m refers to the sample index; τ refers to the time lag; and $\bar{x}_{l,k}$ and $\bar{x}_{r,k}$ refer to the mean values in one frame. Considering the radius of the human head and the speed of sound, the range for τ is $[-1, 1]$ ms. As the sampling rate used in this study is 44.1kHz, the range of the sample lag is $[-45, 45]$; resulting in a dimensionality of the CCF feature being 91.

As can be seen in Figure 6.1, the CCF peaks at a different lag time depending on the source location.

Interaural Cross-Correlation Coefficient

The Interaural Cross-Correlation Coefficient (IACC) is a measure of similarity between the signals arriving at the left and right ears. It is taken as being the maximum value of the

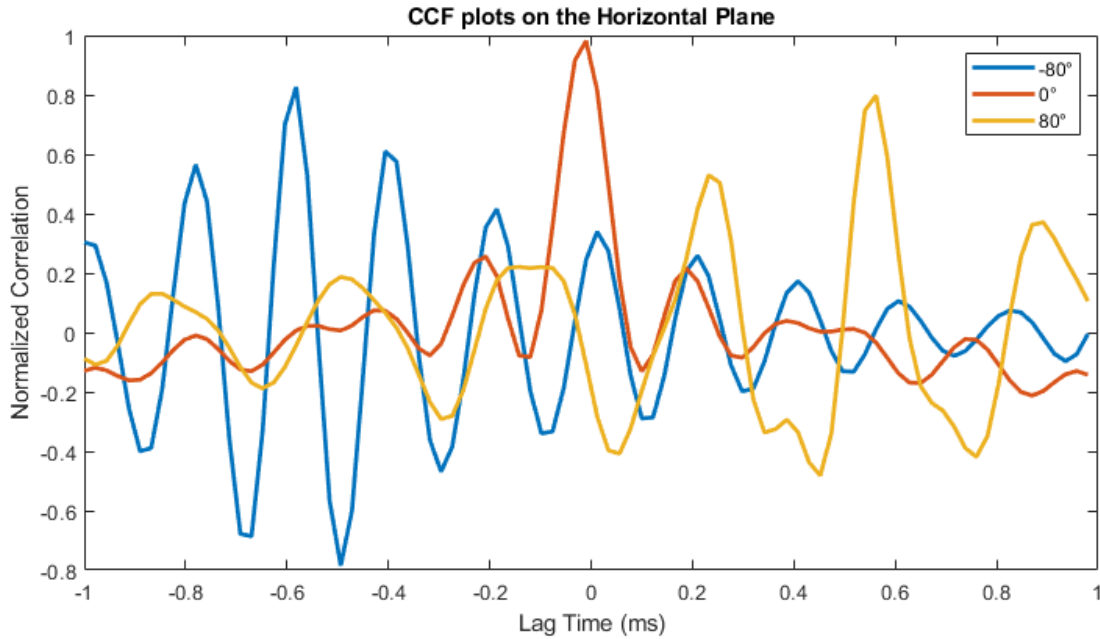


Fig. 6.1 Examples of CCF plots between binaural signals at $-80^\circ, 0^\circ, 80^\circ$ on the horizontal plane. These measurements were made using random speech signals from the TSP database convolved with pairs of HRIRs from CIPIC subject 165.

normalized CCF and can be either positive or negative lying in the range $[-1, 1]$. A value of 1 indicating that the signals are the same and -1 indicating that the same signal has arrived at each ear 180 degrees out of phase. An IACC value of 0 implies there is no similarity in the signals arriving at both ears.

$$IACC = \max |CCF(\tau)| \text{ where } |\tau| \leq 1 \text{ ms} \quad (6.7)$$

In horizontal stereophony, we know that Interchannel Correlation (ICC) relates to the horizontal spread of a phantom auditory image. When two correlated signals are presented between a stereophonic pair of left-right loudspeakers, a decrease of ICC increases the horizontal image spread (HIS) of the phantom auditory image [67, 216]. In 2018, Gribben and Lee showed that for vertically spaced loudspeakers, Vertical Image Spread (VIS) increased as ICC decreased for frequencies greater than 500 Hz [68]. This leads to the question, what does interaural cross correlation (IACC) between left and right signals tell us about the elevation of a sound source and vertical image spread.

In addition to this, Gribben and Lee showed the effect that ICC has on binaural signals. BRIR pairs were measured at two different heights and at 5 different azimuths. These BRIRs were then convolved with correlated and decorrelated signals. The difference in frequency response between the correlated and decorrelated signals was then found which shows an increase in various peak and notch frequencies for different azimuth positions.

Specifically, at 0° azimuth, vertical decorrelation results in a boost of 8-9kHz. For 30° azimuth, boosts can be seen at 7-8kHz and 16kHz in the contralateral ear and 8kHz and 12kHz in the ipsilateral ear. At 110° azimuth, decorrelation seems to generally result in a boost of high frequencies in the contralateral ear, while the ipsilateral ear features two boosts around 8-9kHz and 14-15kHz. The study goes on to say that a reduction of notch depth may decrease localization accuracy and as such increase the perceived VIS.

Apparent Source Width

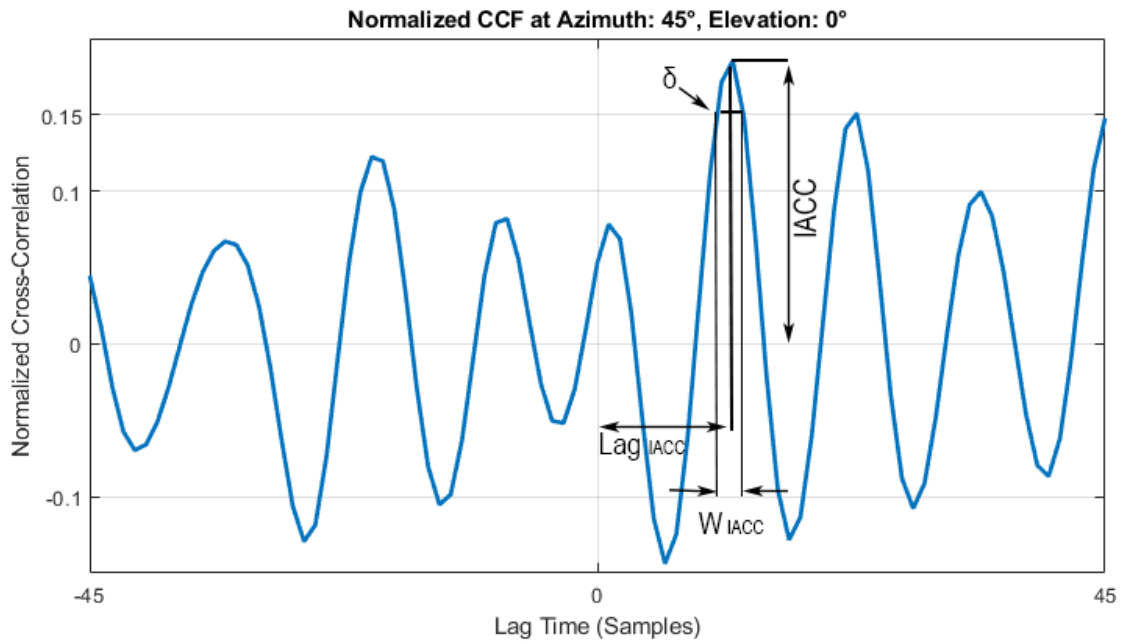


Fig. 6.2 Finding W_{IACC} , an approximation for ASW. Reproduced after Sato [167].

Apparent Source Width (ASW) is a term used to describe the acoustical width of a sound source as perceived by the listener and is a defining attribute of a sound field such as a concert hall. In 2002, a study by Sato & Ando demonstrated that ASW could be approximated effectively as the width of the interaural cross-correlation function, W_{IACC} , which is defined as the interval of delay time at a value δ , 10% below the IACC [167]. The W_{IACC} of two signals can be calculated using the CCF by measuring the width of the peak at 90% of the value of the IACC. In Figure 6.2, we can see how the width of the interaural cross-correlation function is found and thus used to approximate ASW. In this study we will use the W_{IACC} as an approximation of ASW. From this point onward this cue will simply be referred to as ASW.

Cross Power Spectral Density and the Relative Transfer Function

The power spectrum of a signal in the time domain describes the distribution of power into frequency components that make up that signal. The power spectrum is computed over a finite amount of time and can be used to determine information about an audio signal, for example the pitch and timbre of a piece of music. Cross-power spectral density (CPSD) is the cross-correlation between the power spectrum of two signals, showing the power shared by the two signals at a specific frequency.

In [146], Opoichinsky et al. showed the relative transfer function (RTF) of two signals could be a cue used to determine sound source azimuth in two microphone localization tasks. The RTF used in this instance was based on the CPSD of signals from these microphones. As the directional information in a binaural signal is independent of the actual source signal, we are mostly interested in capturing the information contained in the HRTF. The RTF is a representation of the spectral information of one binaural signal relative to the other and can be calculated as follows:

$$RTF_{\Theta}(f) = \frac{\hat{S}_{x_{l_{\Theta}}x_{r_{\Theta}}}(f)}{\hat{S}_{x_{l_{\Theta}}x_{l_{\Theta}}}(f)} \quad (6.8)$$

where $\hat{S}_{x_{l_{\Theta}}x_{r_{\Theta}}}$ is the cross-power spectral density of the left and right binaural signals; $\hat{S}_{x_{l_{\Theta}}x_{l_{\Theta}}}$ is the auto power spectral density (APSD) of the left binaural signal. Using an FFT size of 512 results in the calculation of 257 point RTF.

Cepstral Coefficients - MFCCs and GTCCs

Cepstral Coefficients are common features used in computational auditory scene analysis (CASA). In this thesis we will examine the use of both Mel-Frequency Cepstral Coefficients (MFCCs) and Gammatone Filter Cepstral Coefficients (GTCCs). While both have proven to be invaluable cues in tasks such as speech recognition [213], their effectiveness in SSL tasks has yet to be fully determined.

While there are several variations in the methods used to calculate MFCCs and GTCCs from an audio signal, this thesis will use the methods implemented in MATLABs ‘mfcc’ and ‘gtcc’ functions respectively. The calculation of MFCCs can be described using the following process:

- The input signal is pre-emphazied using a high-pass filter.
- A Short-Time Fourier Analysis is performed on the signal using a hamming window with a window size of 50ms and a 25ms overlap. When using a sampling rate of 44.1kHz, this equates to a window size of 2205 samples while the overlap size is rounded up to 1103 samples.

- The magnitude spectrum is then wrapped into mel-spectrum bands using triangular overlapping windows whose center frequencies are equally distributed on the mel scale.
- The sum of these mel filtered signals are then gathered and the log energy of each filterbank is calculated.
- The Discrete Cosine transform (DCT) is applied to these log energies to derive the cepstral features. This is nonlinear rectification.
- Cepstral Liftering can also be applied in the calculation of MFCCs whereby higher dimensional coefficients are amplified to balance the magnitudes and variances across dimensions.

GTCCs are gathered in the following way when the Filter Domain is set to ‘Time’ in MATLAB :

- The signal is passed through an N-channel gammatone filterbank
- For each channel in the filterbank, the response is then fully rectified, i.e. its absolute value is taken and the signal is split into n-frames according to window size and overlap.
- This creates a time-frequency T-F representation that is a variant of a cochleagram.
- Nonlinear rectification is then calculated by finding the cubic root of the sum of each T-F representation.
- The DCT is then applied to these values to derive the cepstral features.

In this study, 32 MFCC and GTCC values are used for each ear measured at center frequencies from 0-20kHz. Additionally the log energy of each signal is also included with each MFCC and GTCC measurement resulting in a total of 33 values. As these are measured for both ears, the final vector length for both GTCC and MFCC cues is 66. The central frequencies for both MFCCs and GTCCs can be seen in Figure 6.3.

6.3 Single Feature Performance for SSL Tasks

In this section we present the results of DNNs trained on the binaural features discussed in the previous section. In each instance, performance along a particular plane is examined. The entire CIPIC HRTF measurement consists of 1250 individual HRTF measurements. These consist of 50 vertical measurements ranging from -45° to 230.625° measured at 25

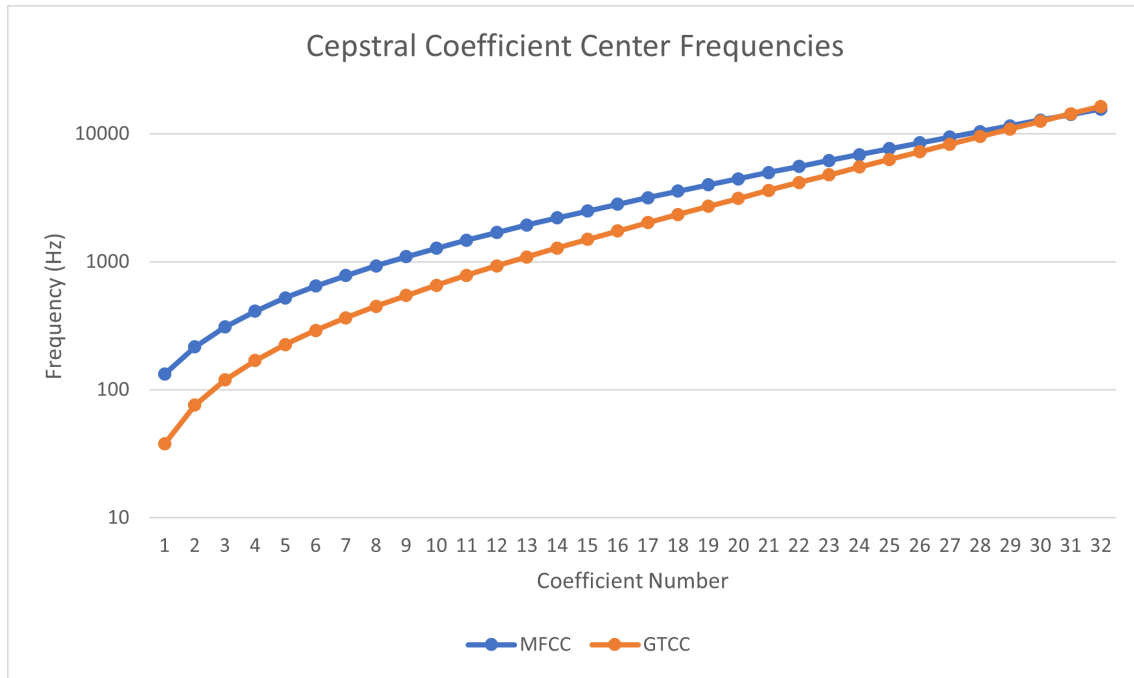


Fig. 6.3 The center frequencies used to calculate MFCC and GTCC values.

different sagittal planes from -80° to $+80^\circ$. Each vertical measurement is evenly spaced 5.625° apart while the horizontal measurements are spaced 5° apart from -45° to $+45^\circ$ with subsequent measurements at $\pm 55^\circ$, $\pm 65^\circ$, and $\pm 80^\circ$. For this study, only the frontal hemisphere is considered and therefore only vertical HRTF sets from -45° to 90° are used. This conveniently amounts to 25 vertical HRTF sets meaning that sound source localization performance is examined around 25 vertical and 25 horizontal HRTF locations. Each HRIR pair is convolved with 100 randomly selected speech samples from the TSP database resulting in a total dataset of 2500 binaural speech samples. As each feature varies in length, the size of the input layer and subsequent hidden layers in the network vary, the number of nodes in each hidden layer is given alongside the performance of the specific feature.

Of particular interest in this study is the localization performance for both matched and mismatched HRTF sets. Each of the 45 HRTF sets of the CIPIC database are used to train an independent network which is then tested on all 45 HRTF sets. This allows for the examination of the performance of each trained network for 1 matched and 44 mismatched HRTF conditions which gives a more accurate view of the performance of each feature set. While in general, HRTF sets will perform best in the matched condition, a variation in the results of the mismatched condition allows us to compare the performance of different HRTF sets and examine an overall performance which can be measured by two means, the percentage of test samples correctly classified to within a particular threshold and the mean absolute error (MAE) of all the test samples.

6.3.1 Method

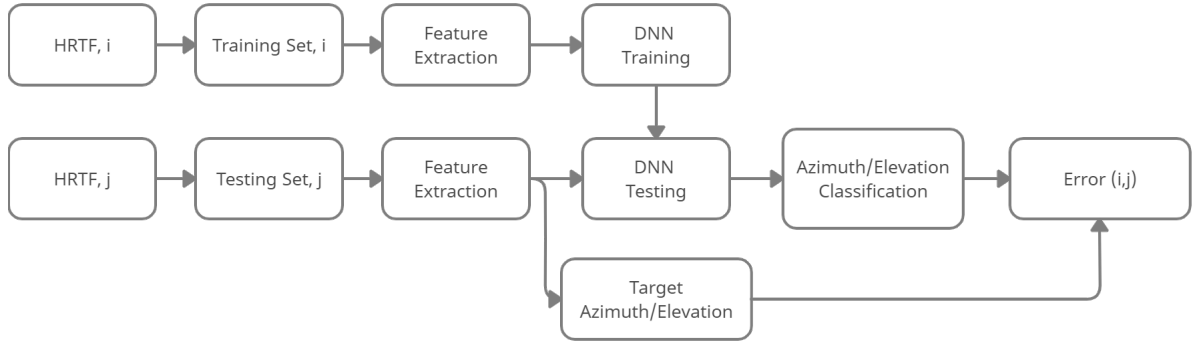


Fig. 6.4 A flowchart of the DNN based localization method.

The method used in this study can be seen in Fig 6.4. Firstly, a database of binaural speech signals is created using HRTF sets from each individual of the CIPIC database and random speech signals from the TSP database. For each HRTF, 100 speech samples are used and an average value for each cue across all frames is taken. These frames are of length 50ms with a 25ms overlap and are created using a Hamming window. From these frames, the selected features discussed in the previous section are extracted. Namely these include ITD, ILD, IACC, ASW, CCF, MFCCs, GTCCs, & RTF. The creation of binaural signals and the subsequent extraction of features for training is performed using a front-end processor in MATLAB.

Test samples are then saved to a .csv file for DNN training which is performed in python using Keras [36] as a high-level API to run a Tensorflow backend [1]. The corresponding azimuth or elevation class for each sample is converted to a one-hot encoded vector, a method used for DNN classification tasks. These vectors are then used as the output features which the network is trained on. Before training the DNNs, each binaural input feature is normalized to have zero mean and unit variance.

The network itself is customizable with layer sizes varying depending on the feature used to train it. Its architecture can be seen in Figure 6.5. The input layer is equal to the length of the input feature while the output layer is equal to 25, the number of classes in this classification task. There are three hidden layers consisting of a variable number of nodes which use a rectified linear unit (ReLU) activation function. ReLU activation is a piecewise linear function that will output the input directly if it's positive or will output 0 if it's negative. Following each hidden layer is a Dropout layer where 20% of the nodes are randomly removed to avoid overfitting. The DNN training algorithm uses the 'Adam' optimizer [95], with a learning rate of $10e^{-4}$. The output layer uses 'softmax' activation to select an output class. The network is compiled using 'categorical_crossentropy' as its loss

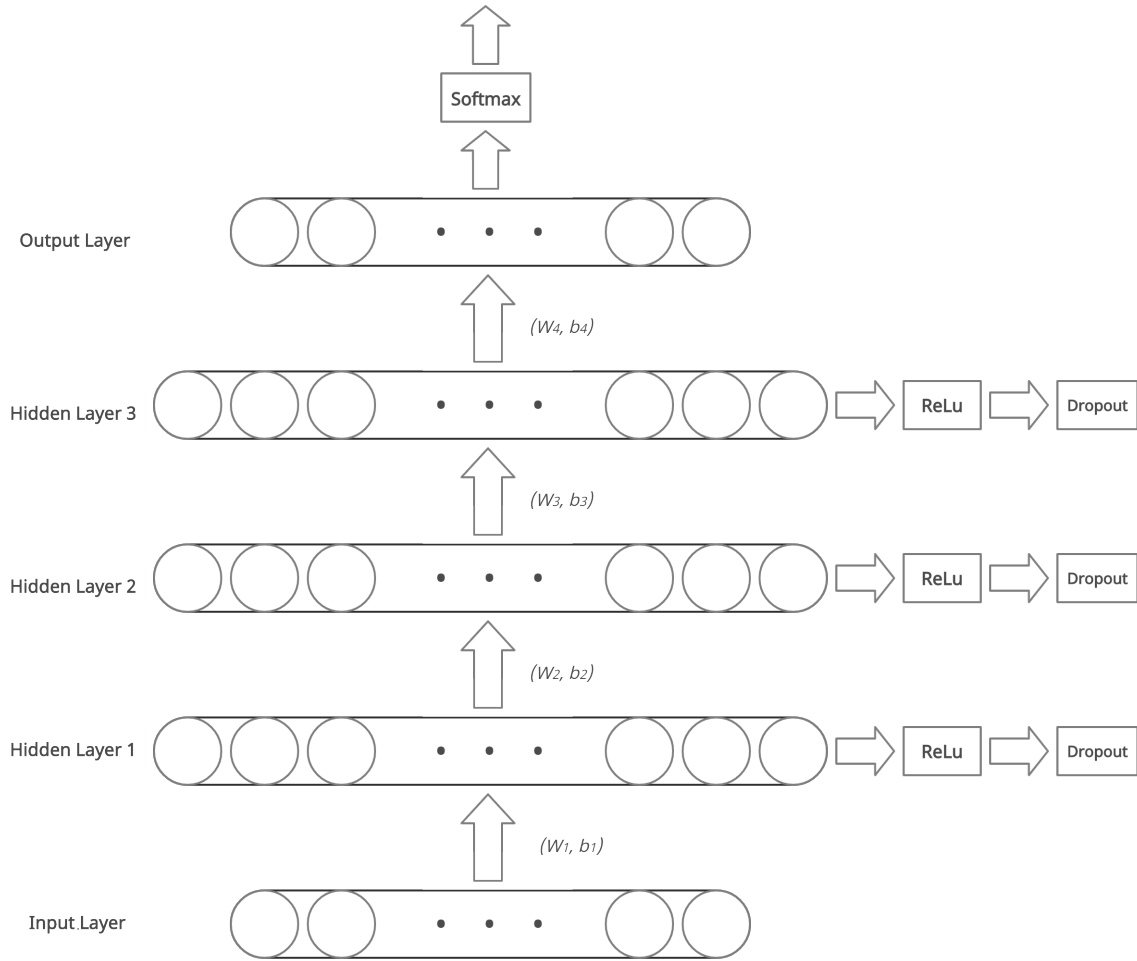


Fig. 6.5 The architecture of the DNN used to perform sound source localization using single feature inputs.

function. The network is trained for a total of 300 epochs with early stopping employed if there is no change in model performance for 10 epochs. Mini-batch gradient descent is used with batch sizes of 200.

