3D audio reproduction using frontal projection headphones

Kaushik Sunder

2016

Kaushik Sunder. (2016). 3D audio reproduction using frontal projection headphones.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/66303>

<https://doi.org/10.32657/10356/66303>

957
5A
D/M/11

3D AUDIO REPRODUCTION USING FRONTAL PROJECTION HEADPHONES

KAUSHIK SUNDER

SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2015

Acknowledgments

I would like to first thank my supervisor Assoc. Prof. Dr. Woon-Seng Gan for all the support he has given in the last 4 years. Dr. Gan has always been encouraging me to do better at every stage of my research. He has often pushed me to my limits so that I can better myself every single time. His never ending passion for research is a quality that I have deeply imbibed from him. Dr. Gan has indeed been a big source of inspiration and I am indeed lucky to get an opportunity to work under him.

I would also like to thank my post-doc Dr. Tan Ee-Leng, who has again been a true source of inspiration in my work. Dr. Ee-Leng's ability to tackle any issue be it technical or otherwise, is exemplary. I have had many hours of discussions on a daily basis with Dr. Ee-Leng on various topics of 3D audio. The development of 3D audio headphone in this thesis has been a result of one such discussion with him. On a personal level too, Dr. Ee-Leng has taught me several things, which has helped a great deal in coping up with the PhD.

I owe a special thanks to our lab technician Mr. Yeo. I could go to him anytime and have a discussion on any topic under the sun, and he would time and again prove that he knows literally about everything. I have had discussions with him on topics ranging from music to politics to sports. The kind of technical support, he has provided me, made my journey through my PhD a lot more easier. The running skills that Mr. Yeo imparted in me has been my biggest stress buster during my PhD days. I have been extremely fortunate to have got a great lab atmosphere who are good friends as well as colleagues.

Acknowledgments

I would like to extend my sincere thanks to my other labmates: Kushal, Abhishek Seth, Dileep, Rueben, Tong Wei, Shi Chuang, Ko Ko, Mu Hao, Du Hai, Lam Bhan, Yi Yang, Santi, Stefano, Tatsuya, Valentin and JianJun. I have worked together with Jianjun in the 3D headphones project and owe him a lot of credit to the success of this project. Jianjun also co-authored and contributed equally to the paper on natural sound rendering, which was published in the IEEE Signal processing magazine.

Words are not enough to thank my best labmates Rishabh and Apoorv. My work would have been incomplete if not for these two amazing colleagues. In my most difficult times, they held my hands and supported me in my most difficult days. There have been several days when they have helped me with my experiments tirelessly leaving their sleep aside. I owe a lot to them and will definitely miss the amazing moments I spent working together.

I would like to thank Mr. Mark Jay, an experienced audio researcher and moderator of the perceptual audio group in LinkedIn, for his unstinted support. Mark has been my go-to man for any doubts related to audio, and he would patiently explain me each and every point. An instance which I can never forget is when Mark retrieved his 10 year old notes to show me some equalization curve. I have never seen such a dedicated audio researcher. Needless to say, I have immensely from Mark and a huge thanks to him. I would also like to thank Dr. Durand Begault who is currently a researcher and scientist at NASA Ames. Dr. Begault's book on 3D audio was an eye opener for me. Dr. Begault has been an inspiration in this field and I would like to express my gratitude towards him for everything.

It is easy to get lost in your own world while doing a PhD. My amazing friends both in Singapore as well as in India have helped and supported me unconditionally. I have received a lot of love from friends in NITK. I would like to first thank the 6A group (Chaitu, Praveen, Saurabh, Sharath, Kaushik Jayaram, Shreyas, Sagar, Sangameshwar,

Acknowledgments

Akash, Adithya KD, Niranjana) for all their support right from National Institute of Technology Karnataka (NITK), where from I graduated. My biggest support system back in India has been my adopted sisters, Teju, Mahima, Revathi, Shruthi, Pallavi, lakshmi, Ranjitha, Monal, Poonam, Amrita. I have received more love from them than I could ever imagine. And a big thank you to my amazing nitk juniors: Anusha, Hema, Tejovanth, Abhinav KR, Dhruv, Tushar, Sachin, Sandeepa, Sreevatsa, Pranav, Dayasagar.

I would like to thank Dr. Laxmi (Lacchu) and soon to be Dr. Lavi for all the love and care they have showered upon me during the last 4 years. Words cannot describe how much they supported me. I would like to thank my amazing lunch group: Abhishek, Raja, Sumedh, Handiru, Sheetal, Areetha, Saurabh, Supriya, Nishanth and Nikitha for putting up with me and for the most enjoyable lunch time moments all these years. Special thanks to Nikitha and Soumyabrata, who helped me port my thesis into LaTeX. I would like to thank my friend Karthik, from whom I have learnt so much about music and audio. My perspective of sounds and music has definitely changed for good ever since.

Nothing could have been possible without the wishes and support of my parents and other members of the family. Whatever i achieve today and will ever achieve in future is because of my parents Sunder and Sudha, and my dear brother Keshav. I have received unconditional love and support from everything throughout the life from them. I am truly grateful for everything. Finally, I would like dedicate this thesis to my grandmother Kalyani, who has been watching over me and blessing me in everything.

Table of Contents

1	Introduction	1
1.1	Spatial Hearing	1
1.2	3D Audio Technologies	5
1.2.1	Surround Sound Systems	6
1.2.2	Sound field Technology	7
1.2.3	Binaural Technology	9
1.3	Motivation of the Thesis: Binaural audio over headphones	12
1.4	Main contribution of Thesis	13
1.5	Applications of 3D audio	15
1.6	Coordinate System	16
1.7	Overview of Thesis	17
2	Background	21
2.1	Head Related Transfer Functions	21
2.2	Binaural Synthesis/Recording	24
2.2.1	Binaural recording of the sound scene	25
2.2.2	Creating an artificial sound scene by binaural synthesis	27
2.3	Measurement of HRTFs	27
2.4	Directional localization cues	30
2.4.1	Interaural cues	30
2.4.2	Spectral cues for localization: Role of pinnae	36
2.5	Binaural playback over headphones	43
2.6	Headphone equalization	43
2.6.1	Non-decoupled equalization technique	45
2.6.2	Decoupled equalization technique	47
2.7	Circumaural headphones : A theoretical model	51
2.8	Open headphones	54

TABLE OF CONTENTS

2.9 Drawbacks of using non-individualized HRTFs 55

2.10 Summary 59

3 Measurement of Head Related Transfer Functions 60

3.1 System identification methods 61

3.2 Measurements of HRTFs 63

3.3 Excitation signal 65

3.3.1 Exponential sine sweep 67

3.4 Measurement/Recording point 73

3.5 Near-field HRTFs : Measurement constraints 75

3.6 Measurement procedure 76

3.7 Calibration 79

3.8 Minimum phase representation of HRTF 79

3.9 Creation of HRTF database 83

3.10 Summary 84

4 Individualization of Non-Individualized Binaural Synthesis using Frontal Projection Headphones 85

4.1 Individualization techniques available in literature 86

4.1.1 Individualization using ITD 87

4.1.2 Individualization of spectral cues 90

4.2 Individualization using Frontal Projection Headphones 93

4.3 Frontal projection headphones : Design 94

4.3.1 Transmission in free-field 95

4.3.2 Frontal Projection headphone listening 98

4.3.3 How the frontal projection headphones provide 3D cues 102

4.4 Measurement of headphones transfer functions (HPTF) 103

4.5 Equalization of headphones: Type-1 and Type-2 105

4.5.1 Type-1 equalization (conventional headphones) 107

4.5.2 Type-2 Equalization (frontal emitter headphones) 108

4.6 Objective assessment 112

4.6.1 Comparison of HPTFs with the measured HRTFs 113

4.7 Subjective experiments 116

4.7.1 Effect of equalization on frontal projection headphone playback . 116

4.7.2 Results 117

4.8 Localization experiment 120

TABLE OF CONTENTS

4.8.1	Results	121
4.9	Superposition experiment	122
4.9.1	Results	125
4.10	Additional processing blocks	125
4.10.1	Decorrelation	127
4.10.2	Front-back biasing	127
4.11	Subjective test: Additional processing blocks	128
4.11.1	Results	129
4.12	Summary	131
5	Effect of headphone equalization on distance perception	133
5.1	Introduction	134
5.2	Measurements of HPTFs and equalization	135
5.3	Subjective experiments	137
5.4	Experiments : Phase 1	139
5.5	Experiments : Phase 2	140
5.6	Effect of headphone repositioning	142
5.7	Analysis of variance	143
5.8	Discussion	145
6	Modeling distance-dependent individualized HRTF using frontal projection headphones	148
6.1	Literature survey	149
6.1.1	Auditory cues for distance perception	150
6.1.2	Models of distance-dependent HRTFs	151
6.2	Role of Auditory Parallax and ILD Cues in distance perception	155
6.2.1	Subjective experiments	160
6.2.2	Results	163
6.3	Modeling distance-dependent individualized HRTFs	170
6.3.1	Objective analysis	181
6.3.2	Subjective experiments	184
6.4	Summary	191

TABLE OF CONTENTS

7 Individualization of head-related transfer functions in the median plane using frontal projection headphones 195

7.1 Spectral cues in the sagittal plane 196

7.2 Models for simulating elevation effects 196

7.3 Measurements of HRTFs and HPTFs 201

7.4 Estimation of the position of the notches $N1$, $N2$ 201

7.5 Modeling HRTFs in the median plane using frontal projection headphones 207

7.6 Objective analysis 211

7.7 Subjective analysis 216

7.8 Summary 220

8 Natural sound rendering for digital media content using 3D audio headphones: An integration of signal processing techniques 222

8.1 The need for natural sound rendering 223

8.2 Challenges 224

8.3 Signal processing techniques 225

8.4 Sound source decomposition using BSS and PAE 230

8.4.1 Decomposition using BSS 231

8.4.2 Decomposition using PAE 233

8.4.3 A comparison between BSS and PAE 236

8.5 Individualization of HRTFs 238

8.6 Equalization 239

8.7 Integration of natural sound rendering techniques 239

8.8 3D Audio headphones 240

8.9 Subjective experiments 241

8.10 Conclusions 246

9 Conclusions and Future Works 248

9.1 Conclusions 248

9.2 Future works 255

Author's Publications 257

Bibliography 258

Appendices 292

TABLE OF CONTENTS

A	Spatial oriented format for acoustics (SOFA)	293
A.1	Objects	295
A.2	SOFA Conventions	295
B	Range dependent spherical head model	299
C	Inverse filtering	304
D	Theoretical model for headphones	306
D.1	Pressure-chamber principle	306
D.2	Analogies between different domains :	307
D.3	Components of headphones :	308
D.4	Model of a headphone-earcanal system :	311
D.5	Analysis of leakages in headphones :	314
D.6	Closed Vs Open-back designs :	316

List of Figures

1.1	Rendering a virtual source environment very close to the real environment. The top panel shows the person in a museum listening to the virtual sounds of a waterfall. When the rendering is realistic, the listener feels that he is teleported to the real waterfall as in the bottom panel.	3
1.2	3D Audio technologies can be categorized generically in terms of Surround sound and 3D sound technology.	5
1.3	Crosstalk in Binaural audio playback using loudspeakers.	11
1.4	Coordinate system followed in this thesis	17
1.5	Overview of the thesis	19
2.1	HRTF of KEMAR dummy head for an azimuth angle of 30° in the horizontal plane.	22
2.2	Binaural synthesis and playback over headphones.	24
2.3	a) B & K 4128 HATS dummy head [1]; (b) KEMAR by G.R.A.S [2]; (c) KU-100 by Neumann [3]; (d) HMSIV by Head Acoustics [4]; (e) MK2B by 01 dB-metravib [5]; (f) VALDEMAR [6].	26
2.4	ITD model (Equation 2.8)	34
2.5	ILD model (Equation 2.13)	35
2.6	Non-decoupled mode of equalization. BIR denotes either a HRIR or a binaural recording. ‘M’ denotes the contribution of the microphone and ear canal to the measurement that does not depend on the direction of incidence. Note that * represents convolution and % represents deconvolution. This figure has been adapted from [7].	46