Feature performance is measured under three criteria; the percentage of test samples correctly classified, the percentage of test samples classified to within 5° in the case of horizontal localization and 6° in the case of vertical localization, and the mean absolute error (MAE) of the feature performance. Each feature is trained on 45 separate networks, one for each subject in the CIPIC database. Each of these 45 networks is tested using 45 sets of testing data, one set for each subject. The performance is thus taken as the average performance over all networks and testing conditions, i.e. 45x45 tests. The weights and biases of each layer are reset before re-training the network on a new subject.

6.3.2 Feature Performance on the Horizontal Plane

In this section, the performance of features when estimating sound source location in the horizontal plane is examined. Features are trained to estimate the azimuthal direction of the binaural signal using classification to one of the 25 HRTF locations around the horizontal plane from -80° to $+80^\circ$. The DNN uses classification to estimate the azimuth of the signal and an estimation is only considered correct if the right class of angle is given by the classifier. In Table 6.1, the input layer length and hidden layer length for each feature can be seen along with the features performance in estimating source azimuth along the horizontal plane. Feature performance is examined under three headings; the percentage of samples correctly classified, i.e. Error = 0° ; the percentage of samples correctly classified to within 5° , i.e. Error = $\leq 5^\circ$; and finally the mean absolute error (MAE). As this task uses classification, an error of $\leq 5^\circ$ is equivalent to a classification to one class in either direction between -45° to $+45^\circ$. Beyond this, HRTF measurements are separated by at least 10° and are not considered as correct classifications. These feature performance values are taken as an average over all 45x45 tests.

Table 6.1 Performance of Individual Features for Localization on the Horizontal Plane

Feature	Input Layer Length	Hidden Layer Length	Error = 0°	Error $\leq 5^\circ$	MAE
ITD	1	50	45.62%	76.32%	8.3°
ILD	1	50	33.14%	63.5%	11.14°
IACC	1	50	7.55%	18.46%	52.98°
ASW	1	50	6.4%	15.22%	56.97°
CCF	91	200	45.24%	77.74%	13.3°
MFCCs	66	200	13.75%	25.28%	35.43°
GTCCs	66	200	18.82%	34.95%	24.8°
RTF	257	300	27.46%	46.1%	25.6°

In Figure 6.6, the performance of each feature for all test conditions can be visualized as separate images presented as performance matrices. These are the percentage of test samples correctly classified with an error of 0° . Note that the scale of the accuracy attained for each feature is different and is defined in the colorbar directly to the right of each image. In Figure 6.7, the accuracy of each feature in both the matched and mismatched HRTF conditions is shown as bar plots with the standard deviation for each also displayed. Note that these results are an average of 45 tests in the matched condition and 1980 tests in the mismatched condition.

When examining localization on the horizontal plane; the best performance was found when using temporal cues, ITD and CCF. These results are in agreement with those found

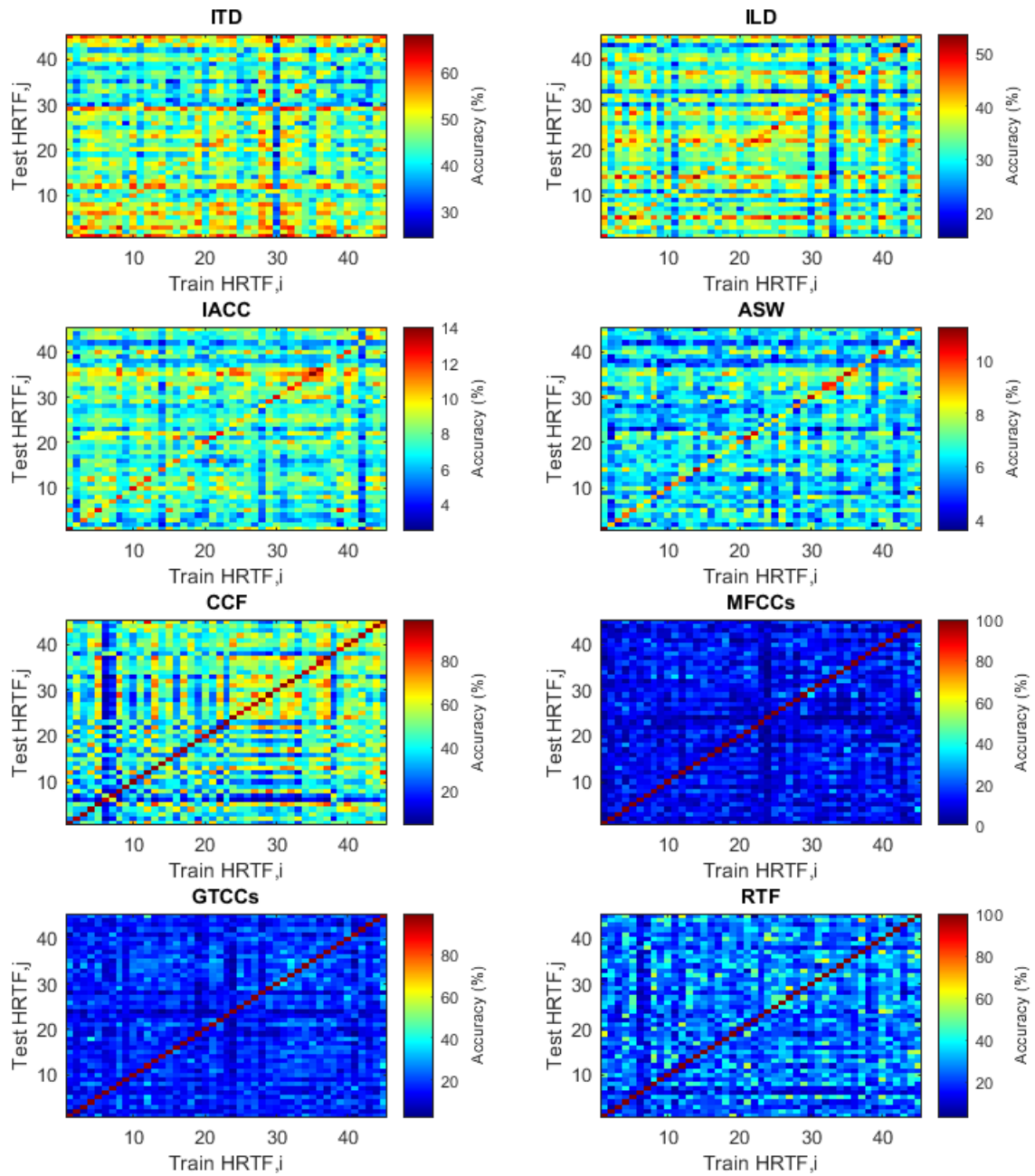


Fig. 6.6 The performance of each feature along the Horizontal Plane, correctly classified. A single training HRTF is used to train the DNN and subsequently tested on each of the 45 HRTFs from the CIPIC dataset.

in many other studies including [120]. Temporal cues also performed best in terms of generalisation with ITD and CCF achieving results of 76.28% and 77.26% when estimating

to within 5° in the mismatched condition respectively. While ILD performed better than most features in this task, its results were poorer than both ITD and CCF. CCF proved to be the best single feature in this task performing best in both the matched and mismatched condition. Neither IACC or ASW performed well while spectral cues MFCCs, GTCCs, and RTF performed poorly in terms of generalisation. All three of these cues performed very well in the matched HRTF condition and comparatively poorly in the mismatched HRTF condition showing signs of overfitting. This indicates that these spectral cues are highly individualised. GTCCs outperformed MFCCs slightly indicating that the ERB scaling of frequencies is more proficient in horizontal localization tasks than the Mel scale.

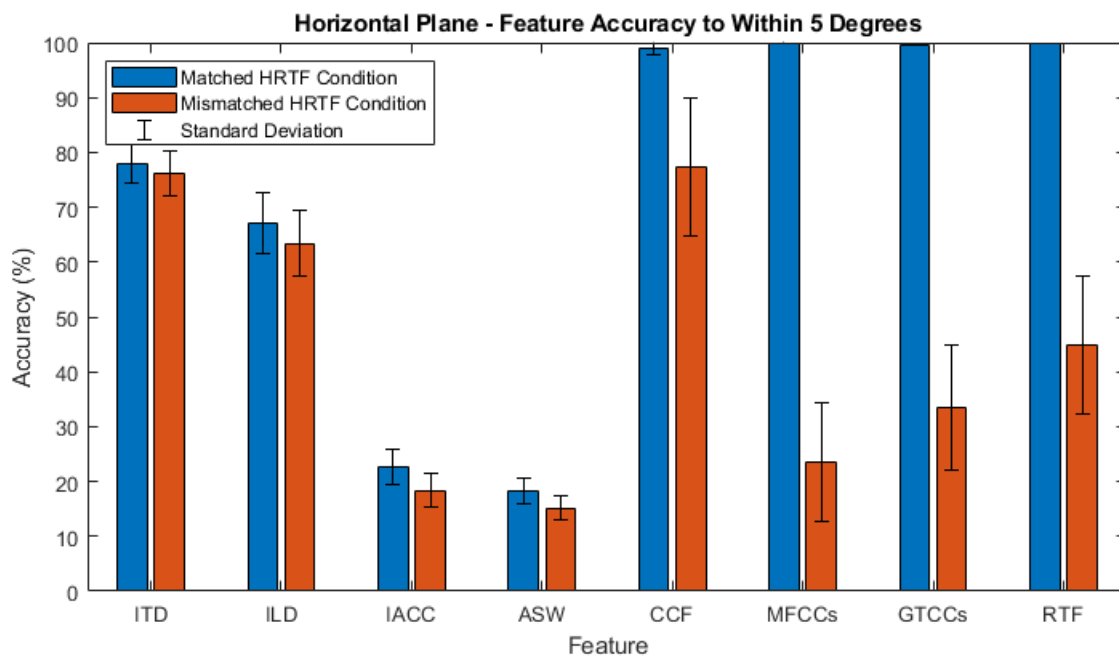


Fig. 6.7 Horizontal plane feature performance to within 5° . Accuracy is shown as the average performance in both the matched and mismatched HRTF condition.

6.3.3 Feature Performance on the Median Plane

In this section, features are trained to estimate the vertical direction of the binaural signal. While the entire HRTF in the CIPIC database consists of 50 vertical HRTF measurements, here the study is confined to the first 25 HRTFs ranging from -45° to $+90^\circ$. These measurements are uniformly spaced 5.625° apart. Consequently we are concentrating only on the frontal hemisphere and ignoring sound source localization from the rear of the listener. This will reduce error associated with front-back confusion. The DNN uses classification to estimate the elevation of the signal and an estimation is only considered correct if the right class of angle is given by the classifier. In Table 6.2, the input layer length and hidden layer length for each feature can be seen along with the features performance in estimating source elevation along the median plane. As in the case of horizontal plane localization, performance is measured under three criteria; the percentage of samples correctly classified, i.e. Error = 0° ; the percentage of samples correctly classified to within 6° , i.e. Error = $\leq 6^\circ$; and finally the mean absolute error (MAE). As this task uses classification, an error of $\leq 6^\circ$ is equivalent to a classification to one class in either direction between -45° to $+90^\circ$. As shown by Blauert in [20], this is less than the just noticeable difference for elevation which at its lowest is $\pm 9^\circ$ at 0° . These feature performance values are taken as an average over all 45x45 tests.

Table 6.2 Performance of Individual Features at Vertical Localization on the Median Plane.

Feature	Input Layer Length	Hidden Layer Length	Error = 0°	Error $\leq 6^\circ$	MAE
ITD	1	50	4.42%	12.26%	59.75°
ILD	1	50	4.65%	13.06%	58.5°
IACC	1	50	4.32%	12.25%	59.7°
ASW	1	50	4.3%	12.15%	59.52°
CCF	91	200	6.28%	14.92%	56.54°
MFCCs	66	200	8.82%	21%	47.06°
GTCCs	66	200	9.72%	23.31%	45.5°
RTF	257	300	7.02%	15.96%	55.72°

In Figure 6.8, the performance of each feature for all test conditions can be visualized as separate performance matrices. Note that the scale of the accuracy attained for each feature is different and is defined in the colorbar directly to the right of each image. In Figure 6.9, the accuracy of each feature in both the matched and mismatched HRTF conditions is shown as bar plots with the standard deviation for each also displayed. Accuracy is determined as the percentage of samples classified to within 6° of the correct source position. This is equivalent to approximately one class in either direction in terms of

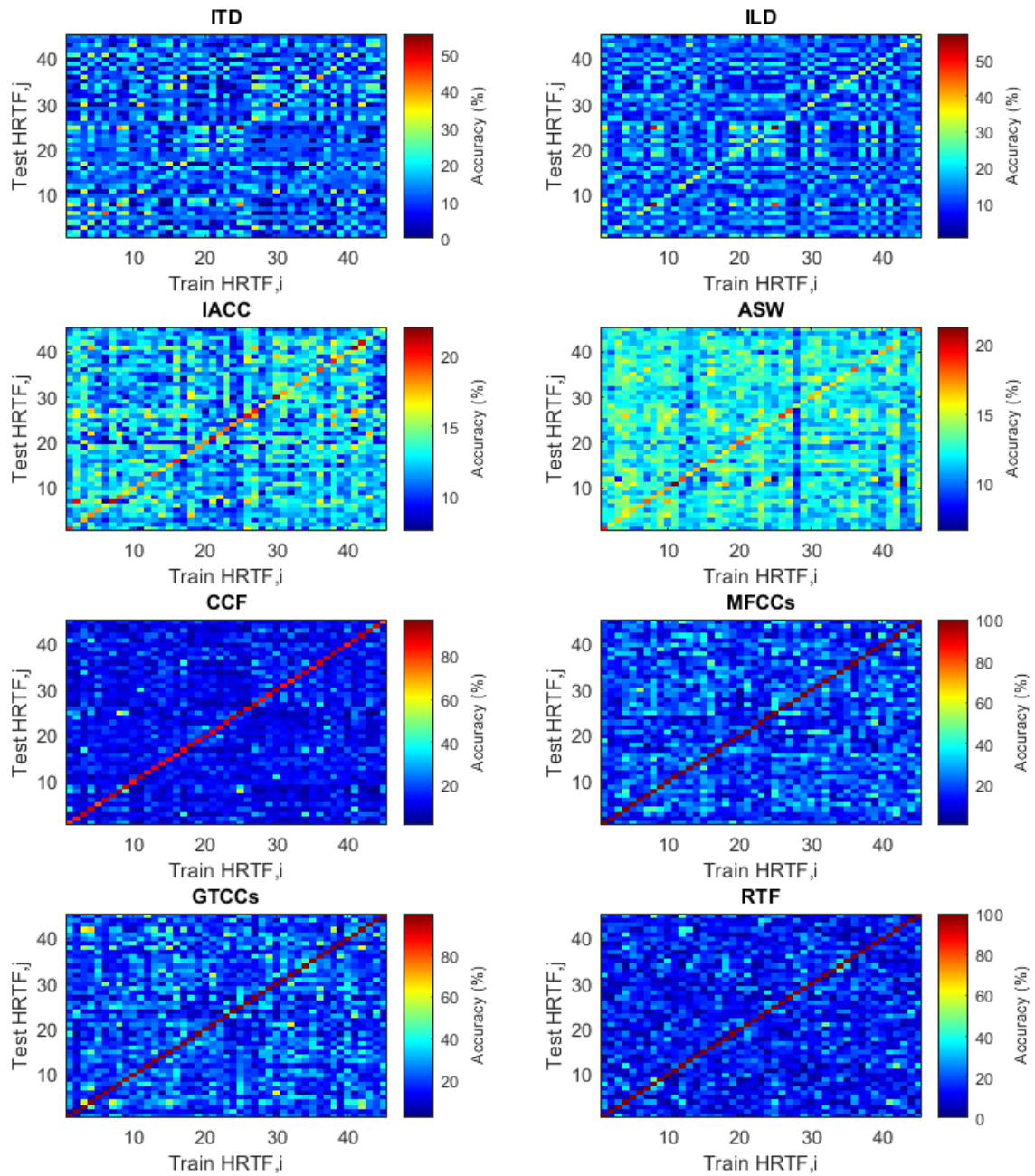


Fig. 6.8 The performance of each feature along the Median Plane, correctly classified to within 6° . A single training HRTF is used to train the DNN and subsequently tested on each of the 45 HRTFs from the CIPIC dataset.

the HRTF measurement positions along the vertical plane. Note that these results are an average of 45 tests in the matched condition and 1980 tests in the mismatched condition.

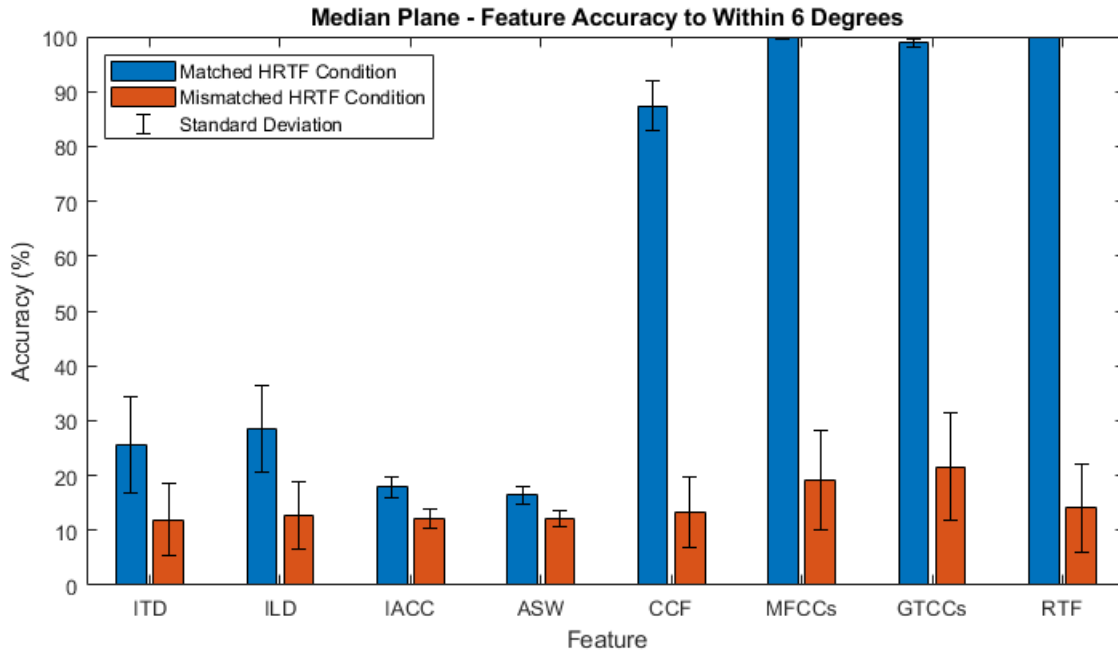


Fig. 6.9 Median plane feature performance for estimating elevation to within 6°. Accuracy is shown as the average performance in both the matched and mismatched HRTF condition.

As shown in Fig 2.9, localization blur for sound events along the median plane is mostly $> 10^\circ$, particularly in the rear hemisphere. With this in mind it should be noted that an error of $\leq 6^\circ$ is more accurate than an average human could localize sound along the median plane. With this in mind, an error rate relative to 2 classes, i.e. to within $\pm 11.25^\circ$ would be more representative of human hearing. However, because the algorithm being trained uses classification and not regression that would give the algorithm a 5 out of 25 chance of simply guessing a class within our target range. Such a large range of accepted classes could lead to a testing performance which is misrepresentative of actual feature performance.

In terms of localization along the median plane, no features perform well in the mismatched HRTF condition despite a strong performance of CCF, RTF, and cepstral features in the matched condition. Interaural cues as expected perform very poorly under all performance criteria and it's clear from the results shown here that they are unsuitable features for elevation estimation. Performance by the CCF, RTF, and cepstral features in the matched condition highlights the individuality of the HRTF in terms of vertical localisation. It's evident that there is information in these cues that can be used to accurately predict the vertical location of a sound source in binaural audio but that information is not transferable to other HRTFs. GTCCs performed best in terms of the mismatched test condition with an accuracy of 21.6%.

Table 6.3 Performance of Individual Features for Localization on the Sagittal Plane at 45°

Performance of Individual Features for Localization on the Sagittal Plane at 45°					
Feature	Input Layer Length	Hidden Layer Length	Error = 0°	Error ≤ 6°	MAE
ITD	1	50	4.89%	13.09%	63.05°
ILD	1	50	4.85%	13.76%	57.8°
IACC	1	50	4.65%	12.82%	60.97°
ASW	1	50	4.6%	12.81%	59.43°
CCF	91	200	8.11%	20.14%	48.74°
MFCCs	66	200	9.13%	21.6%	48.37°
GTCCs	66	200	10.53%	25.49%	45.13°
RTF	257	300	8.32%	19.21%	51.08°

6.3.4 Feature Performance on the Sagittal Plane at 45°

In addition to examining feature performance on the Median plane at 0° along the horizontal, in this section the performance of features in estimating source elevation along the sagittal plane at 45° is examined. It is understood that interaural cues have a smaller variance along the median plane than they do at sagittal planes moving towards the side of the head as was shown in Figures 5.3 and 5.4. Along this plane, there is increase in interaural cues ILD and ITD as well as a shift in the CCF towards a lag time around 0.3ms. This also results in a decrease in the IACC and ASW while the cepstral cues, MFCCs and GTCCs see a shift in energies towards the right ear coefficients. These changes in cues relative to the median plane may be more or less pronounced at different elevations resulting from individual head and pinna geometry and as such it may be easier to estimate source elevation along this plane using these particular cues.

As in the case of median plane localization, performance is measured under three criteria; the percentage of samples correctly classified, i.e. Error = 0°; the percentage of samples correctly classified to within 6°, i.e. Error = ≤ 6°; and the mean absolute error (MAE). As this task uses classification, an error of ≤ 6° is equivalent to a classification to one class in either direction between -45° to +90°. These feature performance values are taken as an average over all 45x45 tests.

In Figure 6.10, the performance of each feature for all test conditions can be visualized as separate performance matrices. Note that the scale of the accuracy attained for each feature is different and is defined in the colorbar directly to the right of each image. In Figure 6.11, the accuracy of each feature in both the matched and mismatched HRTF conditions is shown as bar plots with the standard deviation for each also displayed. Accuracy is determined as the percentage of samples classified to within 6° of the correct source position. This is equivalent to approximately one class in either direction in terms

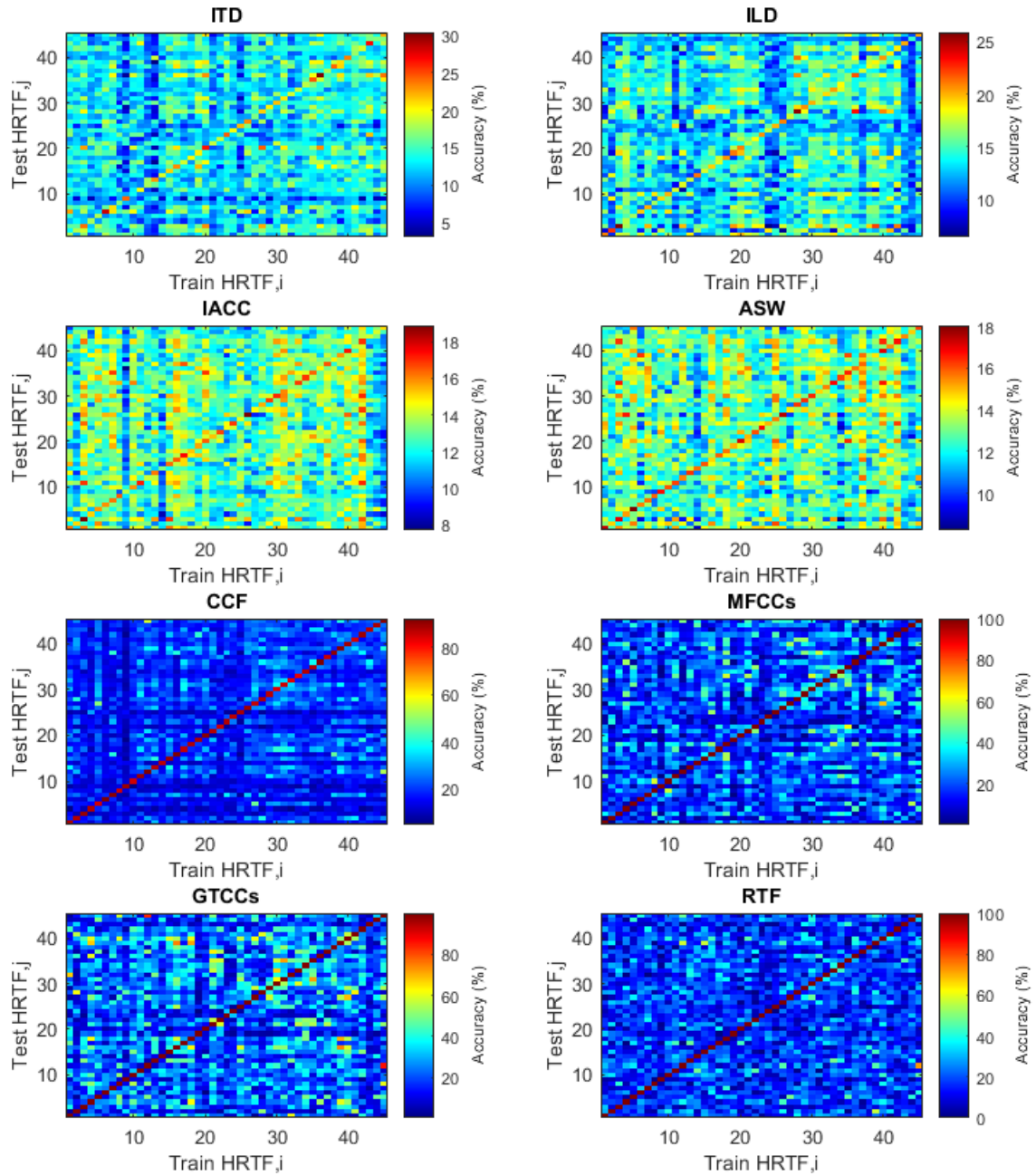


Fig. 6.10 The performance of each feature along the Sagittal Plane at 45°, correctly classified to within 6°. A single training HRTF is used to train the DNN and subsequently tested on each of the 45 HRTF sets from the CIPIC dataset.

of the HRTF positions along the plane. Note that these results are an average of 45 tests in the matched condition and 1980 tests in the mismatched condition.

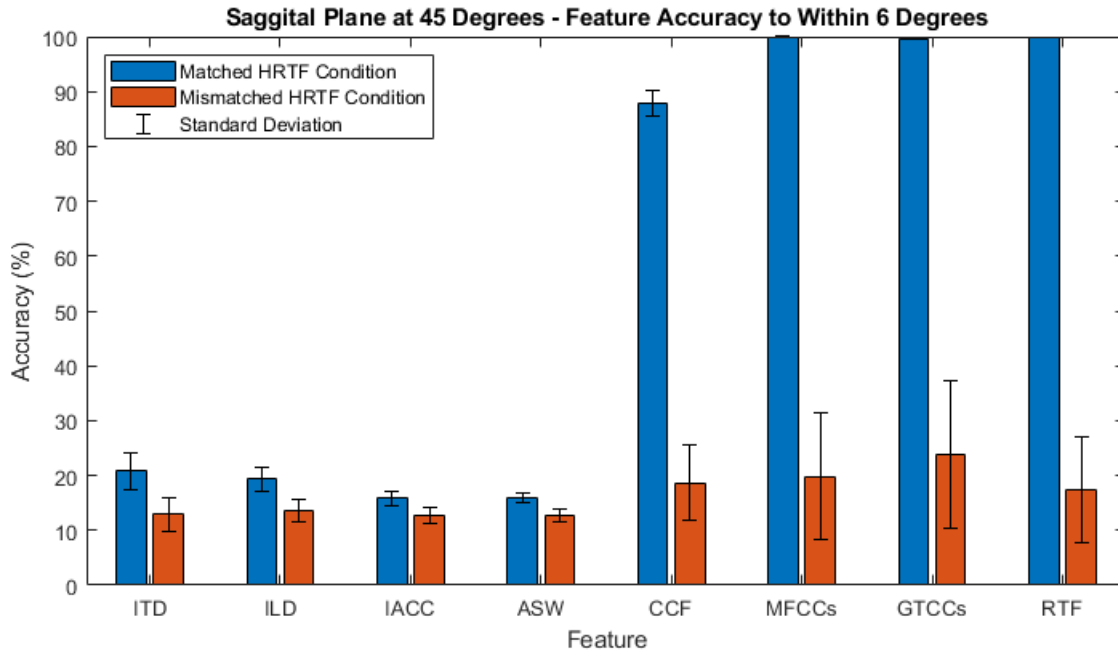


Fig. 6.11 Sagittal plane at 45° feature performance for estimating elevation to within 6°. Accuracy is shown as the average performance in both the matched and mismatched HRTF condition.

When moving away from the median plane and performing the same testing on the sagittal plane at 45°. Performance decreases slightly in the interaural cues in both the matched and mismatched HRTF condition. There is also a slight but noticeable increase in performance for CCF, RTF and cepstral coefficients in the mismatched HRTF condition. This would indicate that vertical localization is can be performed better away from the median plane which is to be expected.

In terms of vertical localisation, of particular interest from these results is the performance of GTCCs which were capable of finding similarity's between HRTFs that are known to be similar. This is the case of HRTFs 12 and 45 of the 45 HRTFs tested which were both measured using a KEMAR dummy head microphone but with different ears. On the sagittal plane at 45°, GTCCs found an accuracy of 86.3% when training on HRTF 45 and testing on HRTF 12 and 83.6% when training on HRTF 12 and testing on HRTF 45.

6.4 Multi-Feature Performance for SSL Tasks

Most ML algorithms for SSL tasks are trained using a combination of features. This is true in the case of deep neural networks trained by Ma et al. to perform azimuth localization using a combination of CCF and ILD features [120]. Vecchiotti et al. used raw left and right audio signals to predict sound source azimuth using convolutional neural networks

[188]. In 2019, Wu and Talagala showed that random forests could be trained to estimate both azimuth and elevation using a combination of ILD and IPD features [204]. While all of these studies presented promising results in both noisy and reverberant environments, their performance in the mismatched HRTF scenario was not examined.

In this section, an examination of multi-feature training algorithms is presented in both the matched and mismatched condition. A new feature combination of CCF, ILD, and GTCCs is introduced and its performance is evaluated alongside other current methods. As in the previous section, training and testing is performed on a single plane to examine the performance in estimating both sound source azimuth and elevation.

6.4.1 DNN training using CCF and ILD, the Ma Method

Interaural cues, ITD and ILD are the most commonly used cues in binaural source localization models. However, Ma et al. found that the normalized cross-correlation function contains more information relating to source localization than the ITD [120]. This is consistent with the results found in this work both in terms of horizontal location as shown in Figure 6.7 and in terms of vertical localization as shown in Figures 6.9 and 6.11. Ma's work showed that the performance of the CCF can be further improved by combining it with a single value ILD measurement taken as an average ILD value across the entire frequency range. In this work, the binaural signals are filtered by a gammatone filterbank to obtain several subbands, each with their own measurements of CCF and ILD. For each subband, a DNN was trained with the source azimuth determined by the average estimation from each DNN. This study was concerned with locating multiple sound sources, hence the employment of the gammatone filterbank. As we are only interested in locating a single sound source in this study, a single CCF and ILD measurement is taken across the entire frequency range and used to train a single DNN. While Ma's method is a robust method for localization in the horizontal, it is not intended specifically to be used for vertical localization. It should also be noted that the experiments performed in this study were performed using a sampling rate of 16kHz meaning that the CCF of length $\pm 1ms$ had a length of 33 and no frequencies above 8kHz were present in the stimuli.

6.4.2 Random Forest Training using ILD and IPD, the Wu Method

In 2019, Wu and Talagala presented a binaural SSL model capable of localizing both the azimuth and elevation of a single sound source [204]. This work combined the interaural features ILD and IPD and used random forest classification. The advantage of using a random forest over a neural network is that they can use information gain to determine which features are best for splitting the data into subsequent classifications. With this in

mind, one can feed a large amount of input data into a random forest algorithm and it will decide which individual features are best to use to train the ensemble while ignoring the features which have the least relevance. The amount of features each tree within the forest will use to make its estimation can be predefined using the max depth parameter in the sklearn RandomForestClassifier in python. This enables the user to train a network on large arrays of data knowing that only the most important cues will be considered. Wu's method takes advantage of this by training the algorithm on the phase and magnitude differences of a full FFT of the binaural signal. Their study used a sampling rate of 16kHz with an FFT of size 512, resulting in 256 values for IPD and ILD ranging from 0-8kHz.

To replicate this method, IPD and ILD can be calculated from a binaural signal using the following formulae:

$$v_f^p = \angle \left| \frac{X_{l,k,f}}{X_{r,k,f}} \right| \quad (6.9)$$

$$v_f^m = 20 \log_{10} \left| \frac{X_{l,k,f}}{X_{r,k,f}} \right| \quad (6.10)$$

where k and f represent the frame and frequency index, X_l and X_r are the FFTs of the left and right binaural signals respectively. While this method was shown to outperform the probabilistic piecewise affine mapping (PPAM) proposed by Deleforge in [43]. It was not tested for performance in the mismatched HRTF condition.

6.4.3 Comparing Multi-Feature Localization Performance

The methods employed by both Ma and Wu can be seen as benchmarks for localisation tasks using ML. Implementing their methods allows the same results to be achieved and also to examine the two methods in the mismatched scenario - something neither study investigated. In addition to this, a new multi-feature array for SSL in proposed comprised of a combination of CCF, ILD, and GTCC cues. This array can be seen as an extension of the Ma method which incorporates the best performing spectral feature, GTCCs from the single feature study presented earlier in this chapter. It is believed that this will allow for a better accuracy in the mismatched HRTF condition for vertical localization. While both the Ma and Wu methods were originally established using a sampling rate of 16kHz, in this work, a consistent sampling rate of 44.1kHz is used to fairly compare methods. Thus these methods are being applied to a larger frequency range and the size of the input vectors for each network is increased.

Firstly, both methods are implemented and their performance examined for localization in the horizontal plane. The Ma method uses a CCF with a lag range of $\pm 1ms$ producing a 33-dimensional binaural feature combined with a single value ILD measurement. The

resultant vector is of length 34 and is used to train a DNN with two hidden layers, each with 128 nodes. The network performs classification with a 'softmax' activation function applied to the output layer. The number of hidden layer nodes in the replication of the Ma method is increased from 128 to 300 due to the increase in the length of CCF when using a larger sampling rate.

The Wu method uses a combination of ILD and IPD values extracted from the magnitude and phase components of the FFT of a binaural signal. A 512 point FFT size is used in this method resulting in 256 values for both ILD and IPD measured between 0-8kHz. This feature space is fed into two separate random forest algorithms, one for determining the source azimuth and the other for determining the elevation. The RF algorithm for azimuth estimation consists of 50 trees and a max depth of 32 while the one for elevation estimation uses 100 trees and a max depth of 64. Both algorithms use information gain to determine feature importance and reduce the number of individual ILD and IPD points used to train the network. In replicating this method, the max depth of each tree in Wu's method is increased from 32 to 64 for horizontal plane localization and from 64 to 128 for vertical plane localization.

The combined feature set of CCF, ILD, and GTCCs is of length 158 and is used to train both a DNN and an RF algorithm. The extraction of these features is as described earlier in this chapter. When training the DNN with this feature set, the same architecture is used as shown in Figure 6.5 using a hidden layer size of 300 nodes. When training the RF algorithm with this feature set, 50 trees and a max depth of 32 is used for horizontal estimation while 100 trees with a max depth of 64 is used for vertical localization. These two new algorithms are compared with the Ma and Wu methods in terms of performance in the horizontal plane, median plane and sagittal plane at 45°. The average result of these algorithms along these planes can be seen in Figures 6.12, 6.13, and 6.14.

6.4.4 Discussion

From Figure 6.12, it can be seen that the novel feature set of CCF, ILD, and GTCCs performs worse along the horizontal plane than the methods presented by Ma and Wu when used to train a DNN. This is true for both matched and mismatched HRTF conditions. However when used to train an RF algorithm, this feature set performs on par with both methods in both matched and mismatched conditions. This is to be expected given that information gain is used to define feature importance in the RF algorithm. It is likely in this case that the interaural cues CCF and ILD were used most often among the trees in the RF as they contained more relevant information towards determining the source azimuth. The significantly poorer performance of the feature set when using a DNN algorithm is somewhat surprising given that a subset of this feature set is used to greater performance in

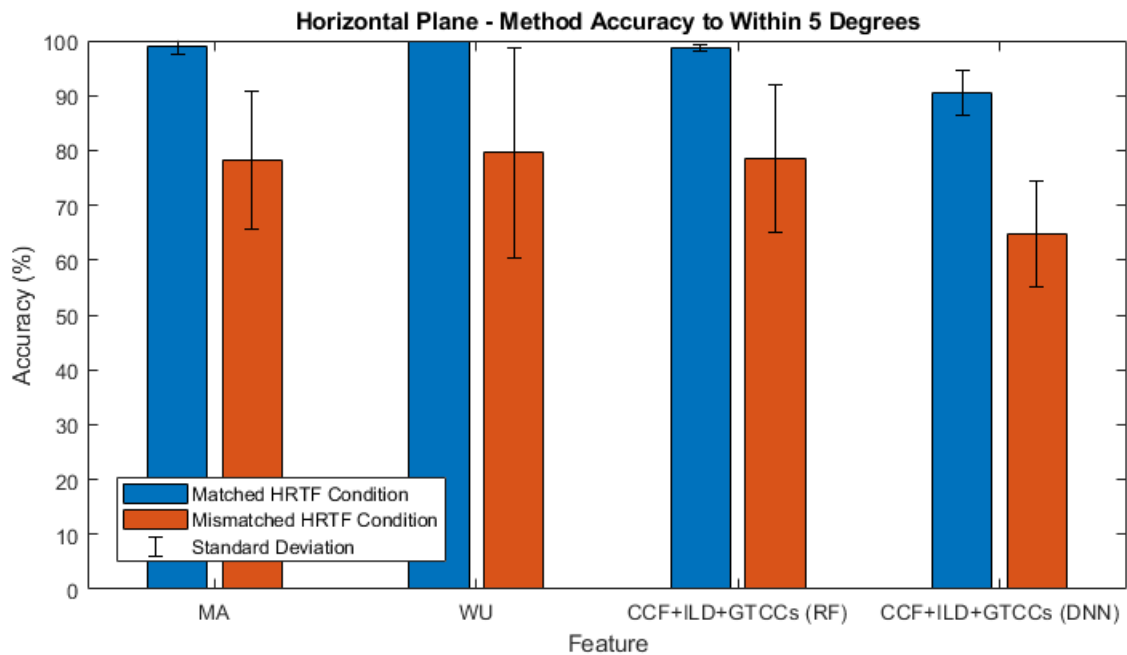


Fig. 6.12 Horizontal Plane accuracy in the matched and mismatched condition for azimuth estimation to within 5 degrees.

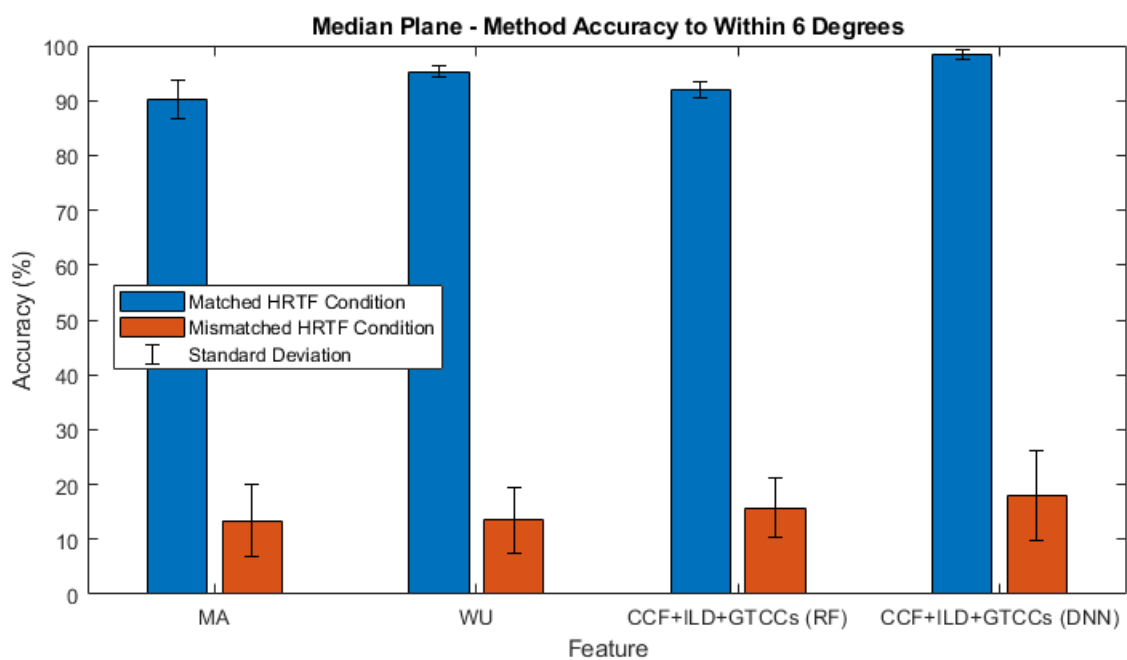


Fig. 6.13 Median plane accuracy in the matched and mismatched condition for elevation estimation to within 6 degrees.

the Ma method. This decrease in performance is most likely due to the network architecture which was the same as that used in replicating the Ma method. Its possible that increasing

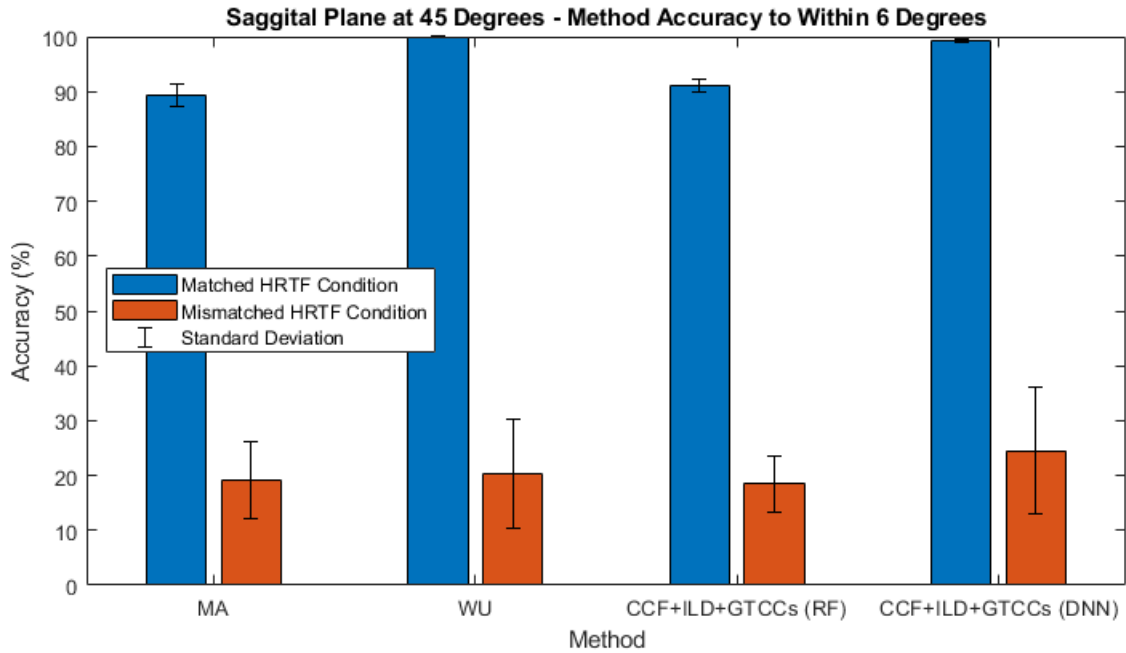


Fig. 6.14 Sagittal plane at 45° accuracy in the matched and mismatched condition for elevation estimation to within 6 degrees.

the number of nodes in the hidden layers or adjusting the learning rate would yield a better performance.