LIST OF FIGURES

2.7 Decoupled Equalization. The recording and the reproduction systems are equalized separately with a common reference field. Thus, it allows elimination of the contribution of the transducers from the recording system. This method is a common choice when the measurement/recording equipments are different from the listening session. The different types of reference fields used are given in Table 2.2. This figure has been adapted from [7]. 48

2.8 Examples of closed-back, open-back and the open headphones. 52

2.9 Simple model for an open-back headphone. Figure adapted from [302]. . 53

2.10 Acoustic circuit analogue for an open-back headphone. 53

3.1 HRTF measurement system employed in this study. 64

3.2 A basic input/output system. 68

3.3 A more complete representation of the system. The loudspeaker contributes to the non-linear effects and the propagation in the air represents a linear system. 69

3.4 An example of an exponential sine sweep. Parameters : $Fs = 44,100\text{ Hz}$, Start Frequency = 10 Hz , Final frequency = $1,000\text{ Hz}$, Duration = 1 s . . 70

3.5 Magnitude Spectrum of the exponential sine sweep. Parameters : $Fs = 44,100\text{ Hz}$, Start Frequency = 10 Hz , Final Frequency = $1,000\text{ Hz}$, Duration = 1 s 71

3.6 Magnitude Spectrum of the inverse exponential sine sweep. Parameters : $Fs = 44100\text{ Hz}$, Start Frequency = 10 Hz , Final Frequency = 1000 Hz , Duration = 1 s 73

3.7 Setup for the measurement of near-field HRTFs in the anechoic chamber. The speaker used is a Cambridge Soundworks speaker and the dummy head is the B & K HATS dummy head. 77

3.8 Free-field response of the speaker for various distances. 78

3.9 Measured HRTFs of Subject K, A and R for the frontal ipsilateral, frontal contralateral and rear ipsilateral locations. 80

3.10 Raw HRIRs and their minimum-phase representations. (a) HRIR of contralateral ear of azimuth 60° and 0° elevation (b) HRIR of contralateral ear of azimuth 80° and 0° elevation 82

4.1 Comparison of a frontal projection and a conventional side projection headphone. 86

LIST OF FIGURES

4.2 Sound transmission through the external ear in free-field a) Sketch of the anatomy b) an analogue model. 95

4.3 Sound transmission through the external ear in presence of headphone a) Sketch of the anatomy b) an analogue model. 97

4.4 Transmission in a binaural system with frontal projection headphones. . 99

4.5 How frontal projection provides the 3D cues. 102

4.6 Measurement of HPTFs using a HATS dummy head. 104

4.7 Top : Variability in the side emitter headphone response based on 10 trials. Bottom : Standard Deviation within the 10 trials. 105

4.8 Top : Variability in the frontal emitter headphone response based on 10 trials. Bottom: Standard Deviation within the 10 trials. 106

4.9 Collapsed view of headphone response 107

4.10 Type-1 Equalization process. 108

4.11 Example of a measurement of free-field response of the frontal projection headphone. 109

4.12 Type-2 Equalization process. 110

4.13 The free-field response based on 10 trials and their average FFR taken as the representative transfer function. 111

4.14 Comparison of Frontal emitter response and Side emitter response for Subject ‘K’ and HATS dummy head (Bottom). Note that the bottom curves have been shifted by -40 dB for better clarity. 113

4.15 Solid Line: SD between FER and SER; Dotted Line: Blauert’s observed pattern of SPL differences between front and rear directions. 114

4.16 (a) Comparison of FER and Frontal HRTFs (0° to 50°) (b) Comparison of SER and Rear HRTFs (100° to 140°). 115

4.17 Front-back confusions (%) for the side emitter (Type-1 equalization) and frontal emitter (Type-1 and Type-2 equalization). 118

4.18 Relative distance from the head for different equalization conditions. . . . 119

4.19 Localization Response Chart. 120

4.20 Percentage localization accuracy in Frontal and Rear directions using both frontal and side projection emitters. 121

4.21 Localization test responses for (a) Side Emitter (b) Frontal Emitter. . . . 123

4.22 Frequency response of the modified HRTF spectrum after spectrum flattening. 124

4.23 Percentage rate of reversals for various experimental conditions. 126

LIST OF FIGURES

4.24 Process of obtaining decorrelated signals for both left and right channels. 127

4.25 Frontal bias filter specifications. Bands (1, 3, 5) are frontal boosted bands; Bands (2, 4) are rear boosted bands. 128

4.26 Subjective test response chart. IH = In-head localization, NF = Near front, F = Front, FF = Far front. (1, 1'), (2, 2'), (3, 3'), (4, 4') are all symmetrical pairs of same ITD and ILD values. 129

4.27 Relative distance from the head (Subjective test 2). 130

5.1 The closed-back, open-back and the open headphones used in this study. 135

5.2 Headphone transfer function and equalization filters for various headphone types. 136

5.3 Stimulus-Response pairs for distance (in centimeters) using Non-individualized HRTFs for the two cases a) without and b) with equalization. The data points are shown on a log-log scale. The solid line in each panel of the figure represents the best linear fit of the stimulus location to the response location. Also shown is the correlation (ρ) between the log-stimulus and the log-response distances and the compressive function fit parameters k and m 139

5.4 Stimulus-Response pairs for distance (in centimeters) using Individualized HRTFs for the three cases:a) No equalization, b) Non-individual equalization, c) Individual equalization. The data points are shown on a log-log scale. The solid line in each panel of the figure represents the best linear fit of the stimulus location to the response location. Also shown is the correlation (ρ) between the log-stimulus and the log-response distances and the compressive function fit parameters k and m 141

5.5 Stimulus-Response pairs for distance (in centimeters) using Individualized HRTFs and Individualized equalization to investigate the effect of headphone repositioning. The data points are shown on a log-log scale. The solid line in each panel of the figure represents the best linear fit of the stimulus location to the response location. Also shown is the correlation (ρ) between the log-stimulus and the log-response distances and the compressive function fit parameters k and m 143

6.1 Comparison of measured and modeled proximal range HRTFs. Column (a) HRTF at a distance of 35cm and 50° azimuth; Column (b) HRTF at a distance of 45cm and 60° azimuth. 159

LIST OF FIGURES

6.2 Headphone responses of Sennheiser HD600 and the frontal projection head-
phone measured on the blocked ear-canal of a B & K HATS dummy head. 162

6.3 Stimulus-response pairs for distance (in cm) using non-individualized HRTFs.
Note the log-log scale. The solid line in each panel represents a linear re-
gression fit. The dashed line in each panel corresponds to the “correct”
responses. Also shown is the correlation coefficient (ρ) between the log-
stimulus and the log-response distances and the compressive function fit
parameters k and m (Eq. 6.10). Left column corresponds to the lateral an-
gles and the right column represents the data from the near-medial angles.
Note that only the distance responses greater than 15 cm are displayed for
visual clarity. 165

6.4 Stimulus-response pairs for distance (in cm) using individualized HRTFs.
The layout is as in Fig. 6.3. 166

6.5 Top: Correlation coefficient between the logarithm of actual and responded
distances; Bottom: Slope of the linear regression line best fitting the log-
stimulus data to the log-response data for azimuth angles (between 0°
and 180° for both (a) Non-individualized HRTFs, and (b) Individualized
HRTFs. 168

6.6 Approaches to model distance-dependent individualized frontal (a) and
rear (b) HRTFs. 172

6.7 Rear directional filter for all the azimuth angles for the ipsilateral ear.
Black: rear directional filters of all the subjects of the CIPIC database.
Grey (Bold): average rear directional filter of all the subjects. 174

6.8 Type-2 equalization filter calculated by taking the inverse of an average
FFR based on 10 trials. 176

6.9 Low-frequency compensation filter to compensate for the deviation in the
low-frequency response of the modeled HRTF from the measured individ-
ualized HRTF. 178

6.10 Different steps involved in the modeling process. (a) Subject K, source is
placed at 40° azimuth angle and 70 cm from the center of the head (b)
Subject A, source is placed at 60° azimuth, 50 cm from the center of the
head (c) Subject R, source is placed at 100° azimuth, 60 cm from the
center of the head. The azimuth angle is measured with respect to the
sagittal plane. 180

LIST OF FIGURES

6.11 Comparison of modeled individualized frontal ipsilateral HRTFs and measured individualized frontal HRTFs of all the three subjects for different azimuthal angles and distances, as indicated in each panel. 182

6.12 Comparison of modeled individualized frontal contralateral HRTFs and measured individualized HRTFs of all the three subjects for different azimuthal angles and distances, as indicated in each panel. 183

6.13 Comparison of modeled individualized rear ipsilateral HRTFs and measured individualized HRTFs of all the three subjects for different azimuthal angles and distances, as indicated in each panel. 184

6.14 Root mean spectral difference (RSD)scores of the modeled HRTF with respect to the measured individualized HRTFs for the ipsilateral and contralateral ear. 185

6.15 Localization responses for modeled individualized, measured individualized, and measured non-individualized HRTFs for all the three sets of distances. The distances are divided into three categories: near (20 cm, 30 cm, 40 cm), medium (50 cm, 60 cm, 70 cm), and far (80 cm, 90 cm, 140 cm). Modeled distance dependent individualized HRTFs were played back using frontal projection headphones, while the measured individualized and non-individualized HRTFs were played back using Sennheiser HD600 headphone. 187

6.16 Stimulus-Response pairs for distance (in cm) using modeled individual HRTFs (with and without intensity cues) and real sound sources (without intensity cues). The layout is as in Fig. 6.3. 189

6.17 Correlation and Slope of the subject’s responses for the modeled HRTFs without and with intensity cues 190

7.1 The spectral notches and peaks formed due to the interaction of the acoustic wave with the pinna. The peaks and notches are labelled as $P1$, $P2$; ... and $N1$, $N2$ and so on, according to the increasing order of frequencies. . 199

LIST OF FIGURES

7.2 Frontal projection headphone responses measured on human subjects. Black: Average frontal projection headphone response of the five trials of measurements. Blue: Frontal projection headphone response based on five different trials. Frontal projection headphone responses were measured using a B & K 4101 binaural microphone placed at the blocked ear canal of the human subject. The notches at the high frequencies are due to the interaction of the frontal projected sound with the pinna. Also note that the low-frequency response of the frontal projection headphone response is poor due to its open structure. 202

7.3 Illustration of the pinna extraction algorithm developed by Raykar *et al* (2006) for the CIPIC HRIR of Subject 127.(a) : Signal, (b) : LP residual, (c) Windowed LP residual, (d) Autocorrelation, (e) Windowed autocorrelation. (f), (g),(h),(i),(j) are their corresponding magnitude spectrum. (k) corresponds to the group delay spectrum of the windowed autocorrelation. Note that the pinna spectral notches are prominent in the group delay spectrum due to its high-frequency resolution property. 204

7.4 The different steps in the pinna extraction algorithm developed by Raykar *et al.* [8]. 208

7.5 Variation of center frequencies of the peak *P1* and notches *N1*, *N2* for Subjects 10, 11, 48, and 138 taken from the CIPIC database. The extraction of the peak and notch positions were carried out using Raykar’s algorithm. Note that the notch positions increase monotonically with elevation. On the other hand, the center frequency of the peak *P1* is more or less constant with respect to elevation. 209

7.6 Rear directional filter for different elevation angles. Blue: Rear directional filters of all the subjects. Red (Bold): Average rear directional filter of all the subjects. 210

7.7 Comparison of measured and modeled HRTFs in the median plane for subject A for different elevations. 212

7.8 Comparison of measured and modeled HRTFs in the median plane for subject R for different elevations. 213

7.9 Comparison of measured and modeled HRTFs in the median plane for subjects K for different elevations. 214

LIST OF FIGURES

7.10 Modeled and measured rear-directional HRTFs in the median sagittal plane. The rear-directional filter is applied to the frontal HRTFs to obtain the rear-directional HRTFs. HRTFs for 20°, 60°, and 80° are shown for all the three subjects. 215