Along the median plane, the proposed feature set outperforms both Ma and Wu methods in the mismatched condition for both RF and DNN implementations with the DNN implementation performing best. This observation can again be made along the sagittal plane at 45° where the DNN implementation of the new feature set exhibits an accuracy of 24.54% in the mismatched condition, over 4% better than the Wu method which achieves an accuracy of 20.3%. Another observation that can be made in this instance is that the standard deviation is highest for this method which is actually indicative of better generalization showing this method can perform better for HRTF sets which are alike and worse for those which are different. For instance, training on HRTF set 1 achieves an accuracy of 63% when testing on HRTF 32 indicating these two HRTF sets have similar properties. Conversely when testing on HRTF 45, an accuracy of just 6.4% can be observed showing that these two HRTF sets are very different. These testing performances can be seen in Figure 6.15. It is evident on observing these performance matrices that the proposed feature set employed using a DNN is the most suitable method for estimating source location along the vertical. While there are many training and testing conditions which perform badly, there are also many which perform well above the average implying that this method makes use of a similarity between HRTF measurements evident in the GTCC cue.

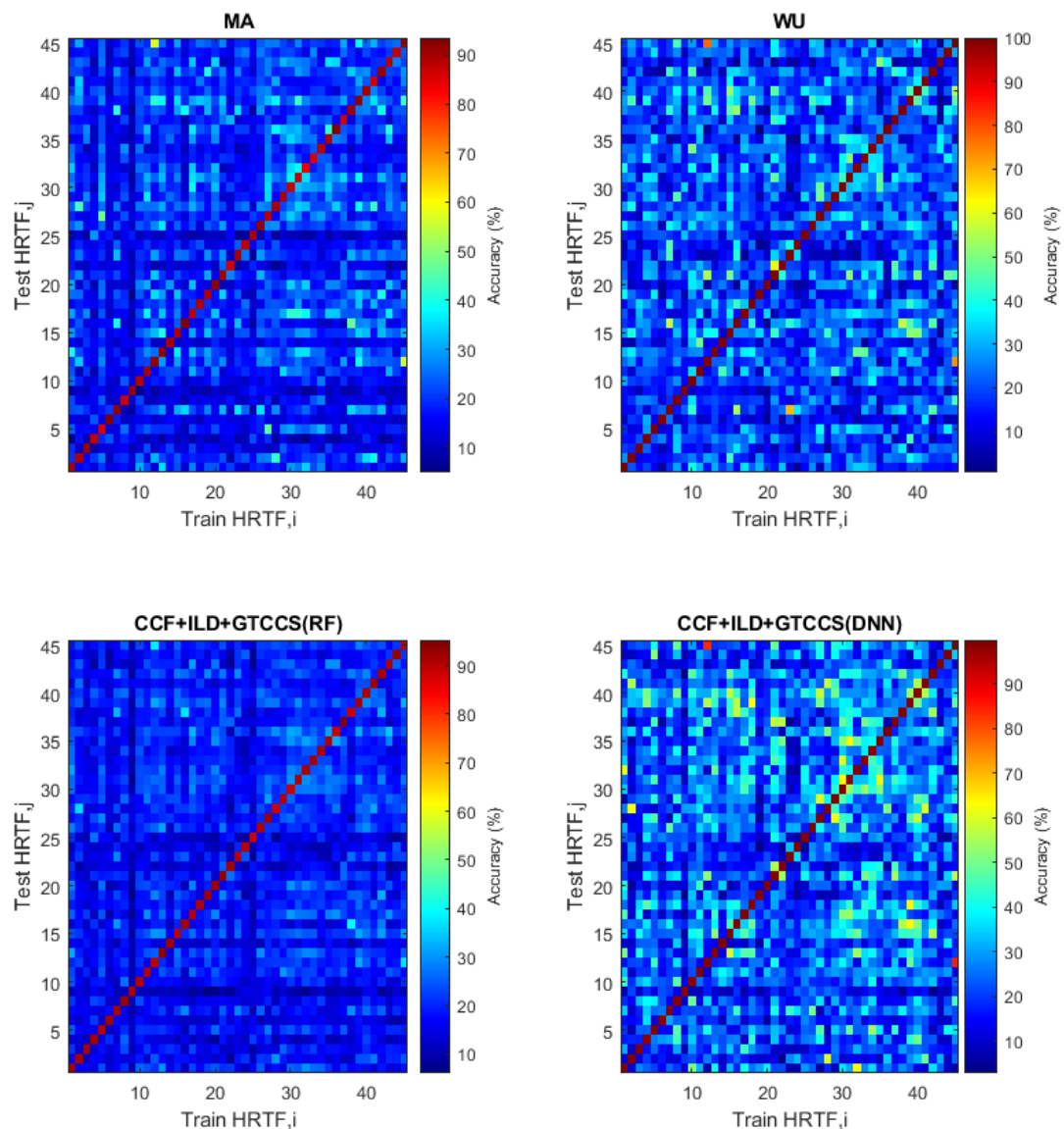


Fig. 6.15 Sagittal plane at 45° method performance for estimating elevation to within 6°. A single training HRTF is used to train the algorithm and subsequently tested on each of the 45 HRTFs from the CIPIC dataset. Two popular methods of binaural localization are compared to two new proposed methods.

6.5 Conclusion

In this chapter, a series of ML algorithms were trained to identify the location of a single speech sound source in binaural audio signals. Using features extracted from these audio signals, an examination of performance under both matched and mismatched

HRTF conditions is made. Results show that while interaural cues such as interaural time difference, interaural level difference, and the normalized cross-correlation function perform best when localizing along the horizontal plane, spectral cues are required to improve performance for elevation estimation along vertical planes.

From this single feature analysis, a new feature combination is proposed for improving vertical localization in the mismatched HRTF condition. This method uses a combination of CCF, ILD, and gammatone cepstral coefficient cues trained with a deep neural network to improve vertical localization in the mismatched HRTF condition by up to 4% compared to recent state of the art methods. The performance of GTCCs over other spectral cues validates the use of the equivalent rectangular bandwidth filter in tasks related to sound source localization. This cue modelled on the human auditory system via cochlea simulation has been shown to capture vertical cues in the HRTF.

This method also shows an ability to identify similar HRTF sets and equally dissimilar HRTF sets with regards to vertical localization. While this results in a large variance in testing results, this is indeed a positive finding which will be expanded upon in the following chapter.

Chapter 7

HRTF Clustering for Robust Full Sphere Sound Source Localization

Overview: In this chapter, the adaptive affinity propagation algorithm is introduced and used to cluster HRTFs from the CIPIC database into groups of similar HRTFs. Results show how training an ML algorithm using the central HRTFs from these clusters results in an algorithm that is much more robust for localizing sound sources in the mismatched HRTF condition and also within noisy environments. This performance is measured on the sphere where an angular error is required to determine the algorithms error accurately. Using anthropometric data from the CIPIC database, a random forest algorithm is trained to determine the cluster of a subject based on their pinna geometry. Using this model, the feature importance of each pinna measurement is determined. This reveals that the cavum concha width is the pinna feature which has the largest determination of HRTF clustering. This finding is in agreement with many physical pinna models which have been developed in the past.

7.1 Introduction

The problem of localization error in the mismatched condition can be targeted by including more or even all of the HRTF sets in the training data. However, increasing the amount of training data has a direct effect on the increase in time it takes to train the network and may also reduce the overall accuracy of the algorithm as weights struggle to fit to a large number of HRTF sets. This is a brute force solution to the problem and serves as a sub optimal approach.

A more streamlined approach to improving performance in the mismatched HRTF condition is to include a select few HRTF sets in the training data which are representative of the entire population of HRTF sets. These optimal HRTF sets can be determined using

a clustering algorithm which clusters HRTF sets into a discrete number of sets based on their relative training and testing performance.

Localization similarity between HRTF sets can be seen in a number of figures in the previous chapter which examined localization performance for different features. For instance, if we examine the performance of CCF along the horizontal plane in Fig 6.6, the localization accuracy at (12,45) and (45,12) is high. This indicates that the similarity between these two HRTFs is high and that the DNNs trained on them respectively can be substituted for each other. Not surprisingly, these particular HRTF sets are those both measured on the KEMAR dummy head microphone using different ear sizes. Conversely, the localization accuracy at (6,27) is the lowest at 2.64% showing a great difference in HRTF 6 and HRTF 27. These similarity matrices can be used to identify clusters of HRTF sets which are alike and also find the HRTF at the centre of the cluster which best defines the entire cluster.

In this study, a clustering analysis based on the affinity propagation (AP) algorithm is employed. Compared with other common clustering methods such as K-means and hierarchical clustering, the AP algorithm has many unique advantages. For instance, unlike K-means, the AP algorithm does not need any prior information on the number of clusters and will determine the most appropriate number of clusters from the data. The AP algorithm also selects a real HRTF to define a cluster centre whereas K-means clustering selects a centre relating to an average of all sample points which most likely will not correspond to a real HRTF. Finally, the AP algorithm can be applied to problems where the similarity matrix is not symmetrical, i.e. the performance of the DNN trained on HRTF set i and tested on HRTF set j is not equal to the performance of the DNN trained on HRTF set j and tested on HRTF set i .

The AP algorithm was used by Wang et al. to perform HRTF clustering based on horizontal localization performance only [194]. Building upon this work in this chapter, an Adaptive form of the AP algorithm is used to cluster HRTF sets together based on performance in the vertical plane.

7.2 The Affinity Propagation Clustering Algorithm

7.2.1 The AP algorithm

The *Affinity Propagation*, AP clustering algorithm identifies clusters in a data set $X : [\mathbf{x}_1, \dots, \mathbf{x}_k, \dots, \mathbf{x}_N]$ where, we will assume, $\mathbf{x}_k \in R^n$. The criterion underlying the clustering is a distance metric $d(i,k) = \text{distance between } \mathbf{x}_i \text{ and } \mathbf{x}_k$. Here we will use $d(i,k) = \|\mathbf{x}_i - \mathbf{x}_k\|^2$

In AP the data points are viewed as the 'nodes' of a network and the algorithm takes as

an input a matrix of real-valued similarities between the nodes. This *Similarity* matrix $\mathbf{S}$ with elements $s(i, k)$ is derived from the distance metric and so is symmetric.

The matrix $\mathbf{S}$ is

$$s(i, k) = -d(i, k) = -\|\mathbf{x}_i - \mathbf{x}_k\|^2 \text{ for } i \neq k \quad (7.1)$$

$$s(k, k) = -\sigma_k^2 \text{ these are called 'preferences'} \quad (7.2)$$

The choice of a larger value for a *preference* would prioritise that data value as a potential *exemplar* i.e. as a cluster centre. If, *a priori*, all data values are equally suitable as exemplars, then, the advice from the developers [59] of AP, is to select a common value of σ^2 for all the *preferences*. This could be the median of all $s(i, k)$ for $i \neq k$ to obtain a moderate number of clusters or their minimum to have a small number of clusters.

The AP algorithm takes this initial knowledge of the Similarity and proceeds to exchange 'messages' between the nodes. The messages are numeric and of 2 types: (i) There is the '*Responsibility*' stored in a matrix $\mathbf{R}$ with elements $r(i, k)$. The Responsibility values $r(i, k)$ reflect how well-suited the data point $\mathbf{x}_k$ is to be an *exemplar* or 'cluster centre' for the data point $\mathbf{x}_i$. There is also (ii) the '*Availability*' stored in a matrix $\mathbf{A}$ with elements $a(i, k)$. The Availability values $a(i, k)$ reflects how appropriate it would be for the data point $\mathbf{x}_i$ to choose data point $\mathbf{x}_k$ to be its exemplar. The responsibilities and availabilities are computed iteratively using the following rules, starting with the initial availability values $a(i, k) = 0$.

$$r(i, k) \leftarrow s(i, k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i, k') + s(i, k')\} \quad (7.3)$$

$$a(i, k) \leftarrow \min\{0, r(k, k) + \sum_{i' \text{ s.t. } i' \notin \{i, k\}} \max\{0, r(i', k)\}\} \quad (7.4)$$

$$a(k, k) \leftarrow \sum_{i' \neq k} \max(0, r(i', k)) \quad (7.5)$$

Iteration through these rules, till there is little change to the elements of the $\mathbf{R}$ and $\mathbf{A}$ matrices, or for a fixed number of iterations gives Responsibility and Availability values that can be combined to identify the exemplars of a clustering and an associated labelling of the data set. The combination and exemplar identification for data point $\mathbf{x}_i$ are done using

$$\hat{k} \text{ is } \max_{\text{wrt } k} E_x = a(i, k) + r(i, k) \quad (7.6)$$

Then if $\hat{k} = i$ we have $\mathbf{x}_i$ is an exemplar, otherwise $\mathbf{x}_{\hat{k}}$ is the exemplar for $\mathbf{x}_i$.

To dampen oscillatory behaviour, when these rules are applied recursively, the developers of the algorithm advise they should be embedded in a simple first-order recursion. Selecting a damping factor $0 \leq \lambda \leq 1$ and with an iteration counter α we get

$$r^{\alpha+1}(i, k) = \lambda r^{\alpha}(i, k) + (1 - \lambda) \Gamma_1^{\alpha} \quad (7.7)$$

$$a^{\alpha+1}(i, k) = \lambda a^{\alpha}(i, k) + (1 - \lambda) \Gamma_2^{\alpha} \quad (7.8)$$

Where Γ_1^{α} and Γ_2^{α} are

$$\begin{aligned} \Gamma_1^{\alpha} &= [s(i, k) - \max_{k' \text{ s.t. } k' \neq k} \{a^{\alpha}(i, k') + s(i, k')\}] \\ \Gamma_2^{\alpha} &= \min\{0, r(k, k) + \sum_{i' \text{ s.t. } i' \notin \{i, k\}} \max\{0, r(i', k)\} \text{ for } i \neq k \\ \Gamma_2^{\alpha} &= \sum_{i' \neq k} \max(0, r(i', k)) \text{ for } i = k \end{aligned}$$

7.2.2 An adaptive AP algorithm

The AP algorithm has been shown to provide a robust engineering solution to clustering problems in audio signal processing such as speaker clustering [205] and acoustic control of crosstalk [105] and in VANETs for mobile telecommunications [173]. But the number of clusters it will identify is associated with the value assigned to *preferences*. As stated earlier, the number of clusters, identified by the algorithm, is adjusted by the choice of value of σ^2 . To address the combined challenges of adjusting λ to enhance convergence and dampen oscillations and adjusting σ to optimize the number of clusters identified an *Adaptive Affinity Propagation*, AAP clustering algorithm has been proposed and released as Matlab code [196],[195]. In the AAP algorithm the value of λ is adapted using a moving window observing the iterations and under the competing influences of damping oscillatory tendencies by $\lambda \rightarrow 1$ and faster convergence with $\lambda \rightarrow 0$. This stabilizing adaptation of λ accelerates convergence and facilitates an adaptation of the assignment of a value to σ . While no analytic relationship is known between σ and NC = number of cluster, it is known that as $-\sigma \uparrow [-\infty, 0]$ then $NC \uparrow [1, N]$. So by initialising $\sigma = \min\{s(i, k)\}$ and using the *Silhouette Index*, as a performance criterion the outer adaptation loop adjusts the value of σ to optimise the number of clusters.

7.2.3 The Silhouette Index

The *Silhouette Index*, SI, is a measure of the compactness and separation of clusters and consequently it is a widely-used. It is an internal validation index as it provides a measure

of goodness of the clustering without recourse to external information. In the Adaptive Affinity Propagation algorithm it is used to estimate the optimum number NC of clusters.

Description: Let a data set of n samples be divided into k clusters C_i , with data points $i, j \in C_i$. For distance $d(i, j)$, the mean distance $a(i)$ between i and all other data points in C_i is

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

The value of $a(i)$ is a measure of how suited data point i is to its assigned cluster C_i .

The *mean dissimilarity* of a data point $i \in C_i$ to some other cluster C_k is defined as the mean of the distance from i to all data points in C_k . For each data point $i \in C_i$ we define

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

So $b(i)$ is smallest mean distance of i to all the points in any other cluster.

The *Silhouette value* of data point i is

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}, \text{ if } |C_i| > 1$$

and $s(i) = 0$, if $|C_i| = 1$ (7.9)

Note from this definition it follows that $-1 \leq s(i) \leq 1$

The *Silhouette Index*, SI of a given clustering of a data set of n samples is the average value of $s(i)$, that is,

$$SI = \frac{1}{n} \sum_{i=0}^n s(i)$$

The range of the SI values is $-1 \leq SI \leq 1$ and the clustering with the highest SI is optimal.

7.3 Clustering Analysis

As the primary concern of this study is to improve the accuracy of vertical localization, HRTF set clustering was performed with respect to the proposed feature set of CCF,ILD, and GTCCs cues and its performance along the sagittal plane. As this feature set contains CCF and ILD cues which generalize well to localization tasks along the horizontal it is believed that after clustering, performance along the horizontal plane will still be

sufficiently accurate. Performing clustering with the *AAP* algorithm reveals an optimum clustering of 4 given the similarity matrix found in Figure 7.1.

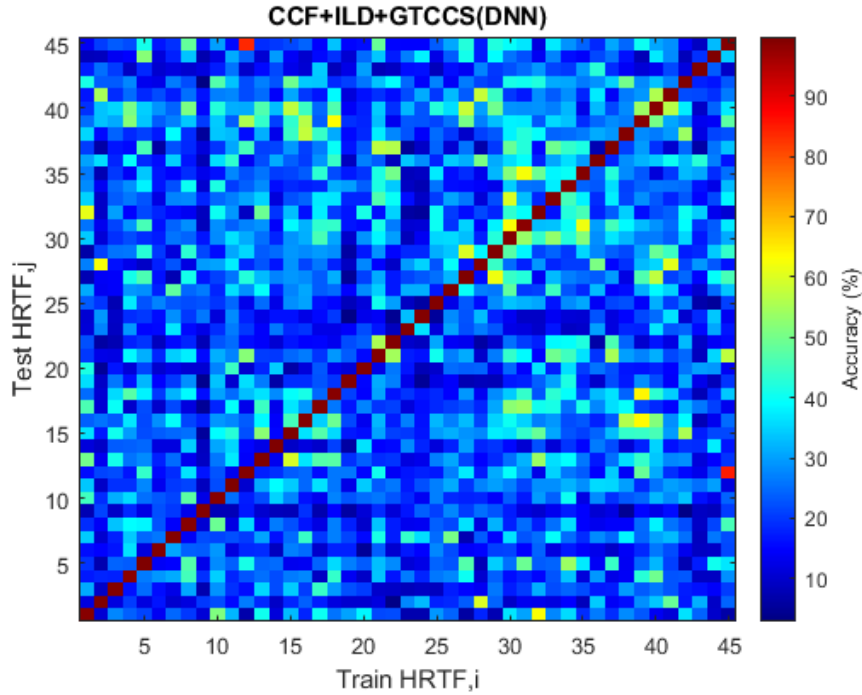


Fig. 7.1 The Similarity Matrix for the proposed method with performance scores based on sagittal plane localization accuracy to within 6° .

These clusters range in size from 9 to 13 and can be seen in Table 7.1. The silhouette index for this particular clustering is $SI = 0.2121$ found using a damping factor, λ of 0.5. This value was the highest SI value for arrangements from 2 clusters up to 10. While this is comparatively smaller than results produced in other clustering studies it should be noted how abstract this clustering is as it is based off localization scores from binaural audio signals. SI values as well as minimum Silhouette values for each cluster number can be seen in Figure 7.2.