7.11 Root mean square spectral differences for the modeled frontal and the rear HRTFs for all the elevation angles. 216

7.12 Subjective responses for measured (a) Conventional headphones) and modeled individualized HRTFs (b) Conventional headphones c) Frontal projection headphones) in the median plane. Responses for the non-individualized HRTFs (d)) is also plotted. Note that the errors are greater for the measured non-individualized HRTFs. 218

7.13 Front-back confusions for the 4 stimulus conditions. 1) Measured individualized HRTFs 2) Modeled HRTFs over conventional headphones 3) Modeled HRTFs over frontal projection headphones 4) Measured non-individualized HRTFs. 220

8.1 A summary of the differences between natural listening and headphone listening and the corresponding signal processing techniques to solve these challenges for natural sound rendering. The main challenges and their corresponding signal processing techniques in each category (source, medium, and receiver) are highlighted and their interactions (not shown here) are further discussed in the chapter. 226

8.2 Virtualization of (a) multichannel loudspeaker signals $x_m(n)$ (adapted from [8]), and (b) multiple sources $s_k(n)$ and environment signals $a_L(n)$, $a_R(n)$, $y_L(n)$, $y_R(n)$ are the signals sent to the left and right ear, respectively. Note that head tracking can be used to update the selected directions of HRTFs/BRIRs. 227

8.3 (a) Human ears act as a natural filter in physical listening. (b) The natural HRTF filter is modeled by a digital filter using various individualization techniques. 238

8.4 Natural sound rendering system for headphones: an integration of all the signal processing techniques reviewed in this paper. 240

8.5 (a) Skeleton of the 3D audio headphones holding both the frontal projection as well as the conventional side emitter (b) A finished 3D audio headphone prototype. 241

LIST OF FIGURES

8.6 A screenshot of the GUI designed for the test 243

8.7 MOS of the 4 measures for each track: (a) sense of direction; (b) externalization; (c) ambience; and (d) timbral quality. 244

8.8 Scatter plot for 4 measures: (a) sense of direction; (b) externalization; (c) ambience; and (d) timbral quality. 245

8.9 Comparison of the overall performance: (a) MOS for 4 measures; (b) scatter plot for all scores; and (c) preference of the tracks. 246

A.1 Typical HRTF/DRIR measurement setup 294

A.2 SOFA attributes contained in the HRTF database 296

B.1 HRTFs at the left ear for various source directions and distances simulated by the spherical head model. 301

D.1 Acoustic impedance analogue of a dynamic driver 311

D.2 A simplified headphone-earcanal model 311

D.3 Acoustic circuit analogue for a simple headphone-earcanal model 313

D.4 Types of headphones and schematic characteristic responses at low-frequencies. a) closed headphones b) open foam cushion c) cushion of impermeable foam with fixed resistance d) Integrated open-back headphones. Figure adapted from reference [301]. 314

D.5 Simple model of headphone-earcanal system with leaks. 315

D.6 Acoustic circuit analogue for a headphone model with leaks. 315

D.7 Simple model for an open-back headphone. 317

D.8 Acoustic circuit analogue for an open-back headphone. 317

List of Tables

1.1	Differences between natural and virtual spatial hearing over headphones .	4
2.1	Different type of artificial heads and their features	28
2.2	List of measured HRTF databases	30
2.3	Directional cues reported in [9]	38
2.4	Equalization techniques for different playback modes (Binaural and Stereophony)	49
2.5	Analogies between Electrical, Mechanical, and Acoustical domains	52
4.1	Comparison of the various HRTF individualization techniques.	91
4.2	Experimental Conditions in Superposition Test	124
5.1	Summary of results	145
7.1	Comparison of the notch frequency positions in the modeled and the mea- sured HRTFs in terms of the notch frequency distance (NFD)	217
8.1	Overview of typical techniques in BSS	232
8.2	Comparison between BSS and PAE in sound scene decomposition. . . .	235
8.3	Stimuli used in the experiments.	242
8.4	ANOVA results for MOS of the four measures.	247
D.1	Analogies between Electrical, Mechanical, and Acoustical domains	308
D.2	Acoustic elements and their electrical equivalent circuits, and R , L , C are expressed in terms of the mechanical dimensions l , b , d (m), $S(m^2)$, V (m^3). Air density is $1.2\text{ Kg}/m^3$, Speed of sound $c = 340\text{ m}/s$, Specific heat of air $= 1.4$, $P_{atmos} = 105\text{ Pa}$	309

List of Abbreviations

AP	Auditory parallax
BEM	Boundary element method
BRIR	Binaural room impulse response
BSS	Blind source separation
BW	Bandwidth
CASA	Computational auditory scene analysis
DF	Diffused-field
DFT	Discrete fourier transform
DRR	Direct to reverberant energy ratio
DVF	Distance variation function
EQ	Equalization
FEC	Free-air coupling equivalent
FEM	Finite element method
FER	Frontal emitter response
FF	Free-field
FFR	Free-field response of the frontal projection headphone
FIR	Finite impulse response
FPN	Frontal perception notch

List of Abbreviations

GUI	Graphical user interface
HATS	Head and torso simulator
HPTF	Headphones transfer function
HRIR	Head related impulse response
HRTF	Head related transfer functions
IDFT	inverse discrete fourier transform
IHL	in-head localization
IIR	Infinite impulse response
IPD	Interaural phase difference
IRS	Inverse repeated sequence
jnd	Just noticeable difference
LP	Linear prediction
LS	Least-squares
MLS	Maximum length sequences
MOS	Mean opinion score
MSE	Mean square error
NFD	Notch frequency distance
NMF	Non-negative matrix factorization
PAE	Primary-ambient extraction
PCA	Independent component analysis
PCA	Principal component analysis
PDR	Pressure division ratio
RDF	Rear directional filter

List of Abbreviations

REF	Reference sound field
RSD	Root mean square spectral difference
SER	Side emitter response
SHM	Spherical head model
SOFA	Spatial oriented format for acoustics
VBAP	Vector based amplitude panning
WFS	Wave field synthesis
WGN	White gaussian noise

Abstract

3D audio reproduction over headphones has become one of the most commonly used forms of playback system. Headphones provide a private listening space and are extremely convenient to use due to portability. However, headphones playback of 3D audio is marred by several challenges.

Head related transfer functions (HRTFs) contain all the spatial information involved in the propagation of an acoustic wave from the source position to the receiver position. HRTFs encapsulate the spectral, temporal as well as timbral effects in the source sound spectrum that are caused due to the interaction of the source wave (diffraction, reflections, and scattering) with the human torso, shoulders, head and the pinna for any given source location. Human ears are highly idiosyncratic and thus, HRTFs are also unique and vary largely among different subjects. Use of non-individualized HRTFs degrade the 3D audio perception by introducing front-back confusions, up-down reversals, in-head localization (IHL) and timbral coloration. Therefore, individualized HRTFs are necessary to provide an accurate and immersive perception of 3D sound. However, measuring these individualized HRTFs are extremely tedious. HRTFs have to be measured precisely for every azimuth, elevation, and distance for every individual, which is practically not feasible. For distances in the near field (< 1 m), the measurement of individualized HRTFs is even more difficult due to the large variation of HRTFs with distances in the near-field. Thus, easier techniques to obtain the individualized HRTFs at any source location have to be developed.

In the past, there have been several attempts at individualizing the HRTFs. The most widely used technique is to acoustically measure the HRTFs at the listener's ears at various source positions around the listener's head. Researchers have also modeled the individualized HRTFs by obtaining the listener's anthropometric features with the help of MRI, laser scanner, 3D imagery, or other 2D imagery techniques. Other techniques involve customizing the generic HRTFs using perceptual feedback. All these techniques to obtain the individualized HRTFs require highly precise individualized measurements, individual anthropometry features, or long hours of training that can cause severe fatigue to the listeners. ***In this thesis, a novel individualization technique is developed using "frontal projection headphones playback" that does not require any individualized acoustical measurements.***

The frontal projection headphones project the sound from the front directly onto the pinnae, unlike the conventional side emitter headphones. Frontal projection of sound during the playback process captures all the individualized pinna spectral features present in the HRTF. The first part of this thesis explains in detail the role of frontal projection headphones playback in modeling the individualized pinna spectral cues. Headphones inherently distort the input sound spectrum and thus, the effect of headphones has to be compensated prior to playback. Headphone transfer function (HPTF) can be considered as a combination of the headphone transducer response and headphone-ear coupling. Headphones-ear coupling depends highly on the human pinnae, thereby making the headphones response highly idiosyncratic. HPTF also displays large variation at high frequencies with repositioning of the headphone over the ear. Thus, a perfect headphone equalization filter cannot be designed that can compensate exactly for the headphone response. For the frontal projection headphones playback, a robust equalization filter known as the Type-2 equalization filter is designed. Type-2 equalization preserves the personal pinna cues that are generated during the playback process, and removes only

the distortion created by the transducer and the resonant modes of the earcup. However, conventional equalization technique would compensate for the entire headphone response and thus the individual pinna cues generated would also be removed. It is found that the individualized pinna cues embedded by the frontal projection headphone after Type-2 equalization model the true individualized pinna cues well. Subjective experiments prove that the frontal projection headphone improves the perception of 3D audio by reducing the localization errors.

The frontal projection headphones playback technique is used to model the distance-dependent individualized HRTFs in the horizontal plane. In order to develop this model, the important cues that affect distance perception are first identified with the help of detailed objective and subjective experiments. In particular, the role of the auditory parallax and the interaural cues are investigated. It is found that the auditory parallax cues play only a minor role for distance perception in the presence of interaural cues. The interaural cues play an extremely critical role in the distance localization process. In the proposed model, the distance-dependent spectral effects of the head and head shadow effects are both modeled by the spherical head model. The highly idiosyncratic pinna spectral cues are modeled by the frontal projection headphones. In this way, the frontal horizontal plane HRTFs can be modeled without any need for external individualized measurements. However, in order to model the rear HRTFs, a rear-directional filter is required. Moreover, Type-2 equalization of the headphones and a low-frequency compensation filter is additionally required in the modeling process. Objective and Subjective analysis showed that the modeled individualized distance-dependent HRTFs using frontal projection headphones match the measured individualized HRTFs.

The modeling of individualized HRTFs in the horizontal plane is then extended to the sagittal plane. The pinna spectral features that characterize the elevation perception are first studied. It is found that the position and center frequencies of the pinna notches vary

monotonically with elevation. The frontal projection headphone responses of the subjects are first captured by a one-time measurement before the modeling process. The important pinna features that affect the elevation are then extracted from the frontal projection response. Elevation perception is simulated by shifting the notch frequency positions corresponding to the elevation variation. With subjective and objective experiments, it is seen that the frontal projection headphones could model the sagittal plane HRTFs close to the individualized HRTFs.

It is observed that the use of frontal projection headphones could naturally enhance the veracity of the 3D audio perception. This approach can also be used in rendering natural sound over headphones especially for digital media content (movies, games, etc.). A 3D audio headphone, which consists of a unique combination of emitters and their orientation is developed. The 3D audio headphone holds both the frontal emitter¹, as well as the conventional side emitter on each side of the earcup. The primary or the directional cues extracted from the digital media content are played through the frontal emitters, while the ambience or the environment signal is played through both the frontal as well as the side emitters. Perceptual experiments carried out showed that the 3D audio headphones can provide a listening experience close to natural listening.

This thesis mainly focuses on the individualization of virtual audio using frontal projection headphones. With the emergence of mobile smartphones with high processing power and the growing market of headphones, personalized 3D audio has the potential to make a great impact with applications ranging from navigation systems, communication systems, assisting the visually impaired and related areas of health, and entertainment.

¹Please note that throughout this thesis, frontal projection headphones and the conventional side projection headphones are alternatively referred to frontal emitter and side emitter, respectively.

Chapter 1

Introduction

Sound plays an important role as a medium for communication and interaction in our day-to-day lives. In a real sound environment, we hear sounds from all the three dimensions perceived in terms of their direction, width, height, and depth. The ability of humans to make sense of their environment and interact with them depends strongly on the listener's ability to perceive several spatial attributes [10]. These attributes can be acoustical as well as non-acoustical cues. The acoustical attributes may contain frequency characteristics, intensity, time delays and the complex reflections, diffractions in the space, while the non-acoustical attributes include the visual cues, familiarity, and awareness of the space. Visual cues greatly influence the perception of the frontal space. However, sound has an added advantage over vision that humans can perceive sounds both from behind as well as top of the head.

1.1 Spatial Hearing

Spatial hearing can be categorized as virtual spatial hearing and natural spatial hearing. Virtual spatial hearing involves synthesizing the spatial auditory imagery of the acoustic event using a 3D-sound synthesis and playback system. Natural spatial hearing

CHAPTER 1. INTRODUCTION

refers to how we hear sounds in our everyday lives. In complex natural sound scenario with multiple sound sources, the perception of “spaciousness” or “locatedness” is often due to the blending or diffusion of sources [10, 11].

The main motivation of this thesis is : *To create a truly immersive and realistic acoustic space, which gives the listener a sense of “being there” at the acoustic event, thereby providing a virtual listening environment that is close to natural hearing*(Fig. 1.1). In this thesis, virtual audio is rendered over headphones due to its ease, portability, and convenience for personal use compared to loudspeakers.

The listening process in humans can generally be considered as a source-medium-receiver model, as stated by Begault [12]. To achieve a virtual listening environment very close to real life listening, it is important that we first look into the differences between virtual spatial hearing (over headphones) and natural spatial hearing (Table 1.1). The analogy of the source-medium-receiver model in this thesis can be elaborated as follows:

Source: Source can involve any number of vibratory sources in the environment. However, it is often convenient to describe the system in terms of a single source in isolation [12]. In virtual listening conditions, the sources are usually sounds that are captured by different recording techniques. In digital entertainment applications, the source typically consists of recording mixtures of movie, game, or music content. However, in natural spatial hearing, the sources can be referred to the physical sound sources and reflections/ambience due to the sound environment.

Medium: The medium refers to the paths by which the source arrives at the listener. In virtual spatial hearing, headphones serve as the rendering medium. Headphones are not acoustically transparent and thus, affect the input sound spectrum. The medium also includes the non-linearities of the sound reproduction system. In natural spatial hearing, the medium is a diffused acoustical space.

Figure 1.1: Rendering a virtual source environment very close to the real environment. The top panel shows the person in a museum listening to the virtual sounds of a waterfall. When the rendering is realistic, the listener feels that he is teleported to the real waterfall as in the bottom panel.

Receiver: The receiver essentially involves the listener's physiology. This includes the highly idiosyncratic filtering of the sound waves with the listener's pinna, head, shoulder and torso. To achieve a spatial perception close to real life listening, these idiosyncratic spectral features in the virtual audio should be exactly replicated at

CHAPTER 1. INTRODUCTION

Table 1.1: Differences between natural and virtual spatial hearing over headphones

	Natural spatial hearing	Virtual spatial hearing
Source	Physical sound sources and reflections/ambience due to the sound environment.	Sound recordings and mixes, mostly designed for loudspeaker playback.
Medium	Diffused-field	Headphones
Receiver	Individual filtering of sound due to listener's torso, shoulder, head and the outer ear. Listener's head can move freely without affecting the directions of the perceived sound.	Individual filtering of sound with headphones placed over the outer ear. The directions of the virtual sound will move along with the listener's head.

the eardrum. However, in virtual spatial hearing, this is not feasible if the source recordings are carried out on a generic dummy head. Moreover, virtual spatial hearing lack important dynamic head movements, which provide important cues for localization.

In this thesis, the challenges involved in the virtual spatial hearing over headphones are first introduced and subsequently, their respective solutions are proposed. In particular, the focus of the thesis remains on modeling the individualized HRTFs using the frontal projection headphones. Individualized HRTFs are modeled across all the three dimensions (azimuth, distance, and elevation). Furthermore, a new headphone equalization technique named Type-2 equalization is proposed for the frontal projection headphones. The advantage of Type-2 equalization technique is that such an equalization technique is subject-independent. In the next few sections, an overview of the existing spatial audio systems and the motivation of this thesis are explained in detail.

Figure 1.2: 3D Audio technologies can be categorized generically in terms of Surround sound systems, Binaural audio and Sound field technology.

1.2 3D Audio Technologies

A 3D audio system can, however, convincingly reproduce the height, azimuth, and the depth information and truly create an immersive virtual auditory space. This technology precisely emulates the propagation of sound field from the source to the eardrum of the listener and provides a very realistic perception of the 3D auditory scene. Olson [18] described the four necessary conditions to achieve a realistic perception of spatial audio:

1. The frequency range of the device should be wide enough to include all the audible components.
2. The dynamic range should be large to permit a distortionless reproduction of spatial audio.
3. The spectral features that correspond to the propagation of the acoustic waves from

CHAPTER 1. INTRODUCTION

the source position to the human ear need to be accurately reproduced.

4. The true acoustic space of the environment should be accurately preserved in the reproduced sound.

Though the first two conditions are well satisfied in all the systems these days, the challenges related to spatial audio reproduction originate from the third and the fourth conditions that form the core of some of the 3D audio technologies. These issues will be addressed in this thesis for the headphone reproduction of 3D audio.

3D audio capture and rendering techniques can broadly be categorized in terms of three technologies namely, surround sound systems, sound field technology, and binaural technology (Fig. 1.2). In this thesis, the focus will be primarily on the binaural audio reproduction over headphones. A brief introduction on the surround sound and sound field technology is presented next.

1.2.1 Surround Sound Systems

Surround sound systems are most commonly used in movie screens, gaming and other home entertainment applications. Surround sound technologies often contain multiple loudspeakers in its playback system and helps in adding a sense of envelopment to the listener's perception of sound. The most common loudspeaker configuration is the 5.1-channel surround sound system. The layout of the loudspeakers and the channel configuration are laid out in the ITU-R BS.775 [13] standards. Though 5.1 surround sound configuration can achieve a degree of realism in its reproduction, there are several limitations in this playback configuration. The main limitation of the 5.1 surround sound format is that it cannot achieve an accurate 360-degree phantom imaging capability.

Other multichannel loudspeaker configurations with more number of channels (7.1, 10.2) provide a wider listening area and give more freedom in positioning the phantom

images. Tomlinson Holman [14] used this concept to design the 10.2-channel surround sound system by additional wider side-front loudspeakers and a center-rear channel to the basic 5.1 system. Two other height channels are added in order to create elevated frontal images. NHK experimented with a 22.2 multichannel surround system for an ultra-high-definition video system in Japan [15]. The 22.2 system contains an upper layer with 9 channels, middle layer with 10 channels, and the lower layer with 3 channels. The main advantage of the NHK system is that it provides spatial realism over a wide area. Moreover due to the upper and lower layer of channels, perception of a sound approaching from above or below is easily achieved. However, due to the requirement of large number of channels, installation of this system in different screens is extremely difficult.

Vector based amplitude panning (VBAP) is a panning technique, which balances amplitude control between various set of speakers to create virtual phantom images [16]. However, VBAP has many limitations, one of them being the inherent implementation of binaural processing mechanisms [16]. The perceptual effect in the VBAP system stems from the ILDs competing with the ITDs. The overall perception is due to a combination of the interaural cues, which can be validated using models of binaural processing [17]. Though the spatial impression with these surround sound systems is quite convincing, it is far from the “true” realism or immersiveness.

1.2.2 Sound field Technology

Techniques like Ambisonics, wave field synthesis exploit the inherent property of sound wave propagation in space [10].

Ambisonics: It is a recording and playback technique, which uses multichannel mixing technology in live and studio applications to recreate a true realistic sound field as compared to traditional surround systems. This technique was invented by

CHAPTER 1. INTRODUCTION

Gerzon in 1976 [19] based on his work on the B-Format or four-channel recording standard. Ambisonics technique makes use of a sound field microphone. One can create a true 3D sound field using six or more speakers arranged around the listener using this technique. The system can be adapted to be used with any number and any variety of speaker arrangements including the ITU 5.1 arrangement. Despite all the advantages, this technology has not yet been accepted in the commercial world.

Wave field synthesis (WFS): WFS was invented by Berkhout in 1988 [20]. A typical WFS system aims to synthesize the true sound field within the entire listening space such that the sweet spot area includes the entire listening space. A WFS system contains two subsystems, namely, the synthesis system function and an analysis system function [21]. In the synthesis system function, the primary source signals are filtered, delayed, and weighted to compute all the driving signals. These driving signals are inputs to the loudspeaker array. The analysis system function analyses the reproduced sound field at different listening positions by computing the sound pressure signals due to contributions from weighted and delayed driving signals. In order to practically realize the WFS reproduction system, some of the approximations that are often carried out are:

1. The loudspeaker plane is reduced to a line array configuration on the horizontal plane due to the infeasibility of covering the entire vertical plane with loudspeakers.
2. Moreover, as the number of loudspeakers cannot be infinite, the continuous line array is reduced to a finite discrete array with uniform spacing between the loudspeakers.

One of the main challenges of WFS is that the high frequencies are not well reproduced beyond a cut-off frequency, known as aliasing frequency. This is because of the spatial

sampling of the continuous loudspeaker array.

1.2.3 Binaural Technology

Binaural technology essentially encompasses a set of tools for recording or synthesizing, and rendering binaural signals at the listener's ears [11, 12]. In this thesis, the virtual spatial hearing is achieved using binaural technology due to the following reasons:

1. Binaural spatialization technique is much more compact and simpler to use compared to other loudspeaker array techniques.
2. Binaural technology enables us to create a virtual source anywhere in the 3D space with just two channels.
3. Binaural playback also results in lower equipment cost in terms of signal acquisition, storage and content delivery in comparison to the multichannel solutions with loudspeaker arrays.
4. The ease of implementation of binaural technology has also led to its use in handheld mobile platforms and multimedia devices that has become extremely popular these days.

Binaural technology aims to provide the auditory system an accurate emulation of how humans localize and perceive sounds in their real life. Binaural signals thus encode the natural cues of auditory localization that results from the interaction of the acoustic wave with the listener's morphology. The captured binaural audio can be reproduced using a pair of loudspeakers or through headphones. However, binaural reproduction over headphones is the main focus of this thesis due to its excellent channel separation and convenience for personal listening. The loudspeaker reproduction of binaural audio and the challenges involved is explained in the next section.

1.2.3.1 Binaural reproduction over loudspeakers

Gardner constructed a 3D sound system using conventional stereo loudspeakers [22]. This is accomplished by combining a binaural spatializer with a transaural audio system. Hence, this technique is also known as transaural 3D audio. An important condition in binaural rendering is that the left and right recorded/synthesized signals have to be routed exactly to the left and right ears, respectively. In a transaural 3D audio system, acoustic crosstalk occurs. This means that the left signal that is dedicated to the left ear is perceived not only by the left ear but also by the right ear, and vice versa. The presence of crosstalk inherently distorts the spatial image as the virtual image collapses to a narrow and frontal image with poor elevation perception. Thus, cross-talk cancellers have to be implemented to cancel the unwanted acoustic crosstalk between the left and the right channels [23] (Fig. 1.3). The crosstalk canceller pre-processes the binaural signal by adding a crosstalk component to the loudspeaker signal in order to cancel the contribution of the contralateral loudspeaker (Fig. 1.3). The following equations describe the derivation of loudspeaker signals:

$$X_L(f) = \frac{B_L(f) \times H_{LL}(f) - B_R(f) \times H_{RL}(f)}{H_{LL}(f) \times H_{RR}(f) - H_{LR}(f) \times H_{RL}(f)}, \text{ and} \quad (1.1)$$

$$X_R(f) = \frac{B_R(f) \times H_{RR}(f) - B_L(f) \times H_{LR}(f)}{H_{LL}(f) \times H_{RR}(f) - H_{LR}(f) \times H_{RL}(f)}, \quad (1.2)$$

where B_L and B_R are the binaural signals; H_{LL} and H_{RR} refers to the direct path between each loudspeaker and its corresponding ipsilateral ear; H_{LR} and H_{RL} corresponds to the crosstalk path between the left and right loudspeakers and their contralateral ears. The pre-processing filter has the following functions:

1. Cancels the acoustic crosstalk between left and the right ears.

Figure 1.3: Crosstalk in Binaural audio playback using loudspeakers.