Following this clustering analysis, sagittal plane localization is repeated at 45° with respect to elevation angle. This time HRTF clustering is employed whereby the 4 central HRTF sets from each cluster are used to train the DNN. This network is then tested on the 45 HRTF conditions whereby 4 are matched and 41 are mismatched. The result of this test shows that clustering reduces performance in the matched condition while enhancing it in the mismatched condition as shown in Figure 7.3. Here performance goes from 99.27% in the matched condition to 92.92% in the matched condition and from 24.54% in the mismatched condition to 45.8%. This improvement of over 20% in the mismatched condition shows a much better generalization in the network to mismatched HRTF conditions. Training the DNN using a single HRTF set condition takes on average

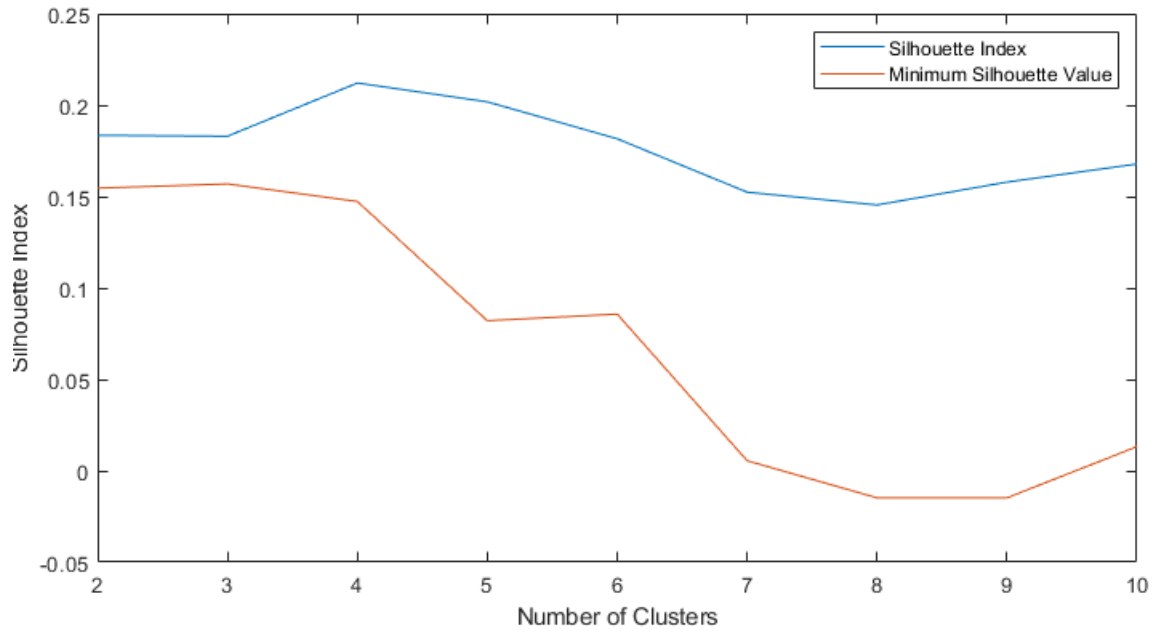


Fig. 7.2 The Silhouette Index values for 2 to 10 clusters found using the Adaptive Affinity Propagation algorithm.

Table 7.1 Clustering result with 4 clusters

Cluster index	No. of HRTF sets	HRTF indices	Central HRTF
Cluster 1	13	1,3,7,10,11,13,15,17,19,20,32,38,42	1
Cluster 2	13	4,5,8,12,14,16,18,27,29,39,40,44,45	27
Cluster 3	10	2,6,9,23,24,26,28,36,41,43	28
Cluster 4	9	21,22,25,30,31,33,34,35,37	37

13.96 seconds while training the DNN using 4 HRTF sets as is the case in this clustered approach takes on average 45.93 seconds.

7.4 Spherical Localization and Angular Error

In the previous chapter, sound source localization was examined in terms of performance along a single plane. However, real world applications of sound source localization apply to localization tasks in three dimensions, i.e. a real world sound source localization algorithm needs to be able to robustly determine the azimuth, elevation, and distance of a sound source.

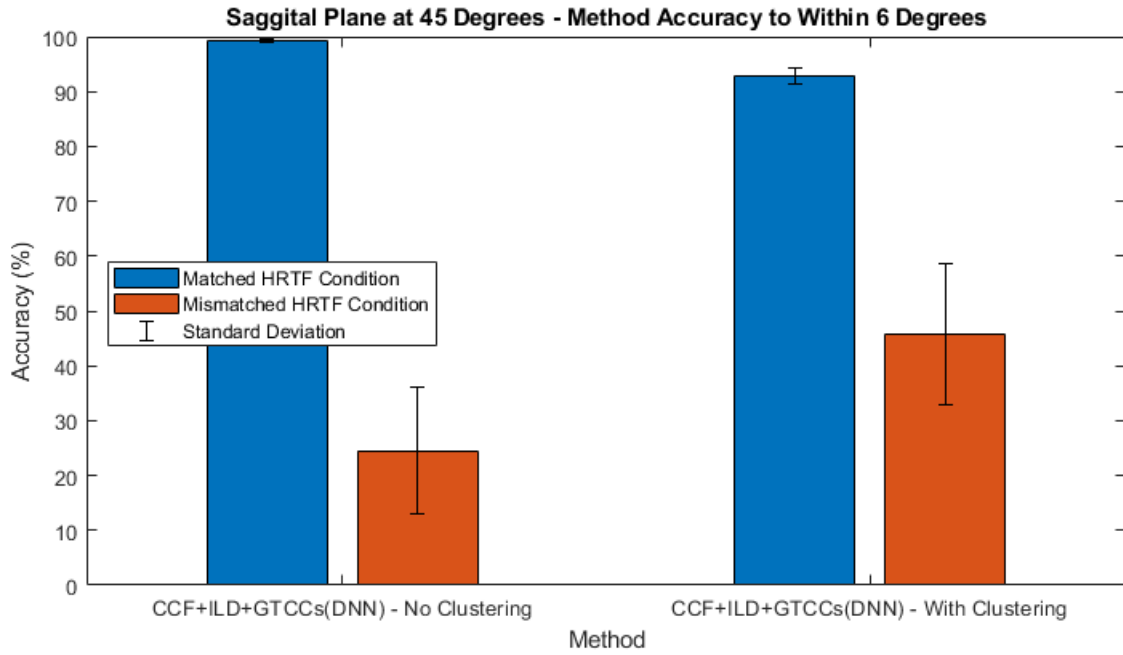


Fig. 7.3 The performance of the proposed feature set along the Saggital Plane at 45°, correctly classified to within 6°. Performance is compared for training on a single HRTF set versus training on 4 central HRTF sets.

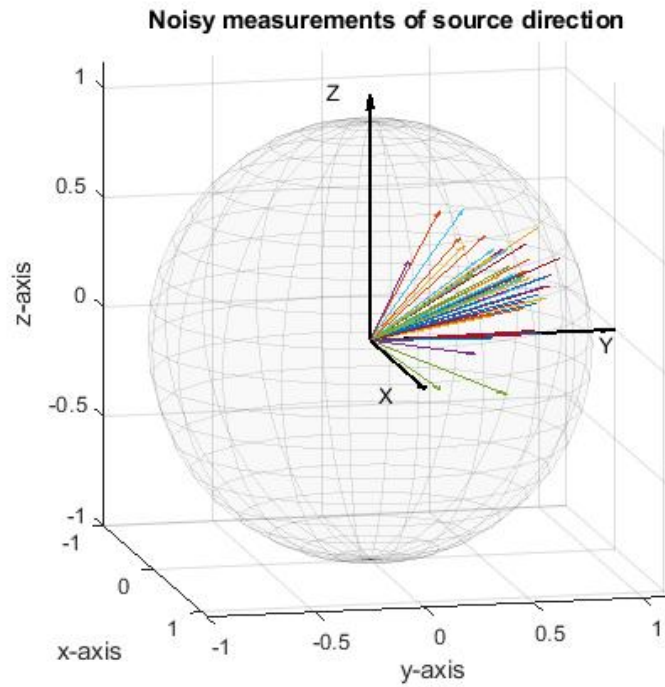


Fig. 7.4 Source direction estimates from noisy signals.

Spherical data, such as estimates of the direction of a sound source derived from binaural signals, lie on the unit sphere, $\mathbf{S}^2 = \{\mathbf{x} \in \mathbb{R}^3: \mathbf{x}^T \mathbf{x} = 1\}$. This is of course assuming that distance is not taken into consideration as is the case in this chapter.

In Figure 7.4, a sample of 50 estimates of the direction of a source are given at azimuth = 30° and elevation = 20°. In this instance, estimates of the source direction can have an error in both horizontal and vertical planes and thus the error can not be easily derived via simple subtraction.

When calculating the error of source estimation on the unit sphere, an angular error which reflects the error in 3D space must be used. In order to perform this calculation of angular error, both ground truth source angles and predicted source angles must first be converted from degrees to radians. Following this, the three dimensional coordinates must be converted from polar to cartesian coordinates subbing in a constant value for distance, in this study defined as 1 metre. This angular error can be defined by the absolute angular difference between two directional vectors in a Cartesian coordinate system and calculated as:

$$\varepsilon = \angle \frac{d_{(\theta,\phi,r)} \cdot \hat{d}_{(\theta,\phi,r)}}{|d_{(\theta,\phi,r)}| |\hat{d}_{(\theta,\phi,r)}|} \quad (7.10)$$

where $d_{(\theta,\phi,r)}$ and $\hat{d}_{(\theta,\phi,r)}$ are the ground truth and estimated source directions respectively. The angular error can then be converted back from radians to degrees.

7.4.1 Localization Performance

Given the feature set proposed for this method of performing sound source localization, spherical localization is performed using a combination of two machine learning algorithms. The first is a random forest network trained to estimate the source azimuth. It consists of 50 trees and has a maximum depth of 32 features. The second algorithm is a DNN used to estimate the source elevation. It has the same architecture as the DNN used in the previous chapter with 3 hidden layers consisting of 300 nodes, a dropout rate of 20% and a learning rate of $10e^{-4}$.

The performance of the proposed method is presented with and without HRTF clustering. As in previous studies, binaural signals are generated by convolving HRTF datasets from the CIPIC database [5] with random speech signals from the TSP speech dataset [90]. Here, the received binaural signals are received from 625 source locations in the front hemifield with 25 sagittal planes each containing 25 HRTF measurements along the vertical. Performance is measured in terms of average angular error in both the matched and mismatched HRTF condition. White gaussian noise is added to the speech signals prior to feature extraction at 4 different signal-to-noise ratios (SNRs); 10dB, 20dB, 30dB, & No Noise. These results can be seen in Table 7.2.

These results clearly show that HRTF clustering significantly decreases angular error in the mismatched HRTF condition across all SNR levels. For all SNR levels the results

found using HRTF clustering are greater than 5° better than those found without HRTF clustering. While performance using HRTF clustering is worse in the matched condition it's important to note that there are 4 matched conditions and 41 mismatched conditions. And while there is a trade-off between matched and mismatched performance using HRTF clustering, it can be concluded that this is worth taking given the improvement of results in a majority of HRTF conditions.

Table 7.2 Average Angular Error in Noisy environments with and without HRTF Clustering

Average Angular Error								
SNR	10dB		20dB		30dB		No Noise	
HRTF Condition	Matched	Mis-Matched	Matched	Mis-Matched	Matched	Mis-Matched	Matched	Mis-Matched
No Clustering	7.04°	19.7°	5.1°	18.72°	3.92°	17.65°	2.55°	16.87°
With Clustering	9.37°	13.8°	7.35°	11.79°	6.17°	10.6°	5.22°	10.23°

In Table 7.3, the azimuth and elevation accuracy's of the algorithms are shown. These are categorised as the percentage of samples correctly classified to within 5° in the case of azimuthal accuracy and to within 6° in the case of elevation accuracy. In terms of Azimuthal accuracy, these results show a consistency in results when using HRTF clustering with performance remaining approximately equal across all noise conditions in both the matched and mismatched conditions. This is not the case when HRTF clustering is not applied however as while performance is greater than with HRTF clustering in the matched condition, it is increasingly worse in the mismatched HRTF condition with increasing levels of noise.

When examining elevation accuracy, we can see that results both with and without HRTF clustering diminish with increasing levels of noise in the signal. As expected, performance in the matched condition is consistently better across all noise levels when there is no HRTF clustering. However, mismatched HRTF performance is better in all noise conditions when HRTF clustering is applied with an accuracy of 28.77% at SNR = 10dB better than an accuracy of 25.87% in the no noise condition when there is no HRTF clustering. Mismatched performance is improved by up to 13% when HRTF clustering is employed.

7.5 Anthropometric Feature Importance

The CIPIC database contains a large amount of anthropometric relating to each subject in the database [5]. More specifically, there are 17 head and torso measurements for each

Table 7.3 Average Azimuth and Elevation Accuracy with and without HRTF Clustering

Azimuth Accuracy Comparison, Error tolerance to $\leq 5^\circ$								
SNR	10dB		20dB		30dB		No Noise	
HRTF Condition	Matched	Mis-Matched	Matched	Mis-Matched	Matched	Mis-Matched	Matched	Mis-Matched
No Clustering	96.252%	68.5%	97.74%	75.84%	97.156%	79.72%	98.752%	80.34%
With Clustering	93.01%	85.32%	93.38%	85.42%	93.59%	85.36%	93.48%	85.57%

Elevation Accuracy Comparison, Error tolerance to $\leq 6^\circ$								
SNR	10dB		20dB		30dB		No Noise	
HRTF Condition	Matched	Mis-Matched	Matched	Mis-Matched	Matched	Mis-Matched	Matched	Mis-Matched
No Clustering	79.13%	20.11%	84.38%	23.1%	89.775%	25.16%	94.64%	25.87%
With Clustering	64.13%	28.77%	73.08%	33.73%	80.19%	35.67%	87.5%	38.75%

subject as well as 10 pinna features measured for each ear. This data is measured in full for 35 of the 43 human subjects while anthropometric data for the two KEMAR measurements are approximated where possible. For instance, there are a full set of pinna geometry measurements but no measurements for height and weight.

As the clustering analysis performed earlier in this chapter was capable of grouping the HRTF measurements of the subjects in this database into 4 separate clusters, it makes it possible to train a decision tree algorithm to determine the correct cluster of a subject based on their anthropometric data. Following the training of an accurate model, the importance of each feature in terms of a coefficient value can be determined. Note that due to the stochastic nature of the algorithm, results may vary and thus, feature performance should be performed several times, comparing the average outcome. In this instance, feature performance was calculated 10 times with the feature values of the left and right ears used 5 times each. In each test, feature performance scores for all features sum to 1.

As there are 37 features to choose from in this study and just 37 subjects with sufficient anthropometric data, it is necessary to reduce the number of features to just a few desired relevant features. As, clustering was performed with respect to vertical localization, an assumption can be made that the most relevant features are those related to the pinna geometry. A random forest classifier was used with 16 trees and a max depth of 4 to model these pinna geometries to the subjects cluster number. Across the 10 tests employed, an average performance of 82.22% was found when approximating the subjects cluster number based on their pinna measurements implying that pinna geometry is quite successful in estimating the cluster number of a subject.

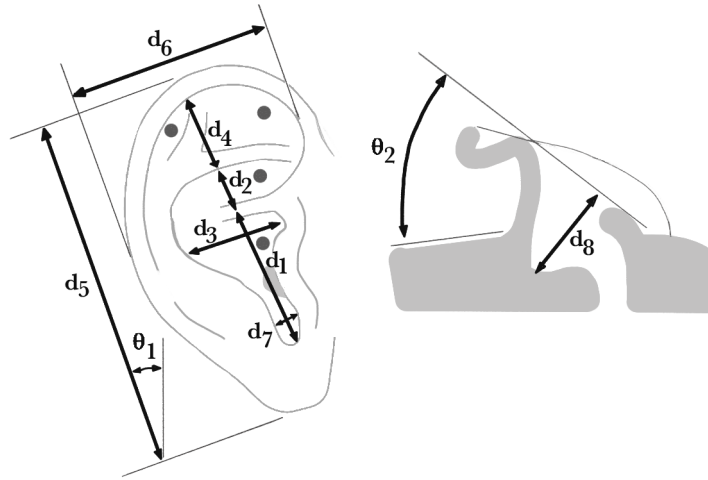


Fig. 7.5 Pinna geometry measurements of the CIPIC database, reproduced after Algazi [5].

The results of this average feature importance performance can be seen in Table 7.4. Here, it can be noted that there is a single stand out feature with regards to determining the cluster number of a subject. This being cavum concha width. Scoring well above all other features, this is the single most important feature in determining subject HRTF cluster based on this analysis. With this in mind, a closer look at how cavum concha width relates to the 4 clusters is derived in the form of two box plots shown in Figure 7.6.

Table 7.4 Pinna Feature Performance

Variable	Anthropometric Feature	Average Performance	Rank
d1	Cavum Concha Height	0.087	8
d2	Cymba Concha Height	0.0886	7
d3	Cavum Concha Width	0.1844	1
d4	Fossa Height	0.0848	9
d5	Pinna Height	0.1023	4
d6	Pinna Width	0.094	5
d7	Intertragal Incisure Width	0.0775	10
d8	Cavum Concha Depth	0.0896	6
θ_1	Pinna Rotation Angle	0.138	2
θ_2	Pinna Flare Angle	0.053	10

While there is no clear distinction between cavum concha widths across all clusters, there is a general trend with cluster 4 having the lowest average value in both ears and cluster 1 having the most compact distribution towards the upper values in both ears. This result is also interesting as it is in agreement with findings by Iida et al. which identified

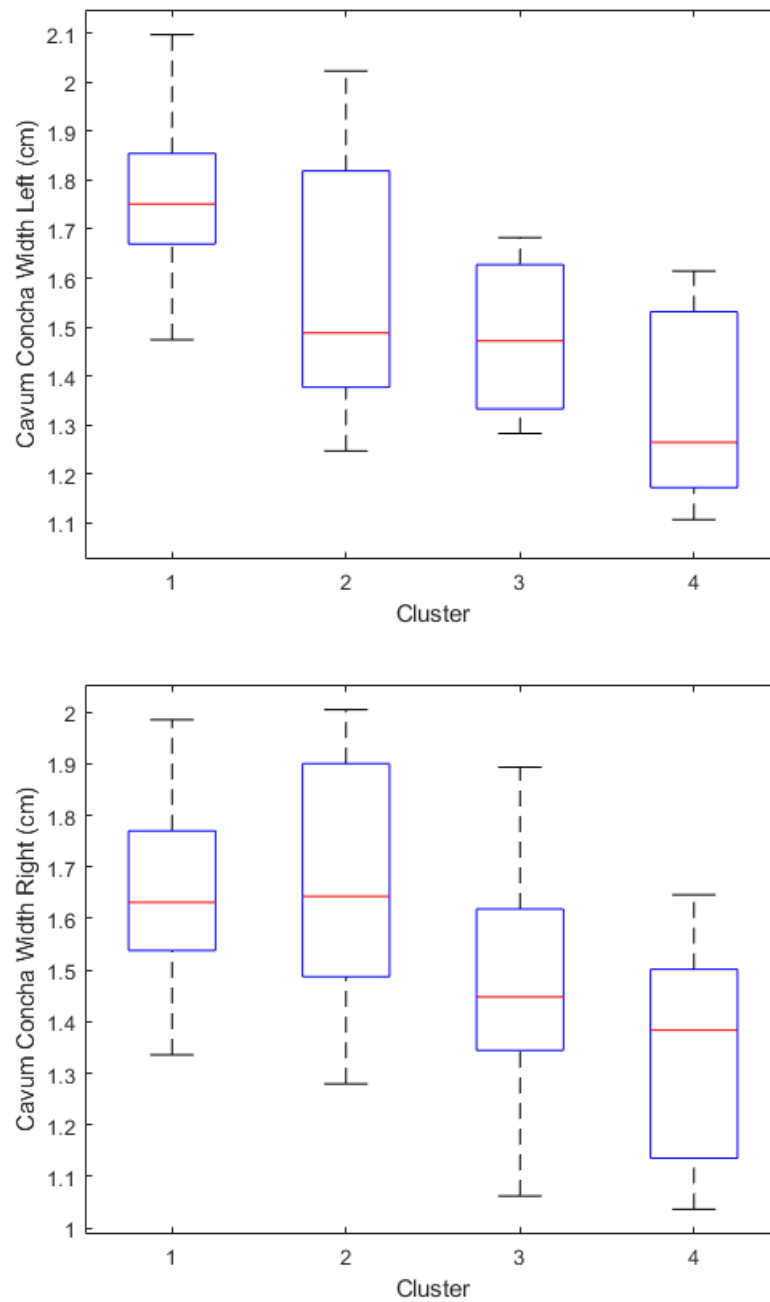


Fig. 7.6 Box plots of cavum concha width in both left and right ears for subjects within each cluster.

the cavum concha width as a feature which has a significant influence on N1, the first notch frequency which informs our sense of vertical localization [84].

Indeed, the concha has been the subject of many models attempting to replicate reflections caused by sound sources at varying heights as described by Algazi in [4]. Genuit's structural HRTF model featured a component replicating the cavum conchae

reflection as it was determined along with the shoulder to produce the most significant reflections from a rising source angle [62]. By attaching artificial pinnae to the ear, Batteau showed experimentally that the pinnae provide substantial elevation cues and even azimuthal cues as well [14]. Batteau proposed a simple model, with the wall of the concha and the rim of the pinna acting as reflectors. Batteau's structural pinna model was formalized by Watkins in [197] and was shown to fit well with experimental data and produce good elevation effects. This Model was extended further by López-Poveda and Meddis who included the effects of diffraction to explain why reflections from the concha posterior wall are observed for sound sources in the lateral median plane as well as the median vertical plane [114]. In another study, López-Poveda examined the effects that filling pinna cavities and removing pinna structures had on spectral notches. With the pinnae studied, one of the major conclusions was that the interaction between the flange and the concha is responsible for both the depth and location of N1 [152].

7.6 Conclusion

In this chapter the adaptive affinity propagation clustering algorithm is introduced and is used to cluster similar HRTF sets together. This clustering is based on the performance of the proposed feature set of CCF, ILD, and GTCC cues in estimating sound source elevation on the sagittal plane at 45° using a DNN as was presented in chapter 7. This clustering algorithm reveals 4 unique clusters ranging in size from 9 to 13 HRTF sets, each with their own central HRTF which is most representative of the entire cluster. Using these 4 HRTF sets, a DNN is trained and used to perform the same localization task along the sagittal plane as was used to determine the HRTF clustering. It is shown that the use of HRTF clustering improves the performance of vertical source estimation by over 20% in the mismatched HRTF condition. Concurrently, HRTF clustering also decreases localization performance in the matched HRTF by approximately 6%.

When examining the effect of HRTF clustering on angular error, it was shown that even in noisy environments, HRTF clustering improves the performance of the sound source localization algorithm in the mismatched HRTF condition. At an SNR of 10dB, the average angular error of the algorithm using HRTF clustering is 13.8° which is superior to the value of 16.87° exhibited in the condition where no clustering is used and no noise is present. On examination of error with regards to azimuth and elevation, HRTF clustering results in a much more robust performance in the mismatched HRTF condition for both. Azimuthal performance is shown to be robust across all levels of noise while elevation performance surpasses no clustering performance by up to 13% in the mismatched HRTF condition.

Finally, a random forest algorithm is trained on anthropometric data from the CIPIC database to determine the cluster number of subjects. Using just pinna measurements, an average accuracy of 82.22% was found for this model which was able to reveal the pinna features which had the most importance in training the model. These results showed that the width of the cavum concha had the most significance in determining the cluster of each subject using this analysis. This implies that the cavum concha is an important feature in the human auditory system for performing vertical localization.

Chapter 8

Conclusion

In the introduction of this thesis it was stated that:

"Modern tools such as deep neural networks can be employed to help us better understand and model the perception of sound source localization, a phenomenon which is highly unique to the individual. This thesis proposes that in using these tools, we can gain a deeper understanding of the cues which allow us to perform sound source localization in three dimensional space. This in turn can allow us to understand the individuality of the HRTF. As a majority of spatial audio is synthesised using transfer functions, there is a need to fully comprehend it both in terms of subjective quality and directional encoding accuracy. In doing so, we may therefore enhance the sense of immersion experienced in virtual auditory environments. "

The work presented in this thesis addresses the human perception of spatial audio with the work presented in chapter 6 and 7 seeking to address the major proposal set forth in the beginning. All of the research however has a common goal in better understanding our perception of spatial audio and how modern technology can be used to better understand, capture and reproduce it.

In this thesis, the following results of research are reported:

- In chapter 4, a comparison of two virtual headphone testing methods was conducted which concluded that their performances were statistically different. While this study found a correlation in the perceived spatial quality and general audio quality preference when using a *filtered* virtual test methodology, the same correlation could not be found when using a *recorded* test methodology.
- In chapter 5, mutual information is used to visualise the importance of frequencies along the spectra of ILD and IPD cues in terms of vertical localization. This analysis spans the range of human hearing and shows that localization cues above 7kHz diminish in the presence of noise.

- In chapter 6, an examination of binaural and spectral cues is performed in terms of their ability to localise both sound source azimuth and elevation in the matched and mismatched HRTF conditions. Combinations of these features are also examined and a novel feature set comprised of CCF,ILD, and GTCC cues is compared with other state of the art feature combinations using both deep neural networks and random forest algorithms. Deep neural networks trained using this feature set are shown to outperform these other methods in terms of vertical localization.
- In chapter 7, the adaptive affinity propagation clustering algorithm is introduced and used to cluster HRTF sets with similar performances in terms of vertical localization using the novel method proposed in chapter 6. HRTF clustering is shown to improve performance in the mismatched HRTF condition both in terms of azimuth and elevation localization and is shown to increase robustness of the algorithm in the presence of noise.
- Also shown in chapter 7 is an examination of anthropometric features and their ability to predict the cluster of a given HRTF set. Using a random forest algorithm, the importance of pinna features in determining HRTF cluster shows that more so than any other feature, the width of the cavum concha is the most significant predictor of HRTF cluster.

8.1 Summary

This thesis was structured in such a way as to provide the reader with an introduction to spatial audio before describing the ways in which the field has been furthered through the work presented in later chapters. In chapter 2, an explanation of spatial audio and its perception is given. Beginning with the human auditory system, the cues which allow us to determine the direction, distance, and trajectory of a sound source are explained. These include both interaural cues relating to both ears and spectral cues. A Head Related Coordinate System was introduced as well as the concept of the Head Related Transfer Function (HRTF) and its time domain equivalent, the Head Related Impulse Response (HRIR).

Chapter 3 describes the various ways in which spatial audio can be presented to listeners. Beginning with 2-channel stereophony and then moving onto multichannel stereophony, the chapter then provides a description of vector base amplitude panning. This is followed by an examination of digital sound projectors and how the image source method can be used to calibrate them. A novel method for calibrating digital sound projectors in rooms with uncertain geometries known as the PRobablistic Image Source Method (PRISM) is

given. An in depth explanation of the ambisonics sound format and how it can be used to play back spatial audio over a variety of different speaker arrays is given next. This is followed by a description of the TCD Audiosphere which was constructed by the author. This 16-channel, Third Order Ambisonics array played a key role in experimental work done on comparing ambisonic microphones which is detailed next. The chapter then concludes with an explanation of binaural reproduction over headphones.

Chapter 4 presented a novel study on virtual headphone testing. This work sought to resolve the disparity between two popular virtual headphone testing methods. These being the *filtered* methodology where the transfer function of each headphone under test is applied to each sound stimuli and the *recorded* methodology where the stimuli is recorded directly from each headphone onto a dummy head microphone. Through a series of MUSHRA listening tests it was concluded that there is a significant correlation between the perceived spatial quality of headphones and general audio quality using the *filtered* methodology. However, this same observation could not be made using the *recorded* methodology for on-ear headphones and thus it is concluded that this method is inconsistent with the more established *filtered* methodology.

Chapters 5 through 7 detail a series of studies on binaural sound source localization using machine learning algorithms. While chapter 5 mostly introduces the reader to the machine learning techniques used in subsequent chapters it also presents a visualisation of spectral importance in terms of vertical localization via mutual information and shows how frequencies above 7kHz are fundamental to this. In the presence of noise it is shown that these higher frequency spectral cues degrade. Chapter 6 examines specific binaural cues in terms of their ability to train machine learning algorithms to determine sound source localization. While interaural cues are shown to be good indicators for horizontal localization, spectral cues perform better at vertical localization. A new machine learning method which combines CCF, ILD, and GTCC cues is presented and compared with other state of the art methods. Using DNNs it is shown that this new feature set outperforms other methods in terms of vertical localization in the mismatched HRTF condition. In chapter 7, the adaptive affinity propagation algorithm is used to cluster HRTF sets based on their vertical localization performance used the proposed method. It is shown that HRTF clustering increases the accuracy of sound source localization in the mismatched HRTF condition and also increases the robustness in the presence of noise. Furthermore the specific clustering of HRTF sets found through this method is used alongside anthropometric features to determine the significance of pinna geometry in terms of HRTF clustering. Through the use of a random forest network, it is determined that the width of the cavum concha is the most significant pinna feature for determining HRTF clusters which are based on vertical localization. This finding is in agreement with many

physical models which replicate the reflection of sound sources off the cavum concha to create a sense of elevation in binaural audio.

8.2 Future Work

The work presented in this thesis has been experimental in nature. While providing a greater insight into the perception of spatial audio, potential practical applications for machine learning to interact with the field have been uncovered. While this thesis has demonstrated several interesting developments in the field, there is still further research to be done based on the results shown here. In this *future work* section, the potential developments based on the results of this thesis are discussed. This discussion is written under the heading of the chapter to which it relates.

8.2.1 Chapter 4

In this chapter, two methods of performing virtual headphone testing were compared. Performing a controlled, comparative double blind listening test on different headphones is logistically challenging. While headphone preference is an entirely subjective measure, a reliable means of performing such a measure is an invaluable tool to both headphone manufacturers and audio designers. While virtual headphone testing methods have offered an alternative to the conventional swap-and-compare method, they have not been widely adapted and are under explored at this current time.

The work in this thesis has shown that for subjective testing using a MUSHRA style listening test, there is a strong correlation between perceived spatial audio quality and general audio quality preference when using a *filtered* virtual test methodology. This is in agreement with findings by [70] which help to validate this particular method of virtual headphone testing. This conclusion bridges the gap between other studies which showed there is a strong correlation between a swap-and-compare type test and one which replicates the frequency response of different headphones [145].

While this work helps to validate one method of virtual headphone testing, it also helps to disregard another. Inconsistencies in results gathered using the [?] methodology show that it is unsuitable for the evaluation of spatial audio presented by on-ear headphones. This method creates a build up of low frequency energy and background noise as shown in spectrograms of the stimuli used during testing. Inconsistencies in headphone seating along with the time-consuming nature of this method make it an undesirable way to replicate the qualities of different headphones. This conclusion also vindicates the use of impulse response generated binaural audio which is the basis of the studies presented on machine learning and sound source localization in this thesis.

Since this study was first presented in 2016, more work has been done in the field of virtual headphone testing. Of particular note is a study published by Welti et al. in 2016 [199]. Here a listening test was employed comparing In-Ear (IE) headphones using two testing methods similar to the ones examined here. The Pearson correlation coefficient between the actual and virtualized headphone ratings was $r = 0.98$ indicating agreement. While this study presents evidence supporting the *recorded* methodology, the many differences between in-ear and on-ear headphones mean that they are relatively incomparable.

Another study by Olive et al. found that the frequency magnitude response of a headphone is a strong indicator for its overall perceived sound quality [143]. The agreement between the predicted and measured headphone preference ratings is $r = 0.91$ with an error of 5.5%. Both of these studies along with the one presented in this thesis present a compelling argument for virtualized audio stimuli to be used in comparative headphone testing. Doing so increases the efficiency, control of nuisance variables, and flexibility in both in-ear and on-ear headphone evaluations.

Future work that could be done to further the results of this chapter include the performance of further listening test using more subjects and more headphones. An argument could also be made against the use of a MUSHRA type listening test for virtual headphone testing for spatial audio where signals are compared with a hidden reference signal as suggested in [179]. It would be of great interest to compare the results of different styles of listening tests and their biases. This could help in the development of standardised virtual headphone test.

8.2.2 Chapters 6 and 7

These chapters present the most novel work performed in this thesis and inform the basis of many exciting possibilities in the areas of machine learning and binaural audio. A novel machine learning algorithm was presented which was capable of out-performing two state of the art methods of binaural sound source localization in the vertical plane in the mismatched HRTF condition. In examining these results it could be seen that this method allows for similar HRTF sets to be found based on their performance in the vertical plane when the algorithm is trained on one HRTF set and tested on another. Using this information it was then possible to group similar HRTF sets together using the adaptive affinity propagation algorithm as was shown in chapter 7. By training the algorithm with central HRTF sets from these clusters, a more robust algorithm could be developed capable of performing well in terms of localization in the mismatched HRTF condition and in the presence of noise. These unique HRTF clusters were then used as outputs to train a random

forest network using anthropometric data. This showed the significance of cavum concha width in terms of vertical localization.

In chapter 6, it was shown that GTCCs are a spectral cue which can capture our innate ability to localize sound on the vertical plane. While this result may not be particularly surprising as many physical models have already suggested this may be the case, it is important to explore the reasons why this may be the case. One such possibility could be that GTCC values from the contralateral ear are responsible as it was shown in Figure 2.12 that notch frequencies are preserved in the contralateral ear when a sound source is lateral to a listener. This would be relatively simple to verify by training a network to perform localization using only contralateral GTCC values.

While this work was conducted both in terms of azimuth and elevation localization, localization in terms of distance was not considered. This was due in part to a lack of a sufficient BRIR database for testing which could examine the performance of an algorithm in matched and mismatched room environments. Perhaps there is a binaural feature or combination of features which can be used to determine source distance in binaural audio from the feature set used in chapter 6. Of particular interest here would be the interaural cross correlation coefficient (IACC) and the approximation for apparent source width (ASW) as these cues are known to be categorically different in varying rooms.

While the anthropometric feature model trained in chapter 7 showed a significance between cavum concha width and HRTF cluster, the results were not exactly conclusive. A further examination between the HRTF and anthropometric features could further verify this result either by statistical analysis or by examining more HRTF data with accompanying anthropometric measurements such as those found in the SADIE II database [8].

The area of machine learning and sound source localization is growing rapidly with many new studies and methods appearing each year. It may be possible to improve upon the results of the algorithms presented in this thesis by tweaking parameters such as learning rate, batch, and hidden layer length. However this work revealed that HRTF clustering greatly improves the performance of machine learning algorithms in sound source localization in the mismatched HRTF condition. This robustness is a take away that can be applied to any novel method proposed in the future by this author or any other.

8.3 Closing Remarks

It is hoped that the results presented in this thesis will be of benefit to anyone working in the fields of spatial audio or audio signal processing. Throughout this work, a study on the validity of virtual headphone testing has been presented as well as a series of works on

machine learning and sound source localization. With regards to headphone testing, it is hoped that a standardized method of virtual headphone testing for spatial audio will be devised perhaps in part due to the work presented in this thesis. The area of sound source localization using machine learning is growing all the time with more and more research groups interested in the topic. While many studies use multiple microphone arrays to attain good performance, localization using binaural audio remains a difficult task. Development of a truly robust binaural sound source localization algorithm capable of operating in real time would have many benefits to the fields of virtual reality, augmented reality, and even in medical devices. Certainly such a model will incorporate HRTF clustering to achieve the most robust results possible.

Appendix A

Comparing Ambisonic Microphones

In this appendix, a series of studies investigating the quality of an array of Ambisonic microphones both in terms of their timbral quality and their directional encoding are presented. While the timbral quality is evaluated using subjective listening tests, the directional encoding of each B-format signal is examined using Directional Audio Coding (DirAC) methods [155]. The results of these studies were published in two separate academic papers at consecutive international conventions for the Audio Engineering Society (AES) [11, 13].

A.1 Ambisonic Microphones

In this section a brief description of each microphone studied is given, highlighting the uniqueness of each in terms of design. Five Ambisonic microphones were evaluated alongside a DPA 4006 omni-directional microphone which was used as a monophonic reference signal for subjective tests. The five microphones assessed were the Soundfield MKV, the Core Sound TetraMic, the MH Acoustics Eigenmike, and the Zoom H2n. All of which can be seen in Figure A.1. The Sennhesier Ambeo as seen in Figure A.2 was introduced. A description of each microphone is given as follows:

DPA 4006

The DPA 4006 is an omni-directional condenser microphone which has an extremely flat frequency response from 10Hz to 20kHz (± 2 dB). It was used in this study to capture monophonic reference recordings as it would supposedly provide the least colouration to the audio. These reference recordings were used in part 1 of this study to determine the timbral audio quality of the other microphones via subjective testing.



Fig. A.1 The Ambisonic microphones compared in Part 1 of the comparing Ambisonic microphones study (from left to right): DPA 4006, Core Sound TetraMic, Soundfield MKV, MH Acoustics Eigenmike, and Zoom H2n [11]



Fig. A.2 The Sennheiser Ambeo microphone which was included in Part 2 of the comparing Ambisonic microphones study [13]

Soundfield MKV

The Soundfield MKV is a tetrahedral microphone containing four sub-cardioid capsules mounted on the vertices of a tetrahedron. The MKV comes with a rack-mounted control unit which contains four calibrated pre-amplifiers that convert the raw A-format microphone signals into a line-level B-format output.

Core Sound TetraMic

The Core Sound TetraMic is a tetrahedral microphone consisting of four cardioid electret-condenser capsules. It was first released in 2007 as both a cheaper and smaller alternative to the Soundfield microphone. The TetraMic outputs a raw A-format signal for each of the four unbalanced microphones. These signals flow through a breakout cable that also

contains four Phantom Power Adaptors (PPAs). These A-format signals are converted to B-format signals using a customized version of the VVMic VST plugin developed by David McGriffy [126].

MH Acoustics Eigenmike

The Eigenmike developed by MH Acoustics is the first commercially-available microphone system that can produce HOA recordings. The spherical microphone array system is comprised of thirty-two 14mm omni-directional electret microphones mounted on an 8.4cm rigid sphere. The pre-amplifiers and 24-bit A/D converters are also contained inside the spherical enclosure of the microphone while the raw 32-channel signal is transmitted to the Eigenmike Microphone Interface Box (EMIB) via a Cat-5 cable. The EMIB produces a Firewire audio stream and appears as a 32-channel audio driver in the host computer. Using the designated EigenStudio application software [132], the 32-channel signal can be calibrated and rendered into both first and higher order Ambisonic signals.

Zoom H2n

The Zoom H2n Handy Recorder is a portable digital recorder which has five built-in microphones capsules and four different recording modes: X/Y, Mid-Side, 2-channel surround and 4-channel surround. These microphone capsules are arranged in two, opposite facing stereo pairs. The H2n is limited as it is unable to produce a vertical Z component. However, due to its low cost and portability, it is a versatile and useful solution for three-dimensional video capture and has been used in Google's JUMP system. For the purposes of this experiment, two stereo files were recorded using the Mid-Side (MS) stereo and 90 degree X/Y stereo modes recorded separately as two stereo files. These two stereo files were then converted to a B-format signal using a conversion matrix. This matrix is given as follows where the MS recording is regarded as the front signal and the XY recording is regarded as the back signal:

$$W = \frac{LF + RF + \sqrt{2}(LB + RB)}{2(\sqrt{2} + 2)} \quad (\text{A.1})$$

$$X = \frac{LF + RF - LB - RB}{\sqrt{2} + 2} \quad (\text{A.2})$$

$$Y = \frac{LF + RF}{2} \quad (\text{A.3})$$

This conversion matrix was only used in part 1 of this study. For part 2, the conversion was performed automatically using the native Ambisonics recording mode released by Zoom as a firmware update in 2016 [214].

Sennhesier Ambeo

Similar to the Soundfield MKV, the Sennheiser Ambeo microphone uses four cardioid capsules mounted on the vertices of a tetrahedron. The microphone outputs a raw A-format signal via a breakout cable. Using the Sennhesier conversion VST plugin, the A-format signal from the microphone is calibrated and converted to B-format. This conversion plugin includes an optional “Ambisonics Correction Filter”. As the default settings for the conversion plugin includes this filter, it was used in all the Ambeo recordings used in the subjective tests.

A.2 Subjective Listening Test

Two experiments were devised to assess the performance of a number of commercially available microphones (shown in Figure A.1) in capturing 360° audio. In line with previous studies which found a significant correlation between localization accuracy, sound quality and a listener’s overall preference [101], the two experiments were devised to evaluate the sound quality and localization accuracy of the different Ambisonic microphones. Experiment A consists of the subjective evaluation of high frequency timbre (Brightness/Dullness), and low frequency timbre (Fullness/Thinness) of B-format recordings of a musical performance. Experiment B consists of an objective directional encoding analysis, the details of which can be found in chapter 3 of this thesis.