2. Corrects for the transfer function between each loudspeaker and its ipsilateral ear to achieve a very transparent reproduction of the binaural signals.
3. Removes the inherent response of the loudspeaker transducer and preserves only the acoustic propagation effects.

An example of a crosstalk cancellation technology is the transauralTM technology [24,25]. Kirkeby [26] proposed a stereo dipole set-up. In this setup, the loudspeaker pair is placed close to each other with a small separating angle of 5° and 10° , which can be approximated as an acoustic dipole. Reducing the loudspeaker spacing would simplify the crosstalk cancellation process as the loudspeaker counteraction is minimized. One of the serious limitations in binaural playback over loudspeakers is that the crosstalk cancellation works well only in a narrow region in between the speakers, resulting in a limited sweet spot area. The system is also highly sensitive to listener movements away

from the sweet spot region.

1.3 Motivation of the Thesis: Binaural audio over headphones

The work carried out in this thesis mainly focusses on the reproduction of binaural audio over headphones. Binaural playback over headphones is the most convenient form of playback mechanism because of its excellent channel separation. Thus, an additional crosstalk cancellation filter is not required. With the headphones market growing enormously and with the increase in preference for personal audio, headphones playback method is the most preferred 3D audio reproduction system. However, binaural playback over headphones though extremely convenient, has several challenges and limitations.

Binaural audio is idiosyncratic and varies highly amongst human subjects. Thus, it is recommended that individualized binaural audio is used for 3D audio playback. When the synthesized or recorded binaural audio is not individualized, a highly distorted 3D audio image is perceived. Some of the main challenges that results with the use of non-individualized HRTFs are:

Front-Back confusions: The synthesized/recorded virtual sounds from the frontal directions are often perceived to be coming from the rear directions. This is referred to as front-back confusions and is commonly found in headphone playback of binaural audio.

Elevation localization error: Use of non-individualized binaural audio affects elevation perception the most. Confusions in the form of up-down reversals are common with non-individualized HRTFs.

In-head localization: Virtual images are perceived inside the head, and lacks exter-

nalization.

Timbral coloration: The use of non-individualized binaural audio leads to coloration of the true virtual sources, since the spectrum at the eardrum deviates from the subject's true spectrum.

Synthesizing binaural audio tailored for a particular listener is challenging, as the HRTF measurements or binaural recordings have to be carried out for every individual at different source positions in the 3D space. Furthermore, the headphones, by itself, is not acoustically transparent as it modifies the input sound spectrum of the binaural audio. Thus, equalization of headphones is necessary. Since, the headphone responses are also highly idiosyncratic, individualized headphone equalization is necessary. The headphone responses also vary with repositioning due to high headphone-ear coupling. The equalization filter obtained can never be accurate based on a single trial and an average of several trials is suggested for the measurement of representative headphone transfer function [27]. An imperfect equalization filter can also lead to perceptual degradation of the spatial audio [28, 29]. In this thesis, a new technique is presented for individualizing the binaural audio using frontal projection headphones playback. The main advantage of the proposed technique is that it does not require any individualized acoustical measurements to model the individualized HRTFs.

1.4 Main contribution of Thesis

The main contribution of this thesis can be summarized as follows:

1. It is found that the frontal projection headphones can emulate the highly idiosyncratic pinna spectral cues. Synthesized binaural audio played back using both frontal projection headphones and conventional side projection headphones proved

CHAPTER 1. INTRODUCTION

that the frontal projection headphones enhance the frontal image. Subjective experiments revealed that the frontal projection headphones reduce the front-back confusions and improves the localization perception.

2. A robust equalization technique, named as Type-2 equalization, is proposed for the binaural audio playback using the frontal projection headphones. The most important advantage of the Type-2 equalization is that it is subject-independent, unlike the conventional equalization methods.
3. The effect of equalization of headphones on distance perception is studied in detail. Low-frequency equalization of headphones is found to be sufficient for good distance perception close to the perception with individualized HRTFs.
4. The perceptual cues that affect distance perception over headphones is investigated. It is found that the perceptual effects of auditory parallax cues are negligible for distance perception in anechoic headphone listening conditions. The ILD cues are found to be the most important cue that aid distance perception in anechoic conditions over headphones.
5. Distance-dependent HRTFs are measured for 9 distances in the proximal region and 72 angles in the horizontal plane. The DSPLab HRTF [30, 31] database is created in the SOFA [32] format and is released online for free.
6. The individualized distance-dependent HRTFs are modeled using the frontal projection headphones along with a cascade of signal processing blocks. Detailed objective and subjective analysis showed that the modeled individualized distance-dependent HRTFs and the measured individualized distance-dependent HRTFs are similar.
7. Individualized HRTFs at various elevations in the median plane are modeled using the frontal projection headphones playback technique [33]. It is found that the

pinna notches vary monotonically with elevation. The frontal projection headphone response is first measured that captures the idiosyncratic pinna spectral features. The notches and other spectral features that depend on elevation are extracted from the frontal projection response using a notch extraction algorithm, proposed by Raykar *et al.* [8]. The extracted notches in the frontal projection response are shifted in frequency to simulate the elevation perception. Detailed objective and subjective experiments show that the modeled HRTFs are close to the measured individualized HRTFs.

8. Finally, a unique four-emitter 3D audio headphones [34, 35] is developed to enable natural sound rendering over headphones. The 3D audio headphone is constructed by mounting a frontal projection emitter and the conventional side emitter on each earcup. The primary and ambient cues are first extracted from the digital media content using the primary-ambient extraction algorithm [36]. The directional cues are presented through the frontal emitters, while the ambience components are played through both the frontal as well as the side emitters. The natural sound rendering system has been validated with the help of detailed subjective experiments.

1.5 Applications of 3D audio

3D audio has found its applications in various fields, like communications (teleconferencing), navigation systems (flight navigation systems), simulations, broadcast entertainment (television, cinemas), and interactive environments (virtual reality games), etc. With the growing popularity of headphones and high processing power of the present day smartphones, various 3D audio apps have also been developed for mobile phones [30]. 3D audio can also be used as an alternative mode to guide the visually impaired while

CHAPTER 1. INTRODUCTION

walking, and in their other daily activities. Frauenberger developed a virtual audio reality system to assist the visually impaired people to use computers [37]. Goose [38] developed an interactive web browser using spatialization to convey hypermedia document structure. Papadopoulos [39] proposed a new approach for abstracting network information in security and management using 3D audio. Researchers have also introduced an immersive spatial audio representation of all the network events. Spatial audio display finds its use in aircraft cockpits and ground combat vehicles as it aids in faster visual search times [40,41]. More recently, several virtual reality headsets, like the Oculus Rift, Sony's Project Morpheus, Microsoft Hololens, etc. have started using head tracked virtual audio to enhance the immersive experience. Another application is in the process of natural sound rendering of the virtual sounds over headphones. One of the outcomes of this thesis is the development of the 3D audio headphone, which is a multi-emitter headphone that helps in natural sound rendering of multimedia content.

1.6 Coordinate System

Figure. 1.4 indicates the coordinate system followed throughout in this thesis. A typical polar coordinate system (r, θ, ϕ) is considered in all the experiments. The distance r refers to the distance between the source and the center of the head, θ and ϕ corresponds to the azimuth angle and elevation angle measured with respect to the mid-sagittal axis.

The angle coordinate $(\theta, \phi) = (0^\circ, 0^\circ)$ corresponds to the location right in front of the listener. Increasing θ corresponds to the source location moving in the clockwise direction. $\theta = 90^\circ$ and 270° corresponds to the position pointing directly to the right ear and left ear, respectively, $\theta = 180^\circ$ corresponds to the source position directly behind the head. Similarly, $\theta = 90^\circ$ and 270° corresponds to a source location directly above and below the head, respectively. $\theta = 180^\circ$ corresponds to the position directly behind

Figure 1.4: Coordinate system followed in this thesis.

the head. Furthermore, mid-sagittal plane or median plane is a plane which passes from anterior to posterior, dividing the body into left and right halves.

1.7 Overview of Thesis

The thesis consists of nine chapters, and its organisation is as follows:

Chapter 2: All the necessary background related to binaural audio reproduction over headphones is explained in this chapter. In particular, the spatial cues required for localizing different directions are presented. The decoupled and non-decoupled modes of equalization are presented in detail. Furthermore, the disadvantages of using non-individualized binaural audio are also discussed.

Chapter 3: The procedures for the measurement of head-related transfer functions are given in detail in this chapter. The development of the DSPLab HRTF database

CHAPTER 1. INTRODUCTION

and the associated spatial oriented format for acoustics (SOFA) is explained in this chapter.

Chapter 4: In this chapter, the frontal projection headphones are first introduced. The comparison between the frontal projection headphone and the conventional side projection headphones is carried out with the help of objective and subjective analysis. The Type-2 equalization for the frontal projection headphones is introduced in this chapter.

Chapter 5: The effect of equalization on distance perception is studied in this chapter. Both objective and subjective experiments are reported. In addition, the effect of headphone re-positioning on the distance perception is also presented.

Chapter 6: The model for the distance-dependent individualized HRTFs using the frontal projection headphones is developed and presented in this chapter. The role of auditory parallax and the ILD cues for distance perception is also investigated in this chapter. Finally, the model is validated with the help of subjective and objective comparisons.

Chapter 7: In this chapter, the modeling for the distance-dependent individualized HRTFs is extended to different elevations. The modeled HRTFs are compared with the measured HRTFs with the help of subjective and objective analysis.

Chapter 8: The signal processing techniques for the natural sound rendering of digital media content over headphones is presented in detail in Chapter 8. A special type of 4-emitter headphone is developed for the natural sound rendering.

Chapter 9: Finally, the conclusions and future work of this thesis is described in this chapter.

Figure 1.5: Overview of the thesis.

Figure.1.5 shows the overview of the thesis. At the top is the virtual spatial hearing using non-individualized HRTFs. Chapters 4, 5, 6 and 7 describe the proposed solution for the virtual spatial hearing to render the perception close to natural hearing. Chapter 8 marks the culmination of this thesis with an application, which makes use of the frontal

CHAPTER 1. INTRODUCTION

projection headphones to enable natural sound rendering.

Chapter 2

Background

Humans have the ability to localize sound objects with ease. The human brain decodes the directional information from the sound pressure level received at the eardrum. A realistic virtual auditory space can thus be created by capturing the signals at any recording point (eardrum or blocked ear canal entrance) and reproducing the binaural signal at the same or some other point within the ear canal of the listener's ear. This forms the basis of "Binaural Technology" [42].

2.1 Head Related Transfer Functions

In the free-field, the sound radiated from a point source reaches the ears after undergoing complex interactions, like diffractions and reflections with the anatomical structures (head, torso, and pinnae) of the listener. The resultant binaural signals at the eardrum due to these interactions, contain several important cues, such as the interaural time differences (ITD), interaural level differences (ILD) and the spectral cues (SC) that the auditory system uses to localize a sound source. The sound propagation from the source to each of the two ears can be seen as a linear time-invariant (LTI) process.

Head related transfer function (HRTF) is defined as the transfer function that de-

Figure 2.1: HRTF of KEMAR dummy head for an azimuth angle of 30° in the horizontal plane.

describes the acoustic propagation between the sound source and the listener’s ears. HRTF includes both the magnitude and phase information of the resultant sound wave after its complex interaction with the head, torso, shoulder, and the pinna. An example of a HRTF magnitude response is shown in Fig. 2.1.

Several definitions of HRTFs have been reported in the literature, as highlighted in [43]:

1. The ratio of the output of a probe microphone located at 1-2 mm from the eardrum of a human subject to the input of the loudspeaker [44].
2. The ratio of the output pressure of a probe microphone near the eardrum to the free-field pressure at the location of the probe microphone with head removed [45, 46].

CHAPTER 2. BACKGROUND

3. The ratio of the sound pressure at the opening of a blocked ear canal to the free-field sound pressure at the center of the head with the head removed [47].
4. The ratio of the sound pressure at the eardrum to the free-field sound pressure at the center of the head with the head removed [48].
5. The ratio of the sound pressure in the ear canal to the sound pressure in the canal with the source directly in front of the listener [49].
6. The ratio of the sound pressure in the left ear canal to the sound pressure in the right ear canal (Interaural HRTF) [46, 50].
7. The ratio of the sound pressure in the ear canal to the maximum sound pressure over all locations at that frequency [51].