A.2.1 Experiment A - Timbral Quality

For the listening test stimuli, a live performance by an acoustic quartet (flute, guitar, saxophone and cello) was recorded using multiple monophonic spot microphones (AKG 414s) and room microphones (AKG C314s, and Neumann TLMs). The quartet performed an original spatial music composition entitled A Round, Around, consisting of tonal, melodic canons, and rotations along with other spatial effects created by passing musical material between consecutive instruments [12]. The instruments and spot microphones were arranged in a circle of 1m radius and at 90° intervals (see Figure A.3). In order to ensure consistent reproduction and microphone position, four excerpts from this recording (samples 1,2,3 & 4) were then reproduced using a subset of a spherical loudspeaker array. B-format recordings were then taken with each of the five microphones under test,

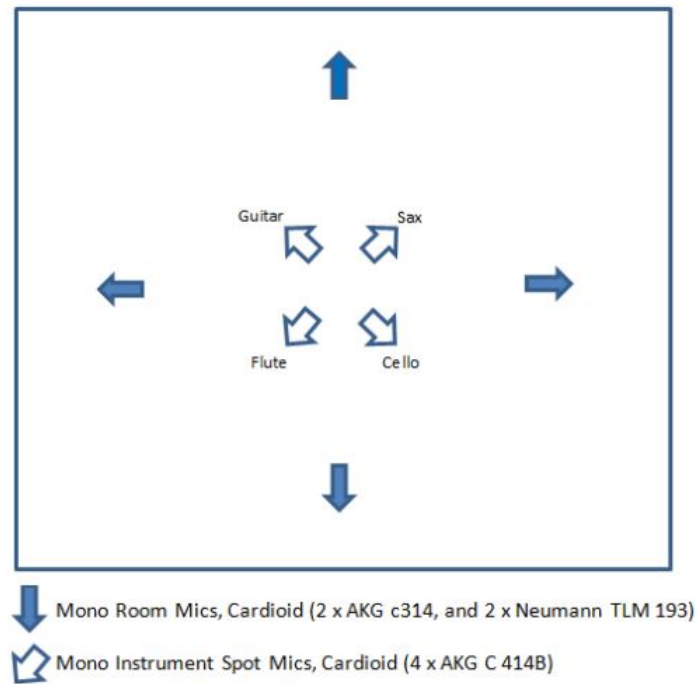


Fig. A.3 Quartet Recording Setup [13]

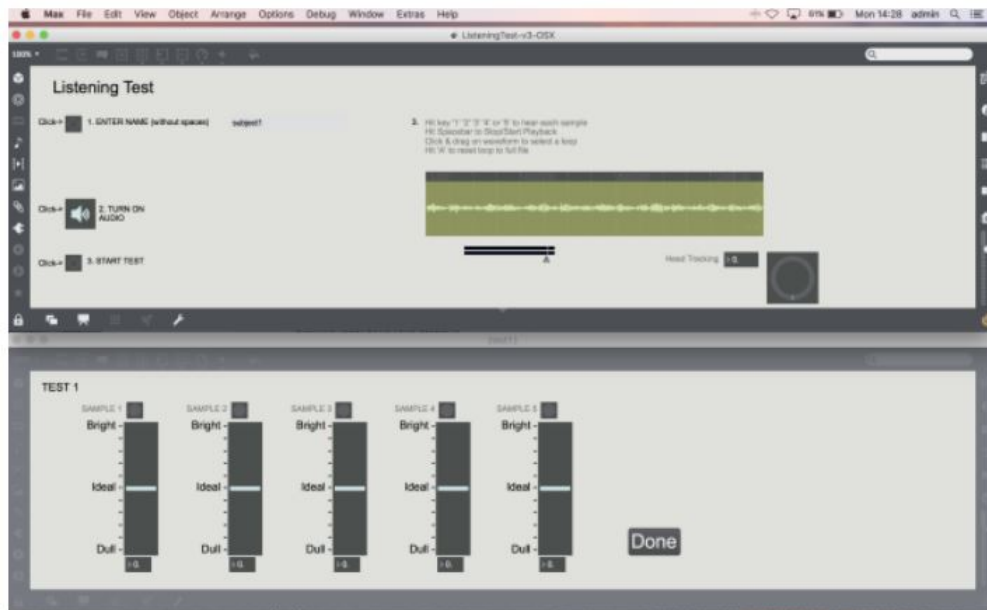


Fig. A.4 Listening Test User Interface [13]

using two tripod-mounted laser pointers to ensure that the same position was used for each recording. All recordings were made at 48kHz/24 bit. The Eigenmike and Zoom recordings were created using their own proprietary interfaces while the Soundfield MKV, Sennheiser Ambeo, and TetraMic microphones were recorded using a MOTU 8M USB

audio interface. The analogue inputs on the MOTU 8M can be digitally controlled in steps of 1dB, however, additional testing revealed variations in input gain of 0 to 0.44dB, which were compensated for prior to the A-to-B conversion. Each recording was converted to B-format using the hardware (Soundfield MKV) or software (EigenStudio for Eigenmike v2.2.5, VVmic for TetraMic v1.6, Sennheiser conversion plugin for Ambeo v1.03) provided by the manufacturer. The Zoom H2n was used in its native Ambisonics recording mode. The Furse-Malham weighting (with the W channel attenuated by 3dB) and channel order (WXYZ) was used throughout. Each stimulus was edited with a 512 sample fade in and out and loudness equalized to -23 dB LUFS as per the EBU R 128-2014 standard [47].

Method

A modified MUSHRA test was implemented to investigate the timbral quality of each microphone. The MUSHRA test procedure (described in ITU-R BS.1534-1) is typically used to evaluate intermediate levels of audio quality in multiple stimuli both through a comparison to a reference recording, and between each stimulus [172]. As no reference recording is available for this particular experiment, all test stimuli are presented to the listener at the same time without a reference. Due to the relative difficulty of the task, and the lack of a suitable reference, all test subjects undertook training prior to taking the test. The benefits of using trained, experienced listeners over naive or untrained listeners has been well documented with trained listeners tending to give more discriminating and consistent preference ratings [144, 169]. Training for this experiment was conducted using Harman's "How to Listen" listener training software [142]. Each subject was required to reach a skill level of 7 or higher in two specific programs focusing on the high frequency timbre (Brightness/Dullness), and low frequency timbre (Fullness/Thinness) of various stereo music samples.

The listening test stimuli consisted of binaural B-format decodes of five different microphone recordings of the same music sample, presented concurrently. Subjects were asked to rate the five recordings in terms of high frequency timbre (Bright/Dull) in tests 1 & 2, and low frequency timbre (Full/Thin) in tests 3 & 4. The same continuous quality scale as in the training program is used, namely from -5 (Dull), to 5 (Bright) in tests 1 & 2, and from -5 (Thin), to 5 (Full) in tests 3 & 4, with 0 (Ideal) as the initial, default value. The test system and user interface (see Figure A.4) was implemented in Max MSP and presented using a laptop, RME Babyface audio interface, Sennheiser HD650 open back headphones, and an Oculus Rift DK2 for head-tracking. As is common in VR and 360° video, the decoding is based around a cube of virtual loudspeakers and mode matched decoder, implemented using the ambiX decoder plugin [99] and the KEMAR HRTF set from the SADIE database [91]. Subjects could switch between different stimuli

in real-time, stop and start playback, and create customized playback loops as required. The order of test stimuli was randomized for each participant, and for each of the four tests. In addition, the initial azimuth orientation of stimuli was varied between each of the four tests. All of the listening tests were carried out in an acoustically treated room with a background noise level measured at less than 25dBA.

Subjects were asked to rate the stimuli in terms of either the high frequency timbre, or the low frequency timbre. 21 participants (20 male & 1 female) performed the test, with ages ranging from 21 to 50 years old, and an average age of 33. The participants included experienced listeners (12 subjects) comprising of professional audio engineers, and academics with prior experience of similar tests, and semi-experienced listeners (9 subjects) comprising of post-graduate students in the Music Technology programme at Trinity College Dublin. The average time taken for both the training and the test was 53 minutes, and a short break was always taken following the initial training and before the test.

Results & Analysis

The results of Experiment A are summarized in Figures A.5 and A.6. Experiment A followed a factorial design with factor A: sample (2 levels, sample 1 & 2, or sample 3 & 4), and factor B: microphone type (5 levels: Soundfield MKV, TetraMic, Eigenmike, Ambeo, and H2n). For Question 1 (high frequency timbre - Bright/Dull), a 2-way ANOVA analysis revealed a highly statistically significant effect of Factor B: microphone type ($F = 59.38$, $p\text{-value} < 0.001$), suggesting significant differences between microphones in terms of high frequency timbre. With respect to Factor A, no statistically significant differences were found between samples 1 & 2 ($F = 0.03$, $p\text{-value} = 0.859$), with no sample/microphone interaction effect ($F = 0.31$, $p\text{-value} = 0.8726$).

For Question 2 (low frequency timbre – Full/Thin), the ANOVA again revealed a highly statistically significant effect of Factor B: microphone type ($F = 50.5$, $p\text{-value} < 0.001$), suggesting significant differences between microphones in terms of low frequency timbre. For this question, a small but still statistically significant effect of Factor A: sample ($F = 5.52$, $p\text{-value} = 0.0198$) was found, with no sample/microphone interaction effect ($F = 1.66$, $p\text{-value} = 0.1605$).

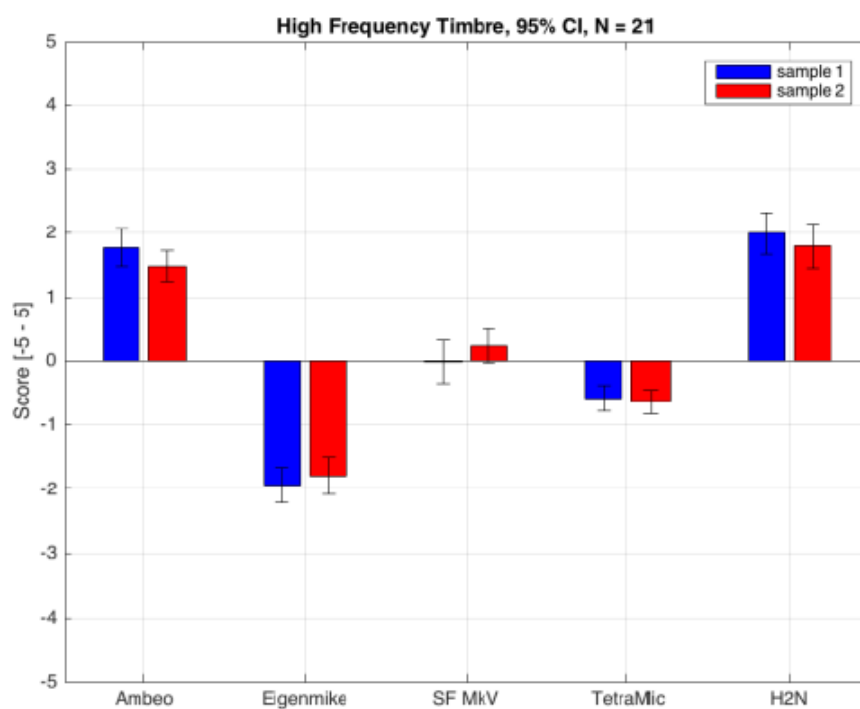


Fig. A.5 Listening Test Results: Q1 Bright/Dull [13]

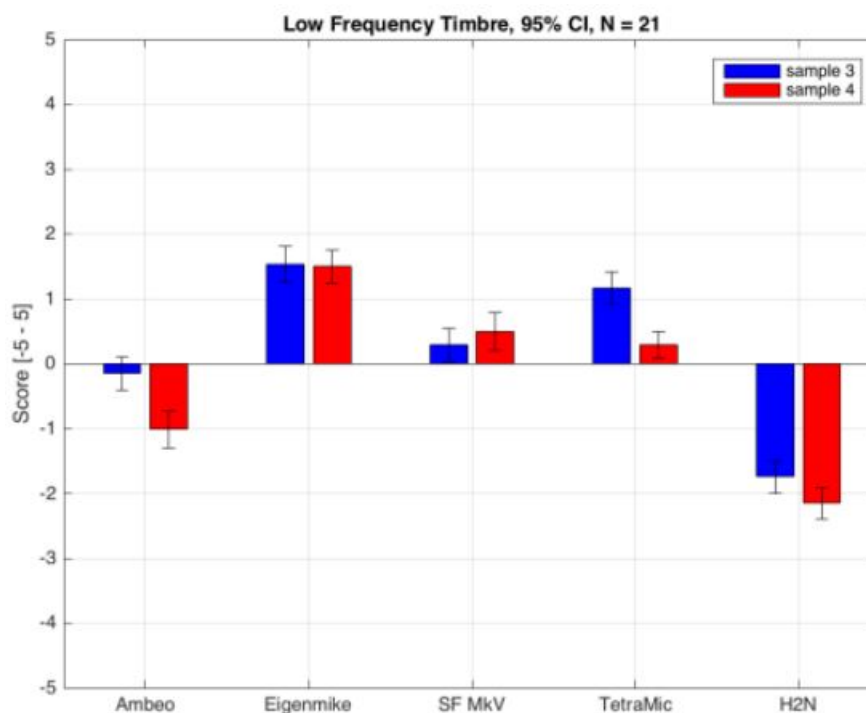


Fig. A.6 Listening Test Results: Q2 Full/Thin [13]

A.3 Conclusions From Comparing Ambisonic Microphones

The results from these subjective experiments are largely similar, with the Soundfield MKV once again producing the best results in terms of overall timbral quality, and with comparable results to the other microphones in terms of directionality, with the exception of the Eigenmike and Ambeo.

The Ambeo performed very well in terms of directionality, on par with the Eigenmike, and significantly better for elevation accuracy compared to both the Soundfield MKV, and TetraMic. In terms of timbral quality however, the Ambeo was evaluated as less ideal and brighter than either the Soundfield MKV or TetraMic. However, this is perhaps an unsurprising result given the noticeable high frequency boost applied in the correction filter used by default in the A-to-B-format conversion plugin. The worse performance of the TetraMic in terms of directionality, particularly elevation, is perhaps explained by the slightly greater inter-capsule spacing of this design compared to other models.

While the use of speech signals and studio recordings in part 1 of this study revealed some issues with noise due to the low signal level outputted by the TetraMic, this was much less evident in this real-world recording in a reverberant acoustic environment. This suggests that given appropriate pre-amplifiers with sufficient gain, excellent results can be achieved with this microphone in terms of timbral quality, albeit with slightly reduced localization accuracy. As might be expected, the more elaborate designs of the Eigenmike produced the best results overall in terms of directional accuracy, but with decreased performance in terms of timbre. As with part 1 of this study, this again suggests a trade-off between timbral quality and directionality, particularly when the number of individual capsules in the microphone is significantly increased. The Zoom H2n performed worse overall compared to all other microphones, however, its performance is still reasonable given its extremely low cost, and that it was not originally designed to produce B-format recordings.

References

- [1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.
- [2] Algazi, V. R., Avendano, C., and Duda, R. O. (2001a). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122.
- [3] Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., and Tang, Z. (2002). Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5):2053–2064.
- [4] Algazi, V. R., Duda, R. O., and Satarzadeh, P. (2007). Physical and filter pinna models based on anthropometry. In *Audio Engineering Society Convention 122*. Audio Engineering Society.
- [5] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001b). The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 99–102. IEEE.
- [6] Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.
- [7] Altoè, A., Pulkki, V., and Verhulst, S. (2014). Transmission line cochlear models: Improved accuracy and efficiency. *The Journal of the Acoustical Society of America*, 136(4):EL302–EL308.
- [8] 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. *Applied Sciences*, 8(11):2029.
- [9] Avendano, C., Algazi, V. R., and Duda, R. O. (1999). A head-and-torso model for low-frequency binaural elevation effects. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452)*, pages 179–182. IEEE.
- [10] Barbour, J. L. (2003). Elevation perception: Phantom images in the vertical hemisphere. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society.
- [11] Bates, E., Gorzel, M., O'Dwyer, H., Ferguson, L., and Boland, F. M. (2016). Comparing ambisonic microphones—part 1. In *Audio Engineering Society Convention 140*. Audio Engineering Society.

- [12] Bates, E. and Boland, F. (2016). Spatial music, virtual reality, and 360 media. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society.
- [13] Bates, E., Dooney, S., Gorzel, M., O'Dwyer, H., Ferguson, L., and Boland, F. M. (2017). Comparing ambisonic microphones—part 2. In *Audio Engineering Society Convention 142*. Audio Engineering Society.
- [14] Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 168(1011):158–180.
- [15] Begault, D. R., Wenzel, E. M., and Anderson, M. 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. *Journal of the Audio Engineering Society*, 49(10):904–916.
- [16] Benbow, A. A. and Anderson, P. L. (2019). A meta-analytic examination of attrition in virtual reality exposure therapy for anxiety disorders. *Journal of anxiety disorders*, 61:18–26.
- [17] Bertet, S., Daniel, J., and Moreau, S. (2006). 3d sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone. In *Audio Engineering Society Convention 120*. Audio Engineering Society.
- [18] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- [19] Blauert, J. (1974). *Räumliches Hören*. Hirzel Stuttgart.
- [20] Blauert, J. (1983). Spatial hearing mit press. *Cambridge, Massachusetts*.
- [21] Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT press.
- [22] Blauert, J. and Hearing, S. (1997). The psychophysics of human sound localization. In *Spatial Hearing*. MIT Press.
- [23] Bloom, P. J. (1977). Creating source elevation illusions by spectral manipulation. *Journal of the Audio Engineering Society*, 25(9):560–565.
- [24] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965.
- [25] Blumlein, A. D. and Patent, U. (1933). 394, 325. *Patent Issued June 14th*.
- [26] Boren, B., Geronazzo, M., Brinkmann, F., and Choueiri, E. (2015). Coloration metrics for headphone equalization. Georgia Institute of Technology.
- [27] Boren, B. B., Geronazzo, M., Majdak, P., and Choueiri, E. (2014). Phona: a public dataset of measured headphone transfer functions. In *Audio Engineering Society Convention 137*. Audio Engineering Society.
- [28] Brandstein, M. (2001). *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media.

- [29] Briolle, F. and Voinier, T. (1992). Transfer function and subjective quality of headphones: Part 2, subjective quality evaluations. In *Audio Engineering Society Conference: 11th International Conference: Test & Measurement*. Audio Engineering Society.
- [30] Bronkhorst, A. W. (2002). Modeling auditory distance perception in rooms. In *Proc. EAA Forum Acusticum Sevilla*.
- [31] Brungart, D. S., Durlach, N. I., and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. ii. localization of a broadband source. *The Journal of the Acoustical Society of America*, 106(4):1956–1968.
- [32] Cabrera, D. and Tilley, S. (2003). Parameters for auditory display of height and size. Georgia Institute of Technology.
- [33] Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H., and Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303.
- [34] Chakrabarty, S. and Habets, E. A. (2017). Broadband doa estimation using convolutional neural networks trained with noise signals. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 136–140. IEEE.
- [35] Cheng, C. I. and Wakefield, G. H. (1999). Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. In *Audio Engineering Society Convention 107*. Audio Engineering Society.
- [36] Chollet, F. et al. (2015). Keras.
- [37] Clark, W. W. and Ohlemiller, K. K. (2008). *Anatomy and physiology of hearing for audiologists*. Singular Publishing Group.
- [38] Craven, P. G. (2003). Continuous surround panning for 5-speaker reproduction. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society.
- [39] Craven, P. G. and Gerzon, M. A. (1977). Coincident microphone simulation covering three dimensional space and yielding various directional outputs. US Patent 4,042,779.
- [40] Csadi, S., Boland, F. M., Ferguson, L., O’Dwyer, H., and Bates, E. (2019). Direct to reverberant ratio measurements in small and mid-sized rooms. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society.
- [41] Davis, H. (1940). The clarification of certain phases of the physiology of hearing. *The Laryngoscope*, 50(8):747–755.
- [42] Defrance, G. and Polack, J.-D. (2008). Measuring the mixing time in auditoria. *Journal of the Acoustical Society of America*, 123(5):3499.
- [43] Deleforge, A., Forbes, F., and Horaud, R. (2015). Acoustic space learning for sound-source separation and localization on binaural manifolds. *International journal of neural systems*, 25(01):1440003.

- [44] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142.
- [45] Dimoulas, C., Kalliris, G., Avdelidis, K., and Papanikolaou, G. (2009). Improved localization of sound sources using multi-band processing of ambisonic components. In *Audio Engineering Society Convention 126*. Audio Engineering Society.
- [46] Duifhuis, H. (2004). Comment on “an approximate transfer function for the dual-resonance nonlinear filter model of auditory frequency selectivity”[j. acoust. soc. am. 114, 2112–2117](1). *The Journal of the Acoustical Society of America*, 115(5):1889–1890.
- [47] EBU-Recommendation, R. (2011). Loudness normalisation and permitted maximum level of audio signals.
- [48] Ege, R., Van Opstal, A. J., and Van Wanrooij, M. M. (2018). Accuracy-precision trade-off in human sound localisation. *Scientific reports*, 8(1):16399.
- [49] Ernst, O., Chazan, S. E., Gannot, S., and Goldberger, J. (2018). Speech dereverberation using fully convolutional networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 390–394. IEEE.
- [50] Eyring, C. F. (1930). Reverberation time in “dead” rooms. *The Journal of the Acoustical Society of America*, 1(2A):217–241.
- [51] Faddis, B. (2008). Structural and functional anatomy of the outer and middle ear. *Anatomy and Physiology of Hearing for Audiologists*, pages 93–108.
- [52] Fahy, F. J. and Salmon, V. (1990). Sound intensity.
- [53] Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*. Audio Engineering Society.
- [54] Fellgett, P. B. (1974). Ambisonic reproduction of directionality in surround-sound systems. *Nature*, 252(5484):534.
- [55] Ferguson, L., Bates, E., O’Dwyer, H., Csadi, S., and Boland, F. M. (2019). Calibration of digital sound projectors with scene uncertainty. In *Audio Engineering Society Convention 146*. Audio Engineering Society.
- [56] Fettiplace, R. and Hackney, C. M. (2006). The sensory and motor roles of auditory hair cells. *Nature reviews neuroscience*, 7(1):19.
- [57] Fletcher, H. (1953). Stereophonic recording and reproducing system. *Journal of the Society of Motion Picture and Television Engineers*, 61(3):355–363.
- [58] Frank, M., Zotter, F., and Sontacchi, A. (2015). Producing 3d audio in ambisonics. In *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology—Cinema, Television and the Internet*. Audio Engineering Society.
- [59] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814):972–976.

- [60] Gardner, B., Martin, K., et al. (1994). *HRFT Measurements of a KEMAR Dummy-head Microphone*, volume 280. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of
- [61] Garofalo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). The darpa timit acoustic-phonetic continuous speech corpus cdrom. *Linguistic Data Consortium*.
- [62] Genuit, K. (1983). Analytic description of average outer ear transfer functions in dependency on direction of sound incidence. In *Proc. 11th International Congress on Acoustics ICA, Paris, Lyon, Toulouse*, pages 9–12.
- [63] 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 *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4463–4467. IEEE.
- [64] Glasberg, B. R., Moore, B. C., and Nimmo-Smith, I. (1984). Comparison of auditory filter shapes derived with three different maskers. *The Journal of the Acoustical Society of America*, 75(2):536–544.
- [65] Google (2019 (accessed October 22nd, 2019)). Google cardboard. <https://arvr.google.com/cardboard/>.
- [66] GRAS (2007). Vibration. *Kemar manikin type 45ba product data*.
- [67] Gribben, C. and Lee, H. (2017). A comparison between horizontal and vertical interchannel decorrelation. *Applied Sciences*, 7(11):1202.
- [68] Gribben, C. and Lee, H. (2018). The frequency and loudspeaker-azimuth dependencies of vertical interchannel decorrelation on the vertical spread of an auditory image. *AES: Journal of the Audio Engineering Society*, 66(7-8):537–555.
- [69] Gulick, W. L., Gescheider, G. A., and Frisina, R. D. (1989). *Hearing: Physiological acoustics, neural coding, and psychoacoustics*. Oxford University Press.
- [70] Gutierrez-Parera, P. and Lopez, J. (2016). Influence of the quality of consumer headphones in the perception of spatial audio. *Applied Sciences*, 6(4):117.
- [71] Haneda, Y., Makino, S., and Kaneda, Y. (1994). Common acoustical pole and zero modeling of room transfer functions. *IEEE Transactions on Speech and Audio Processing*, 2(2):320–328.
- [72] Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., and Lorho, G. (2004). Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society*, 52(6):618–639.
- [73] Hartmann, W. M. (1999). How we localize sound. *Physics today*, 52:24–29.
- [74] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- [75] Heller, A., Lee, R., and Benjamin, E. (2008). Is my decoder ambisonic? In *Audio Engineering Society Convention 125*. Audio Engineering Society.