Among the above list of HRTF definitions, the 2nd, 3rd, and the 4th definitions are most widely used. For any arbitrary source position, $HRTF_L$ and $HRTF_R$ for the left and right ears can be defined as:

$$HRTF_L = HRTF_L(r, \theta, \emptyset, f, a) = \frac{P_L(r, \theta, \emptyset, f, a)}{P_0(r, f)}, \text{ and} \quad (2.1)$$

$$HRTF_R = HRTF_R(r, \theta, \emptyset, f, a) = \frac{P_R(r, \theta, \emptyset, f, a)}{P_0(r, f)}, \quad (2.2)$$

where P_L and P_R represent the complex-valued pressures in the frequency domain at the left and right ears, respectively. The symbols r, θ, f, a represent the distance from the center of the head, azimuth, elevation, frequency, and radius of the head, respectively. P_0 represents the complex-valued free-field sound pressure in the frequency domain at the center of the head with head absent, which is defined as follows [52]:

$$P_0(r, f) = j \frac{k \rho_0 c Q_0}{4\pi r} e^{-jkr}, \quad (2.3)$$

Figure 2.2: Binaural synthesis and playback over headphones.

where ρ_0 is the density of the medium, c is the speed of sound, Q_0 denotes the intensity of the point sound source, respectively, and $k = 2\pi f/c$ represents the wave number. The transmission of outward waves with distance is denoted by a factor of e^{-jkr} . HRTFs are typically measured in anechoic conditions. Sometimes to incorporate room effects, HRTFs are also measured even in relatively diffuse environments. However, since these measurements depend highly on the reverberation in the room, their use case is limited. The measurement of HRTFs in anechoic conditions are explained in greater detail in Chapter 3.

2.2 Binaural Synthesis/Recording

Humans make use of three main cues to localize sound. They are ITD, ILD and SC. There are mainly two methods to obtain the binaural signals:

1. Recording the sound scene at the eardrum or the ear-canal entrance of a human subject or an artificial human head.
2. Synthesizing the virtual audio using HRTF filters.

Sections 2.2.1 and 2.2.2 elaborate the binaural recording and synthesis procedure.

2.2.1 Binaural recording of the sound scene

Binaural signals can be acquired by capturing the signals using a probe microphone placed at the blocked-entrance of the ear canal or at the eardrum of the listener. The recorded signal captures all the diffractions, and reflections of the acoustical waves with the morphology of the individual on its path to the eardrum. The positioning of the microphone within the ear and the stability of the head is very important in capturing the right information for a particular location of the source. Since the anthropometric properties of the ear are unique, the binaural recordings captured at any point within the ear canal are also highly idiosyncratic. Thus, to obtain convincing spatial auditory image devoid of all the confusions for all listeners, there is a need to carry out individualized binaural recordings, which is a highly tedious and time consuming process. Moreover, even a slightest listener head movements during the recording process can degrade the recorded binaural signal.

For generic purposes, researchers have developed artificial head mannequins (or dummy heads) made of specific materials and are fitted with anatomical structures, such as head, torso and pinnae similar to that of real humans in perspective of acoustics [53]. The design of a dummy head is based either on the average dimensions arrived from a certain population of listeners, or on the anthropometric features of a “typical” human subject. These artificial heads are extremely useful in carrying out a variety of acoustical measurements, binaural recordings, sound quality evaluation, measurement of headphone

Figure 2.3: a) B & K 4128 HATS dummy head [1]; (b) KEMAR by G.R.A.S [2]; (c) KU-100 by Neumann [3]; (d) HMSIV by Head Acoustics [4]; (e) MK2B by 01 dB-metravib [5]; (f) VALDEMAR [6].

responses, evaluation of headsets, and hearing aid testing. More importantly, the dummy head can simulate the acoustical contributions of the head, torso and the pinna due to the interaction with the sound waves. Binaural recordings can be carried out using these dummy heads as they contain a pair of microphones often kept at the eardrum, or at any point in the ear canal. Thus, the binaural signals can be captured at its eardrum with great ease and high repeatability.

Figure.2.3 shows a series of some popular dummy heads used for binaural recording/synthesis. The most important features of each dummy head are shown in Table 2.1. However, dummy head binaural recordings (non-individualized binaural recordings) comes with an inherent disadvantage that the spatial cues present in the recording does not match that of the listener, which distorts the spatial image considerably. In Section 2.8, we will explain some of the shortcomings of using dummy head recordings.

2.2.2 Creating an artificial sound scene by binaural synthesis

Binaural synthesis can be carried out by convolving the mono sound source signal ($x(n)$) with the head related impulse responses (HRIR) to obtain the left and right binaural signals (y_L and y_R) [12, 54–57]:

$$y_L(n) = x(n) * HRIR_L, \text{ and} \quad (2.4)$$

$$y_R(n) = x(n) * HRIR_R. \quad (2.5)$$

As explained above, the binaural signals can be obtained either by directly recording using a dummy head/human subject or by synthesizing using the HRIRs. HRIR is computed by taking the inverse Fourier transform of HRTF. In the case of generating binaural signals using HRIRs, the HRIRs have to be first measured at various locations in the 3D space. In the binaural playback using headphones, the non-ideal transfer functions of the microphones, the headphones and the coupling between the ear and the headphone earcup have to be compensated for. These non-ideal characteristics if not equalized, can cause severe distortions in both magnitude and phase of the binaural signal reproduced.

2.3 Measurement of HRTFs

HRTFs are typically measured by playing an excitation signal (e.g. maximum length sequence, exponential sweep, or white noise) through a loudspeaker and recording the impulse response at the eardrum of the listener using a probe microphone. The procedures for the measurement of HRTFs are explained in Chapter 3. When a sound scene is recorded at any point within the ear canal (typically at the eardrum or blocked ear canal entrance), the spatial information is naturally encoded through his or her own set of HRTFs. The characteristics of HRTFs are highly individual, since the morphological

CHAPTER 2. BACKGROUND

Table 2.1: Different type of artificial heads and their features

Type of artificial head	Manufacturer	Features
KEMAR [2]	Knowles manikin acoustic research/ G.R.A.S Sound and Vibration in Denmark	Humanlike appearance, Commonly used, Pair of small pinnae and large pinnae available
HATS [1]	Bruel & Kjaer	Includes head, torso and detailed pinnae; Used for electroacoustic measurements such as HRTF measurements, telephone, headphone testing
HMS IV [4]	HEAD acoustics GmbH	Consists of head, shoulder, simplified pinnae; Evaluation of sound quality and noise
MK2B [5]	01 dB-metravib	Used for sound quality recording in the automobile industry and psychoacoustics
KU-100 [3]	Neumann	Consists only the head; Used commonly for binaural recording
VALDEMAR [6]	Aalborg University	Design for research in binaural and spatial hearing

features of the ears of different people vary considerably. The impracticality of measuring individualized HRTFs for every human subject led to the development of generic HRTFs, where the HRTFs are measured on a dummy head. The price paid for the use of easily available non-individual HRTFs is the degradation of the spatial auditory display.

The early HRTF measurement studies were carried out in the 1940s by Wiener [58, 59]. Most of the measurements in the 1970's and early 1980's were confined to a few subjects in the horizontal or median plane [49, 60–63]. The absence of accurate digital measurement techniques in the early years suffered from low accuracy and inconvenient data storage. The advent of digital measurement systems allow quicker and more accurate HRTF measurements with higher spatial resolutions. Several researchers have measured HRTF

CHAPTER 2. BACKGROUND

databases as shown in Table 2.2.

Empirical measurements are important and an accurate approach to the measurement of HRTFs is necessary to create a realistic perception of virtual audio. Generally, the far-field HRTFs at various spatial directions are obtained from the signals captured at the two microphones at both ears by changing the relative direction between the sound source and subject. Such a measuring method is often very lengthy and can suffer from low-efficiency.

Other novel techniques of measurement of HRTFs have been developed, which uses the principle of microphone-loudspeaker reciprocity [64]. The impulse responses are measured in a single trial by placing a tiny loudspeaker at the entrance of the ear canal and recording the responses at every microphone placed spatially on the sphere (radius = 0.7 m) around the listener. In this way, the HRTFs of all the microphone positions can be derived at a single time. However, one of the main disadvantages of this method is the poor low-frequency response of the micro speaker with a frequency range going up to only 16 kHz. The other disadvantage is the low signal-to-noise ratio of the loudspeaker. All these methods measure the HRTF data at one or a few spatially discrete locations in a single run. HRTFs at other directions can be obtained by interpolation techniques [241] [156]. HRTFs can also be considered as a continuous function of spatial directions. Fukudome *et al.* [65] proposed a continuous measurement method for obtaining HRTFs continuously in the horizontal plane in a single run using a servo swivelled chair. HRTFs for any other arbitrary position can then be extracted from the measured signal by applying an appropriate signal processing method. Other researchers used white Gaussian noise (WGN) and a multi-channel adaptive filtering algorithm to measure the continuous HRTF data across all azimuths at different elevations [66, 67]. Ajdler *et al.* [68] used a moving microphone to measure the HRTFs in the horizontal plane. This method accounts for the doppler effect in the recording due to the movement of the microphone relative to a

CHAPTER 2. BACKGROUND

Table 2.2: List of measured HRTF databases

Owner	(Subjects, Directions)
IRCAM France [69]	(51; 187)
CIPIC , UC Davis [70]	(45;1,250)
University of Maryland [64]	(7;2,093)
Tohoku University, Japan [71]	(3;454)
Nagoya University, Japan [72, 73]	(96;72)
Austrian Academy of Sciences [74, 75]	(77;1,550)
TU Berlin,(3m,2m,1m,0.5m) [76]	(KEMAR;360)
Oldenburg University,(0.8m,3m) [77]	(HATS;365)
Genuit and Xiang [78]	(HMS I and HMS II)
Reiderer [79]	(51; $\phi = -30^\circ, -15^\circ, 0^\circ, 15^\circ, 30^\circ, 60^\circ$ and $90^\circ, \Delta\theta = 10^\circ$)
AUDIS database [80]	(20; 122)
Bovbjerg [81]	(VALDEMAR; 11,975)
Takane [82]	(3; 454)
Grassi [83]	(7; 1132)
Xie <i>et al.</i> , [84]	(52; 493)
Yu <i>et al.</i> , [85]	(KEMAR; 3,889)
Wightman and Kistler [44]	(10; 144)
Mller <i>et al.</i> , [86]	(40; 97)
DSP Lab HRTF database, [31]	(HATS;7,800)

fixed sound source. An important advantage of this technique is that the HRTFs at all angular positions along the horizontal plane can be measured in less than 1 second.

2.4 Directional localization cues

Humans primarily use the ILD, ITD, and SC for localizing directions in the 3D space. Additionally, head movements can provide dynamic cues that can aid in resolving localization confusions.

2.4.1 Interaural cues

The “duplex theory” [87] states that low-frequency sounds are localized by temporal cues and high-frequency sounds are localized by the intensity cues. The ILD cues are

CHAPTER 2. BACKGROUND

more prominent at the higher frequencies, since head-shadow effect becomes significant for high frequencies unless the source is close to the head (< 50 cm). This is because longer wavelengths can bend around the head, thereby minimizing the intensity differences. On the other hand, the ITD is effective in localizing low frequencies typically below 1.5 kHz. Research has also revealed that certain timing differences can be detected in the amplitude envelopes that can be used in localizing high-frequency signals [88–90]. The localization using an ITD envelope cue depends on the ability of the hearing system to extract the timing differences of the onsets of the amplitude envelopes. As the ILD and ITD values are increased, the sound image begins to shift towards the ear leading in time or greater in amplitude. The virtual image, however, stops moving along the interaural axis and remains at the leading ear once a critical value of ILD or ITD is reached. Researchers found that the effective range of ITD and ILD are approximately 0.005 to 1.5 ms and 1 to 10 dB, respectively [91,92]. Von Békésy observed that there could even be a vertical component of lateralization. He noted that listeners could experience the sound image moving upwards as the interaural differences were decreased [93].