- [76] Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE.
- [77] Hidaka, T., Yamada, Y., and Nakagawa, T. (2007). A new definition of boundary point between early reflections and late reverberation in room impulse responses. *The Journal of the Acoustical Society of America*, 122(1):326–332.
- [78] Hirvonen, T., Vaalgamaa, M., Backman, J., and Karjalainen, M. (2003). Listening test methodology for headphone evaluation. In *Audio Engineering Society Convention 114*. Audio Engineering Society.
- [79] Hoffmann, P. F., Christensen, F., and Hammershøi, D. (2013). Insert earphone calibration for hear-through options. In *Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones*. Audio Engineering Society.
- [80] Hong, J. Y., Lam, B., Ong, Z.-T., Ooi, K., Gan, W.-S., Kang, J., Feng, J., and Tan, S.-T. (2019). Quality assessment of acoustic environment reproduction methods for cinematic virtual reality in soundscape applications. *Building and environment*, 149:1–14.
- [81] Hooley, T. (2006). Single box surround sound. *Acoustical science and technology*, 27(6):354–360.
- [82] Houtgast, T. and Aoki, S. (1994). Stimulus-onset dominance in the perception of binaural information. *Hearing research*, 72(1-2):29–36.
- [83] HTC (2019 (accessed October 22nd, 2019)). Htc vive. <https://www.vive.com/eu/>.
- [84] Iida, K., Ishii, Y., and Nishioka, S. (2014). Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener’s pinnae. *The Journal of the Acoustical Society of America*, 136(1):317–333.
- [85] ITU (2003). *ITU-R Rec. BS.1534-1: Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. International Telecommunications Union, Geneva, Switzerland.
- [86] Jager, I., Heusdens, R., and Gaubitch, N. D. (2016). Room geometry estimation from acoustic echoes using graph-based echo labeling. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- [87] Jain, P., Manweiler, J., and Roy Choudhury, R. (2015). Overlay: Practical mobile augmented reality. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 331–344.
- [88] Jeong, C.-H., Brunskog, J., and Jacobsen, F. (2010). Room acoustic transition time based on reflection overlap. *The Journal of the Acoustical Society of America*, 127(5):2733–2736.
- [89] Jeub, M., Schafer, M., and Vary, P. (2009). A binaural room impulse response database for the evaluation of dereverberation algorithms. In *2009 16th International Conference on Digital Signal Processing*, pages 1–5. IEEE.
- [90] Kabal, P. (2002). Tsp speech database. *McGill University, Database Version*, 1(0):09–02.

- [91] Kearney, G. and Doyle, T. (2015). An hrtf database for virtual loudspeaker rendering. In *Audio Engineering Society Convention 139*. Audio Engineering Society.
- [92] Kearney, G., Masterson, C., Gorzel, M., Rice, H., and Boland, F. (2011). Application of hrir factorization to game audio. In *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society.
- [93] Kendall, G. S. (1981). *The simulation of three-dimensional localization cues for headphone listening*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library.
- [94] Keyrouz, F. (2014). Advanced binaural sound localization in 3-d for humanoid robots. *IEEE Transactions on Instrumentation and Measurement*, 63(9):2098–2107.
- [95] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [96] Kistler, D. J. and Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, 91(3):1637–1647.
- [97] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.
- [98] Kolarik, A. J., Moore, B. C., Zahorik, P., Cirstea, S., and Pardhan, S. (2016). Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, 78(2):373–395.
- [99] Kronlachner, M. (2013). Ambisonics plug-in suite for production and performance usage. In *Linux Audio Conference*, pages 49–54. Citeseer.
- [100] Kulkarni, A. and Colburn, H. S. (2000). Variability in the characterization of the headphone transfer-function. *The Journal of the Acoustical Society of America*, 107(2):1071–1074.
- [101] Kurz, E., Pfahler, F., and Frank, M. (2015). Comparison of first-order ambisonics microphone arrays. In *3rd International Conference on Spatial Audio, ICSA*.
- [102] Larsen, E., Iyer, N., Lansing, C. R., and Feng, A. S. (2008). On the minimum audible difference in direct-to-reverberant energy ratio. *The Journal of the Acoustical Society of America*, 124(1):450–461.
- [103] Lartillot, O., Toivainen, P., and Eerola, T. (2008). A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*, pages 261–268. Springer.
- [104] Lee, H. and Lee, B.-H. (1988). An efficient algorithm for the image model technique. *Applied Acoustics*, 24(2):87–115.
- [105] Lee, K. (2013). Position-dependent crosstalk cancellation using space partitioning. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1228–1239.

- [106] Lee, W., Yang, X., Jung, H., Bok, I., Kim, C., Kwon, O., and You, H. (2018). Anthropometric analysis of 3d ear scans of koreans and caucasians for ear product design. *Ergonomics*, 61(11):1480–1495.
- [107] Lehmann, E. A. and Johansson, A. M. (2009). Diffuse reverberation model for efficient image-source simulation of room impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1429–1439.
- [108] Lehmann, E. A., Johansson, A. M., and Nordholm, S. (2007). Reverberation-time prediction method for room impulse responses simulated with the image-source model. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 159–162. IEEE.
- [109] Lindeman, R. W., Noma, H., and de Barros, P. G. (2007). Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–4. IEEE Computer Society.
- [110] Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (1999). The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654.
- [111] Liu, G. K. (2018). Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech. *arXiv preprint arXiv:1806.09010*.
- [112] Liu, J.-M., You, M., Li, G.-Z., Wang, Z., Xu, X., Qiu, Z., Xie, W., An, C., and Chen, S. (2013). Cough signal recognition with gammatone cepstral coefficients. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 160–164. IEEE.
- [113] Löllmann, H. W., Evers, C., Schmidt, A., Mellmann, H., Barfuss, H., Naylor, P. A., and Kellermann, W. (2018). The locata challenge data corpus for acoustic source localization and tracking. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 410–414. IEEE.
- [114] Lopez-Poveda, E. A. and Meddis, R. (1996). A physical model of sound diffraction and reflections in the human concha. *The Journal of the Acoustical Society of America*, 100(5):3248–3259.
- [115] Lovedee-Turner, M. and Murphy, D. (2018). Application of machine learning for the spatial analysis of binaural room impulse responses. *Applied Sciences*, 8(1):105.
- [116] Lui, J. T. and Hoy, M. Y. (2017). Evaluating the effect of virtual reality temporal bone simulation on mastoidectomy performance: a meta-analysis. *Otolaryngology–Head and Neck Surgery*, 156(6):1018–1024.
- [117] Lyon, R. F. (1998). Filter cascades as analogs of the cochlea. In *Neuromorphic systems engineering*, pages 3–18. Springer.
- [118] Lyon, R. F. (2017). *Human and machine hearing*. Cambridge University Press.
- [119] Ma, N., Brown, G. J., and May, T. (2015). Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. In *Sixteenth Annual Conference of the International Speech Communication Association*.

- [120] Ma, N., May, T., and Brown, G. J. (2017). Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(12):2444–2453.
- [121] Majdak, P., Balazs, P., and Laback, B. (2007). Multiple exponential sweep method for fast measurement of head-related transfer functions. *Journal of the Audio Engineering Society*, 55(7/8):623–637.
- [122] Malecki, P. (2014). Sound source localization accuracy of ambisonic microphone in anechoic condition. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 249, pages 1920–1924. Institute of Noise Control Engineering.
- [123] Malham, D. G. (1998). Spatial hearing mechanisms and sound reproduction. *University of York*.
- [124] Martens, W. L. and Marui, A. (2006). Spatial character and quality assessment of selected stereophonic image enhancements for headphone playback of popular music. In *Audio Engineering Society Convention 120*. Audio Engineering Society.
- [125] Mason, R., Brookes, T., and Rumsey, F. (2003). Creation and verification of a controlled experimental stimulus for investigating selected perceived spatial attributes. In *Audio Engineering Society Convention 114*. Audio Engineering Society.
- [126] McGriffy, D. (2012 (accessed August 27, 2019)). Vvtetrvst. <http://www.vvaudio.com/products/VVTetraVST>.
- [127] McKeag, A. and McGrath, D. S. (1996). Sound field format to binaural decoder with head tracking. In *Audio Engineering Society Convention 6r*. Audio Engineering Society.
- [128] MedicalSurgicalNursing (2019 (accessed September 26, 2019)). Sense of hearing. <https://sites.google.com/site/aryamsn/syllabus-communi-education/contact-me/unit---i-ent/1>.
- [129] Menzer, F., Faller, C., and Lissek, H. (2010). Obtaining binaural room impulse responses from b-format impulse responses using frequency-dependent coherence matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):396–405.
- [130] Merimaa, J. et al. (2006). *Analysis, synthesis, and perception of spatial sound: binaural localization modeling and multichannel loudspeaker reproduction*. Helsinki University of Technology.
- [131] Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., and Virtanen, T. (2017). Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.
- [132] MHAcoustics (2014 (accessed August 27, 2019)). Eigenstudio. <https://mhacoustics.com/>.
- [133] Middlebrooks, J. C. (2015). Sound localization. In *Handbook of clinical neurology*, volume 129, pages 99–116. Elsevier.

- [134] Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321.
- [135] Moore, B. C. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The journal of the acoustical society of America*, 74(3):750–753.
- [136] Moreau, S., Daniel, J., and Bertet, S. (2006). 3d sound field recording with higher order ambisonics—objective measurements and validation of a 4th order spherical microphone. In *120th Convention of the AES*, pages 20–23.
- [137] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- [138] Nakashima, H. and Mukai, T. (2005). 3d sound source localization system based on learning of binaural hearing. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3534–3539. IEEE.
- [139] Neidhardt, A., Ignatious-Tommy, A., and Pereppadan, A. D. (2018). Plausibility of an interactive approaching motion towards a virtual sound source based on simplified binaural sets. In *Audio Engineering Society Convention 144*. Audio Engineering Society.
- [140] Nielsen, S. H. (1992). Auditory distance perception in different rooms. In *Audio Engineering Society Convention 92*. Audio Engineering Society.
- [141] Oculus (2019 (accessed October 22nd, 2019)). Oculus rift. <https://www.oculus.com/>.
- [142] Olive, S. (2001). A new listener training software application. In *Audio Engineering Society Convention 110*. Audio Engineering Society.
- [143] Olive, S., Welti, T., and Khonsaripour, O. (2017). A statistical model that predicts listeners’ preference ratings of in-ear headphones: Part 2—development and validation of the model. In *Audio Engineering Society Convention 143*. Audio Engineering Society.
- [144] Olive, S. E. (2003). Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study. *Journal of the Audio Engineering Society*, 51(9):806–825.
- [145] Olive, S. E., Welti, T., and McMullin, E. (2013). A virtual headphone listening test methodology. In *Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones*. Audio Engineering Society.
- [146] Opoichinsky, R., Laufer-Goldshtein, B., Gannot, S., and Chechik, G. (2019). Deep ranking-based sound source localization. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 283–287. IEEE.
- [147] Otani, M. and Hirahara, T. (2008). Auditory artifacts due to switching head-related transfer functions of a dynamic virtual auditory display. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 91(6):1320–1328.
- [148] O’Dwyer, H., Gorzel, M., Ferguson, L., Bates, E., and Boland, F. M. (2016). Virtual headphone testing for spatial audio. *Interactive Audio Systems Symposium 2016*.

- [149] Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). Complex sounds and auditory images. In *Auditory physiology and perception*, pages 429–446. Elsevier.
- [150] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.
- [151] Perrett, S. and Noble, W. (1997). The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4):2325–2332.
- [152] Poveda, E. A. L. (1996). *The physical origin and physiological coding of pinna-based spectral cues*. PhD thesis, Loughborough University.
- [153] Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology*, 13(3):278.
- [154] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, 45(6):456–466.
- [155] Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516.
- [156] Pulkki, V., Laitinen, M.-V., and Sivonen, V. (2010). Hrtf measurements with a continuously moving loudspeaker and swept sines. In *Audio Engineering Society Convention 128*. Audio Engineering Society.
- [157] Rämö, J. and Välimäki, V. (2012). Signal processing framework for virtual headphone listening tests in a noisy environment. In *Audio Engineering Society Convention 132*. Audio Engineering Society.
- [158] Raykar, V. C., Duraiswami, R., and Yegnanarayana, B. (2005). Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *The Journal of the Acoustical Society of America*, 118(1):364–374.
- [159] Rayleigh, L. (1907). Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232.
- [160] Roffler, S. K. and Butler, R. A. (1968a). Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America*, 43(6):1255–1259.
- [161] Roffler, S. K. and Butler, R. A. (1968b). Localization of tonal stimuli in the vertical plane. *The Journal of the Acoustical Society of America*, 43(6):1260–1266.
- [162] Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, 50(9):651–666.
- [163] Sabine, W. C. (1927). *Collected papers on acoustics*. Harvard University Press Cambridge, MA.

- [164] Samarasinghe, P. N., Abhayapala, T. D., Lu, Y., Chen, H., and Dickins, G. (2018). Spherical harmonics based generalized image source method for simulating room acoustics. *The Journal of the Acoustical Society of America*, 144(3):1381–1391.
- [165] Samsung (2019 (accessed October 22nd, 2019)). Samsung gear vr. <https://www.samsung.com/global/galaxy/gear-vr/>.
- [166] Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., and Verhulst, S. (2016). A comparative study of seven human cochlear filter models. *The Journal of the Acoustical Society of America*, 140(3):1618–1634.
- [167] Sato, S.-i. and Ando, Y. (2002). Apparent source width (asw) of complex noises in relation to the interaural cross-correlation function. *J. Temporal Des. Arch. Environ*, 2:29–32.
- [168] Satongar, D., Pike, C., Lam, Y. W., Tew, A. I., et al. (2015). The influence of headphones on the localization of external loudspeaker sources. *Journal of the Audio Engineering Society*, 63(10):799–810.
- [169] Schinkel-Bielefeld, N., Lotze, N., and Nagel, F. (2013). Audio quality evaluation by experienced and inexperienced listeners. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 060016. ASA.
- [170] ScienceABC (2019 (accessed October 1st, 2019)). Basilar membrane: What is it? what are its functions? <https://www.scienceabc.com/humans/basilar-membrane-what-is-it-and-what-does-it-do.html>.
- [171] Series, B. (2011). Itu-r bs.2159-4: Multichannel sound technology in home and broadcasting applications.
- [172] Series, B. (2014). Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*.
- [173] Shahwani, H., Bui, T. D., Jeong, J. P., and Shin, J. (2017). A stable clustering algorithm based on affinity propagation for vanets. In *2017 19th International Conference on Advanced Communication Technology (ICACT)*, pages 501–504.
- [174] Shaw, E. A. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America*, 56(6):1848–1861.
- [175] Shaw, E. A. and Teranishi, R. (1968). Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *The Journal of the Acoustical Society of America*, 44(1):240–249.
- [176] Shinn-Cunningham, B. G. (2000). Distance cues for virtual auditory space. In *Proceedings of the IEEE-PCM*, volume 2000, pages 227–230. Citeseer.
- [177] Shirley, J. W. (1951). An early experimental determination of snell’s law. *American Journal of Physics*, 19(9):507–508.
- [178] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

- [179] Silzle, A., George, S., Habets, E. A., and Bachmann, T. (2011). Investigation on the quality of 3d sound reproduction. *Proceedings of ICSA*, page 334.
- [180] Sony (2019 (accessed October 22nd, 2019)). Playstation vr. <https://www.playstation.com/en-ie/explore/playstation-vr/>.
- [181] Stevens, S. S., Volkman, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.
- [182] Sunder, K. and Gan, W.-S. (2016). Individualization of head-related transfer functions in the median plane using frontal projection headphones. *Journal of the Audio Engineering Society*, 64(12):1026–1041.
- [183] Svensson, U. P., Fred, R. I., and Vanderkooy, J. (1999). An analytic secondary source model of edge diffraction impulse responses. *The Journal of the Acoustical Society of America*, 106(5):2331–2344.
- [184] Temme, S., Olive, S., Tatarunis, S., Welte, T., and McMullin, E. (2014). The correlation between distortion audibility and listener preference in headphones. In *Audio Engineering Society Convention 137*. Audio Engineering Society.
- [185] Tervo, S. (2009). Direction estimation based on sound intensity vectors. In *2009 17th European Signal Processing Conference*, pages 700–704. IEEE.
- [186] Valero, X. and Alias, F. (2012). Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Transactions on Multimedia*, 14(6):1684–1689.
- [187] Valimaki, V., Franck, A., Ramo, J., Gamper, H., and Savioja, L. (2015). Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments. *IEEE Signal Processing Magazine*, 32(2):92–99.
- [188] Vecchiotti, P., Ma, N., Squartini, S., and Brown, G. J. (2019). End-to-end binaural sound localisation from the raw waveform. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 451–455. IEEE.
- [189] Vilkamo, J. (2008). Spatial sound reproduction with frequency band processing of b-format audio signals. *Helsinki University of Technology*.
- [190] Vilkamo, J., Lokki, T., and Pulkki, V. (2009). Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *Journal of the Audio Engineering Society*, 57(9):709–724.
- [191] Voxengo (2016 (accessed September 9th, 2019)). Voxengo deconvolver. <https://www.voxengo.com/product/deconvolver/>.
- [192] Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339.
- [193] Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1949). A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, 21(4):468–468.

- [194] Wang, J., Wang, J., Qian, K., Xie, X., and Kuang, J. (2020). Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched hrtf condition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1):4.
- [195] Wang, K. (2009). Adaptive Affinity Propagation clustering. <https://www.mathworks.com/matlabcentral/fileexchange/18244-adaptive-affinity-propagation-clustering/>. [Online; accessed 16-February-2021].
- [196] Wang, K., Zhang, J., Li, D., Zhang, X., and Guo, T. (2008). Adaptive affinity propagation clustering.
- [197] Watkins, A. J. (1978). Psychoacoustical aspects of synthesized vertical locale cues. *The Journal of the Acoustical Society of America*, 63(4):1152–1165.
- [198] Welti, T. (2015). Improved measurement of leakage effects for circum-aural and supra-aural headphones. In *Audio Engineering Society Convention 138*. Audio Engineering Society.
- [199] Welti, T., Olive, S., and Khonsaripour, O. (2016). Validation of a virtual in-ear headphone listening test method. In *Audio Engineering Society Convention 141*. Audio Engineering Society.
- [200] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123.
- [201] Werner, S. and Liebetrau, J. (2014). Adjustment of direct-to-reverberant-energy-ratio and the just-noticeable-difference. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.
- [202] Wightman, F. L. and Kistler, D. J. (1997). Monaural sound localization revisited. *The Journal of the Acoustical Society of America*, 101(2):1050–1063.
- [203] Woodworth, R. S. and Schlosberg, H. (1954). *Experimental psychology*. Oxford and IBH Publishing.
- [204] Wu, X., Talagala, D. S., Zhang, W., and Abhayapala, T. D. (2019). Individualized interaural feature learning and personalized binaural localization model. *Applied Sciences*, 9(13):2682.
- [205] Xiang Zhang, Jie Gao, Ping Lu, and Yonghong Yan (2008). A novel speaker clustering algorithm via supervised affinity propagation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4369–4372.
- [206] Xie, B. (2013). *Head-related transfer function and virtual auditory display*. J. Ross Publishing.
- [207] Xu, Y., Afshar, S., Singh, R. K., Hamilton, T. J., Wang, R., and van Schaik, A. (2018a). A machine hearing system for binaural sound localization based on instantaneous correlation. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

- [208] Xu, Y., Afshar, S., Singh, R. K., Wang, R., van Schaik, A., and Hamilton, T. J. (2019). A binaural sound localization system using deep convolutional neural networks. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.
- [209] Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., and van Schaik, A. (2018b). A fpga implementation of the car-fac cochlear model. *Frontiers in neuroscience*, 12:198.
- [210] Young, P. T. (1931). The rôle of head movements in auditory localization. *Journal of experimental psychology*, 14(2):95.
- [211] Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846.
- [212] Zhang, Y., Liu, G., and Luo, B. (2014). Finite-time cascaded tracking control approach for mobile robots. *Information Sciences*, 284:31–43.
- [213] Zhao, X., Shao, Y., and Wang, D. (2012). Casa-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1608–1616.
- [214] Zoom (2016 (accessed August 27, 2019)). Zoom ambisonics conversion plugin. <https://www.zoom-na.com>.
- [215] Zotkin, D. N., Duraiswami, R., Grassi, E., and Gumerov, N. A. (2006). Fast head-related transfer function measurement via reciprocity. *The Journal of the Acoustical Society of America*, 120(4):2202–2215.
- [216] Zotter, F. and Frank, M. (2013). Efficient phantom source widening. *Archives of Acoustics*, 38.
- [217] Zweig, G., Lipes, R., and Pierce, J. (1976). The cochlear compromise. *The Journal of the Acoustical Society of America*, 59(4):975–982.
- [218] Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248.