Lateralization experiments involve the manipulation of ITD and ILD to determine the relative sensitivity of physiological mechanisms to these cues. These experiments are often carried out over loudspeakers. “Lateralized” is often referred to indicate that the spatial percept is heard inside the head along the interaural axis between the ears [12]. Variation of overall level of the left and right ear signals (over headphone playback) independent of frequency content can also lead to separating sound sources in most applications. This is because listeners are sensitive to ILD cues for lateralization across most of the audible frequency range [11].

Wightman *et al.* [94] studied the relative salience between ILD and ITD cues by experiments using synthesized stimuli with conflicting ILD and ITD cues. These interaural cues also play an important role in the externalization of the sound image especially

CHAPTER 2. BACKGROUND

when played back through headphones. Weinrich [95] suggested that ITD is the dominant factor in creating a sensation of spaciousness and sound image is localised within the head when there is a discrepancy between the ITD and ILD cues. Thus, only a correct ratio between the ITD and ILD can achieve optimal externalisation. Both ILD and ITD are frequency-dependent. According to the spherical diffraction theory, the ITDs at frequencies below 1.5 kHz are about 1.5 times larger than the ITD above this frequency [96,97]. The dependence of ILDs on frequencies is complex, since it is affected by the anthropometric properties of the individual's ear. Interaural cues cannot uniquely distinguish a location as they are determined only by the angle between the interaural axis and the sound source. Consequently, ITD and ILD information presents ambiguity between source locations on the surface of a cone centred on the interaural axis with its apex at the centre of the head, a locus of points known as the "Cone of confusion" [98]. However, such confusions are lesser in the real acoustical space. This is because the cone of confusion region is disambiguated by the unique monaural spectral cues in these directions [99,100].

2.4.1.1 Interaural time difference

ITD is estimated as the difference of the delays between the HRIRs of left and the right ear. ITD can also be estimated from the phase delay ($\phi_{L,R}$) differences between the left and the right ears, which is estimated from the HRTF phase spectrum:

$$\phi_{L,R} = \frac{\psi_{L,R}(f)}{2\pi f}, \text{ and} \quad (2.6)$$

$$ITD = \phi_L - \phi_R, \quad (2.7)$$

CHAPTER 2. BACKGROUND

where f refers to the frequency and $\psi_{L,R}(f)$ to the phase spectrum of the HRTF. The ITD is also computed from the regression analysis of the excess phase difference between the two ear responses [101]. The linear regression is carried out in the range 0.5 kHz to 2 kHz. The most common method of computing ITD is to calculate the cross-correlation function of the two ear responses, and equating the ITD to the maximum cross correlation value within ± 1 ms. The HRIRs are low-pass filtered with a cut-off frequency of 2 kHz prior to ITD calculation. Larcher [102] proposed the ITD model for a spherical head as (Fig. 2.4):

$$ITD = \frac{D}{2c} (\text{asin}(\cos\varphi \sin\theta) + \cos\varphi \sin\theta), \quad (2.8)$$

where θ corresponds to the azimuth angle, D is the diameter of the spherical head (17.5 cm), φ is the elevation angle, and $c = 344$ m/s is the speed of the sound.

The ITD varies with frequency and is highest for low frequencies. At higher frequencies, the ITD decreases to a value around 2/3 lower than the value at low frequencies, resulting in a low-frequency ITD (ITD_{lf}) and a high-frequency ITD (ITD_{hf}). These ITDs are defined as the low-frequency and the high-frequency asymptotes of the ITD. Kuhn [97] first investigated the interaural phase difference (IPD) for a plane wave diffracted by a spherical head model. At low frequencies with wave number k and a as sphere radius under the assumption that $(ka)^2 \ll 1$ at an azimuth angle ϕ , the ITD is given by:

$$ITD_{lf} = \frac{3a}{c} \sin \phi. \quad (2.9)$$

At high frequencies, the ITD is derived from the path difference, which includes the additional path for circumventing the head:

$$ITD_{hf} = \frac{a}{c} (\sin \phi + \phi). \quad (2.10)$$

Woodworth's formula [103] for modeling ITD is identified to be a model for high fre-

Figure 2.4: ITD model (Equation 2.8).

frequencies as it matches with Eq.2.9. Both ITD_{lf} and ITD_{hf} do not depend on frequency as seen from the equation Eqs. 2.9 and 2.10. At azimuths close to the median plane, $\phi \approx \sin \phi$. Therefore,

$$ITD_{hf} = \frac{2a \sin \phi}{c}, \quad \text{and} \quad (2.11)$$

$$\frac{ITD_{hf}}{ITD_{lf}} = \frac{2}{3}. \quad (2.12)$$

2.4.1.2 Interaural level difference

Interaural level difference (ILD) is defined as the difference of the spectrum magnitude between the left and the right HRTFs. The ILDs are highly dependent on frequency and it is an important cue for localization. The ILD variation is also observed to be

Figure 2.5: ILD model (Equation 2.13)

an important cue for distance perception [104]. Larcher introduced a unique ILD by integrating over a limited frequency range (f_1 to f_2), generally chosen between 1 kHz and 5 kHz (Fig.2.5) [102]:

$$ILD = 10\log_{10} \left[\frac{\int_{f_1}^{f_2} |H_L(f)|^2 df}{\int_{f_1}^{f_2} |H_R(f)|^2 df} \right], \quad (2.13)$$

where $H_L(f)$ and $H_R(f)$ are HRTF magnitude spectrum of left and right ear, respectively.

2.4.2 Spectral cues for localization: Role of pinnae

The pinna cues are the most widely studied amongst all the other localization cues. The pinna acts as a complex resonance cavity, which introduces unique spectral features characteristic to a particular source location. The morphology of the pinna has high inter-individual differences, which greatly affects the frequency spectrum at the eardrum. Due to this high variability of pinna morphology, the spectral characteristics become highly idiosyncratic. Thus, the spatial image of the perceived sound is distorted, if the HRTF used is not the listener's own HRTF [105–108]. The extent of similarity between the features of the individual and the HRTF in order to avoid degradation of the spatial image is still unknown. Some researchers [9, 109–111] proposed that the similarity covers single features, while others [107, 112, 113] proposed that broad ranges of frequencies are necessary for accurate localization. Toledo [114] proposed that the non-individual HRTFs would evoke the direction for which the relevant spectral features that cue localization were present. Toledo noted that the evoked direction is independent of the direction for which the HRTFs were measured, and thus, their spectral features could be identified and parameterized.

Pinna mainly affects the high frequencies, as the wavelength of the sound wave is comparable to the size of the pinna. The directional effects of the pinnae are particularly important for localization in the median plane, where the interaural cues are minimal [100, 115].

Some researchers [9, 110, 116–118] have observed unique spectral notches generated by the interference of direct sound entering the external auditory canal and the time-delayed reflections off the posterior concha wall. Several other researchers have suggested spectral peaks as important cues [46, 109, 119, 120]. Research has also suggested that there are other spectral features covering a larger frequency range that were relevant for sound localization [107, 111, 113, 121]. Poveda and Meddis [122] developed a physical model

CHAPTER 2. BACKGROUND

of diffractions and reflections in the human concha. They also provided evidence that spectral features generated in the concha for sources at all azimuths are similar within the frontal part of the ipsilateral hemisphere.

Some evidences do suggest that perception of direction can be influenced by the spectral cues independent of the location of the sound source. Blauert [109] carried out seminal works by studying in great detail the localization of 1/3-octave band noise played back through loudspeakers. He first divided the hemisphere around the listener into different regions in the horizontal plane as front, above and behind. It was observed that the responses of the subjects were clustered in certain regions depending on the center frequency of the noise stimuli, which led to the concept of “Directional bands” [109]. Blauert proposed that each direction in space is associated with a particular frequency band. A later study reported that the directional bands showed individual differences [123]. Itoh *et al.* noticed that directional bands also occurred for subjects who had naturally degraded free-field performance with broadband noise [123]. Blauert then measured the HRTFs and subtracted the sound pressure levels between the frontal and rear regions of the hemisphere, which led to the theory of “Boosted bands” [109]. Middlebrooks [120] carried out localization experiments for narrow band noise stimuli at high frequencies (6 kHz, 8 kHz, 10 kHz, and 12 kHz). Bias in judgements for elevation up-down, and front or back as a function of center frequency of the narrow band noise was studied. The localization biases were found to be independent of the target direction. These results [109, 120] indicate that directional biases can result from stimuli, mimicking certain spectral features of the pinnae.

Hebrank *et al.* [9] identified directions for which the subjects’ judgements were biased towards a particular direction. They carried out localization experiments with high-pass low-pass, band-pass and band-stop filtered white noise as stimulus and observed that the sound spectra from 4 to 16 kHz are necessary for sound localization [9]. This observation

CHAPTER 2. BACKGROUND

Table 2.3: Directional cues reported in [9]

Front SC	Overhead SC	Rear SC
1-octave BW notch, with a lower cut-off between 4 and 8 kHz.	1/4 - octave BW peak between 7 and 9 kHz.	Small peak between 10 and 12 kHz.
Increased energy above 13 kHz.		Decrease of energy above and below this peak.

implies that the sound source must be broadband for the listener to localize accurately. The directional cues for the frontal, overhead and rear directions as reported in [9] are summarized in Table 2.3. Frontal directions were mainly cued by a 1-octave BW notch with low cut-off frequencies between 4 kHz and 8 kHz [9]. In addition, peaks were found below and above those frequencies with an increased energy above 13 kHz. An increase in elevation in this direction is cued by an increase in the lower cut-off frequency of the 1 octave BW notch mentioned above [122]. The cues for overhead directions were found to be a 1/4-octave BW peak between 7 kHz and 9 kHz. Blauert also found peaks centered around 8 kHz as an important cue for directions above. Rear directions were cued by a small peak between 10 kHz and 12 kHz and decrease of energy above and below this peak. All these results were consistent with Blauert’s directional and boosted bands. Hebrank and Wright [9] also hypothesized that the spectral notches, peaks were formed by the reflections from the back wall of the concha and validated it mathematically by calculating the time delay and center frequency of the notch.

Other psychoacoustical studies have also demonstrated that the high-frequency pinna cues are necessary for accurate localization [9, 115, 124]. These spectral cues along with the robust ITD and ILD facilitate the human auditory system to localise the sound in the 3D space [99, 125, 126].

Butler and Belendiuk [105] conducted experiments, where sound source localization performance in the median sagittal plane were tested under different conditions for real and virtual sound scenes (non-individualized binaural audio playback over headphones).

CHAPTER 2. BACKGROUND

Broadband stimuli were used as test signals in this experiment. They conducted 1/3-octave band measurements centered at 4 kHz, 5 kHz, 6.3 kHz, 8 kHz, and 10 kHz. It was found that as the source moved up in elevation, a notch that coded the frontal region in the median plane moved up in frequency and became narrower in bandwidth. Researchers also observed that subjects did not always perform better using their own individualized recordings. Moreover, all the subjects performed well with particularly two recordings and at chance level with one of the sets of recording. Their results were similar to that of Hebrank *et al.* [9] indicating an elevation dependent notch that shifts with change in elevation. An important point to take note in this experiment is that the microphone frequency response rolled off at 9 kHz; however, its perceptual effect was not reported in literature.

Bloom [110] studied the hearing monaural sensitivity curves and found that the external ear produces specific distinguishable features in the sound spectra that depends on elevation. These spectral features were appropriately modified, and electronically implemented to create an illusion of continuously variable source elevation. The notches in the simplified versions of HRTFs were studied and emulated using a 1-octave band filter centered at 8 kHz with a varying f_c .

Humanski *et al.* [119] studied the spectral cues in the sagittal plane with real sources positioned at $\pm 30^\circ$, $\pm 20^\circ$, $\pm 10^\circ$, 0° in the sagittal plane at $\pm 45^\circ$, $\pm 90^\circ$ and $\pm 135^\circ$ azimuth. Noise signals with 1 kHz bandwidth, which centered from 4 to 15 kHz in 1 kHz steps, were delivered at a fixed intensity from each loudspeaker that was recorded using a miniature probe microphone placed at each ear canal entrance. The analysis was carried out using two types of spectral cues known as overt and covert cues, in which one or both cues could contribute to the localization accuracy in the sagittal plane. Overt features were those that could be identified in a single HRTF signal as its maximum or minimum. Covert features were obtained from comparison of HRTFs across directions. The quality

CHAPTER 2. BACKGROUND

of overt and covert features (peaks and dips) and their relationship to localization was investigated using a high-pass filtered noise with a cut-off frequency at 4.5 kHz. The overt and covert features (peaks and dips) were accounted well for localization with the ipsilateral ear. When the ipsilateral ear was occluded, the covert peaks contributed to the localization with the contralateral ear, which makes the covert peaks as the most critical cue. These results do have a similarity with Blauert's results as both support an increased energy in certain frequency bands as a cue to elevation. Macpherson [112] studied the notch patterns in 6 subjects and predicted the elevation perception on the basis of f_c . Macpherson *et al.* [112] concluded that more spectral features would act as cues apart from the notches alone.

Roffler *et al.* [124] also found that localization responses were not according to their actual position, but based on their spectral features. They conducted localization experiments for six subjects using real sources to study the factors that influence the localization of sound in the vertical plane. The loudspeakers were kept at four locations: -13° , -2° , -9° and -20° in the vertical plane. These experiments employed either noise or tone as raw stimuli both of which were of 10-ms duration repeating 6 times/sec with a 5-ms rise-fall time. Stimuli comprised broadband noise, low-pass filtered noise (with cut-off at 2 kHz), high-pass filtered noise (with cut-off at 2 kHz and 8 kHz), 600 Hz tone burst, and 4.8 kHz tone burst. Subjects were unable to localize correctly either the low-pass or the tonal stimuli. The 600 Hz tone tended to be localized around -2° and the 4.8 kHz tone around -11° . The low-pass filtered noise tended to be located between -10° and -3° . All listeners accurately localized other stimuli at each of the four loudspeaker locations. In another experiment, Roffler and Butler [127] observed from localization experiments that listeners judged the location of tone bursts in the vertical plane, tended to place the stimuli on a vertical scale in accordance with their respective pitch. Higher-pitched sounds were perceived as originating above lower-pitched sounds. Similar to the experi-

CHAPTER 2. BACKGROUND

ments in [109, 120], Roffler and Butler [127] found that tones were not localized according to their actual position but instead they are localized according to their spectral content. Roffler and Butler [124] strongly suggested that in order to localize in the vertical plane correctly; the stimulus must be complex, must contain frequencies above 7 kHz and the listener must have the pinnae un-occluded.

Some researchers [128, 129] studied the detection thresholds of peaks and notches. It was seen that the spectral notches were perceptually less salient than corresponding peaks. Several studies have been carried out to understand the influence of the anatomical features of the external ear. Gardner *et al.* [115] performed localization experiments with a series of loudspeakers with and without the pinna occluded. Stimuli used were filtered broadband noise of 1/2-octave narrow band and full octave bandwidth centered at 2, 3, 4, 6, 8 and 10 kHz. They found that each of the three main parts of the pinna: Scapha, Fossa and Concha are important for accurate localization. The occlusion of concha degraded the localization performance the most. They found that the localization ability improved if the stimuli had high-frequency content. The localization performance was best for broadband stimuli regardless of the degree of pinna occlusion. The performance degraded with increasing occlusion of the pinna cavities. Localization performance with narrow-band noise stimuli (center frequency 8 kHz to 10 kHz) was similar to that with broadband noise stimuli without pinna occlusion. Overall trends in the rear direction were similar except that the localization performance was in general poorer compared to frontal directions.

Furthermore, a 1/2-octave narrow band signal might be sufficient to allow pinna filtering the slopes and center frequency of the frontal notch even though Hebrank *et al.* stated that the notch was 1-octave band [9]. In another experiment, Gardner [130] conducted similar experiments with real sources (9 loudspeakers) in the frontal hemisphere of the medial plane under conditions of different levels of pinna occlusion: no occlusion of pinna

CHAPTER 2. BACKGROUND

cavities, occlusion of only one pinna cavity, occlusion of both cavities. The performance was worse with pinna occlusion and best with no occlusion. Similar to the observations in Gardner and Gardner [115], it was seen here that the localization performance was best with a broadband noise stimuli and degraded for a narrow band noise stimuli with lower cut-off frequency. It was concluded that cues required to localize accurately in the median plane are mainly monaural, however, binaural interaction can improve the performance.

Butler and Planert [131] performed similar experiments to study the influence of stimulus bandwidth on localization of sound in space. In this experiment, the sound sources were placed at various positions in the median plane in order to minimize the binaural cues. To completely eliminate the binaural cues, the left ear was occluded. Localization experiments were carried out with real sources with five loudspeakers positioned from -30° to 30° in 15° steps. Broadband noise was the baseline condition, and the band-limited signals were centered at 8 kHz, with bandwidth changing from 1 kHz to 6 kHz. Results showed that the best performance was for the binaural condition with a broadband noise as the stimulus. Localization performance was found to be still good in binaural condition for signals of bandwidth 4, 5 and 6 kHz. However, localization performance for monaural condition was poor in comparison with binaural condition. Hofman *et al.* [132] changed the shape of the ears by inserting plastic molds in the pinna cavity and measured the localization performance. Elevation localization was severely distorted soon after the modification; however, accurate localization was acquired again after some training. All the above experiments indicate that pinna plays a critical role in localization of sound source in the real auditory space. In the next section, the binaural playback over headphones is discussed in detail.

2.5 Binaural playback over headphones

The primary focus in this thesis is on the spatial sound reproduction over headphones. Playback through headphones is most suitable because it allows independent control of the acoustic pressure exciting the left and right eardrums. However, it lacks the mechanical vibrations perceived by the body in a natural listening condition. It is intended to reproduce the binaural signals recorded or synthesized at the recording point of the listener's ears as accurately as possible. Nevertheless, this is not possible in practice with standard headphones, or earphones without disturbing the acoustic wave propagation, since the emitter is not acoustically transparent. Binaural playback over headphones comes with other inherent issues, such as in-head localization and coloration of the source spectrum. A highly accurate 3D audio synthesis is possible only when we have individualized HRTFs and individualized equalization, which are tedious to measure [133].

2.6 Headphone equalization

Headphone equalization system is an integral part of a 3D audio reproduction system. Headphones distort the input sound spectrum, as they not only color the input sound, but also affect the free-field characteristics at the ear. Typically, the HPTF comprises of the headphones transducer response and the acoustic coupling between the headphones and the listener's ears. Thus, the headphones need to be calibrated to remove the distortion created by the headphones so that the spectrum at the eardrum is the same as the HRTF filtered signal. In order to compensate for the headphone response, the HPTF is first measured at either the blocked ear-entrance or at the eardrum [47]. It is not a necessary requirement that the "reproduction point" (i.e. the point where the binaural signals are played back) is the same as the recording point. However, the location mismatch between

CHAPTER 2. BACKGROUND

the reproduction and the recording point has to be compensated [134].

The HPTFs, similar to HRTFs, are highly idiosyncratic since the HPTF is also a strong function of the pinna. To compensate for the HPTF response correctly for an individual, one needs to measure the headphone transfer functions at the eardrum or the blocked ear-entrance of the individual. The headphone responses are highly individualized since the frequency response involves the headphone-ear coupling [47]. Møller *et.al* [47] measured the headphone transfer functions (at the blocked ear canal) of 40 individuals for 14 different headphones and found that the inter-subject variations at the high frequencies are much higher compared to the low-frequency region. Blocked ear-canal measurements tend to have lesser inter-individual variations compared to the open ear-canal measurements. Pralong and Carlile [135] obtained the HPTFs of the left and right ears of 10 subjects with a Sennheiser HD 250 circumaural headphone. It was found that considerable variability in the responses occurred at frequencies above 6 kHz, with an inter-individual standard deviation peaking upto 17 dB at frequencies around 9 kHz for the right ear. The inter-individual differences are also found to be similar to that in the HRTFs. Since, the high-frequency spectral cues are extremely important for localization perception, individualized equalization is necessary to ensure that headphone does not degrade the spatial perception. The effect of headphone equalization on distance perception is explained in Chapter 5 of this thesis.

HPTF also varies considerably every time the headphone is removed and repositioned. Kulkarni *et al.* [27] obtained 20 transfer function measurements on the KEMAR manikin with the removal and repositioning of the headphone after each measurement. It is observed that the spectral features in the headphone transfer functions are comparable to those observed in HRTFs. Variation of the responses at the low frequencies are lower than that at higher frequencies, since the pressure at low frequencies within the earphone cavity can be approximated to be the same everywhere. Because of this variation in the

measurement of the headphone responses, effective compensation is not possible even with individualized equalization and the spectrum at the eardrum becomes unpredictable. It is not practically feasible to completely account for the variances due to repositioning. The mean response (across the different measurements) is still not adequate because the mean inverse filter cannot equalize the filter functions completely. It is observed that the headphone positioning is better reproducible, if the listener is allowed to position the headphone in a best comfort position [47]. Interestingly, it has been pointed out that the variance produced by the spectral artefacts due to repositioning is lower than the variance of the spectral features useful for the SC [136]. In general, the headphone equalization accounts for the compensation of (Fig. 2.6) a) The frequency response of the acoustic emitter of the headphones. b) The acoustic coupling between the headphones and the listener's ears.

In addition, the equalization may also compensate for the frequency response of the microphone and the mismatch of the radiation impedance looking out from the ear-entrance. The equalization techniques can be basically divided into two categories, namely, non-decoupled and decoupled equalization techniques (Table 2.4) [7].

2.6.1 Non-decoupled equalization technique

Non-decoupled equalization is most commonly used, when the recording or the synthesis is binaural. In this method, HPTF is measured with the same measurement setup used for the recording including the type of head and the microphones used for recording [7](Fig. 2.6). The aim of non-decoupled equalization technique is to ensure that the spectrum at the eardrum is the individualized HRTF features. The non-decoupled equalization technique can basically be divided into categories based on the desired target response:

- Conventional equalization: This mode of equalization has a flat spectrum as the

CHAPTER 2. BACKGROUND

Figure 2.6: Non-decoupled mode of equalization. BIR denotes either a HRIR or a binaural recording. 'M' denotes the contribution of the microphone and ear canal to the measurement that does not depend on the direction of incidence. Note that * represents convolution and % represents deconvolution. This figure has been adapted from [7].

target curve. This is typically used for conventional headphones. The non-uniform transfer function is compensated and the spectrum of the HRTF is preserved at the eardrum. One of the major disadvantages of this technique is that the equalization filter is subject-dependent.

- **Type-2 equalization:** Type-2 equalization technique is proposed in this thesis for the playback from the frontal projection headphones. Unlike the conventional mode of equalization, the target curve is not a flat curve. In Type-2 equalization mode, the aim is to remove only the spectral distortion by the emitters and the resonant

CHAPTER 2. BACKGROUND

modes of the earcup, while preserving the important frontal pinna cues. The most important advantage of the Type-2 equalization technique is that it is independent of the idiosyncratic features of the ear.

In general, if the coupling of the headphones to the ear adequately emulates free-field conditions, this technique has the effect of eliminating the microphone transfer functions as well as the headphones. This method works well if the recording and HPTF measurements are carried out on the same dummy head, since the microphone positions remain fixed inside the ear canal for both cases. However, in the case of human subjects, it is not possible to avoid a mismatch between the microphone positions, thereby leading to equalization errors. The Type-1 (conventional) and the Type-2 equalization are explained in detail in Chapter 4.

2.6.2 Decoupled equalization technique

Decoupled equalization technique can be used in both binaural as well as stereophonic cases (Fig. 2.7). In this technique, typically the recording (or the HRTFs) as well as the headphone is equalized using a reference sound field (free-field, diffuse-field etc.) [7]. The recording system and the reproduction system are equalized in independent sessions using different equipments so that the effect of the microphones and ear canals are eliminated “locally” in each session. If the reference sound field (REF) of the recording environment is well known and reproduced reliably, this method of equalization can result in a very natural perception of sound similar to the non-decoupled equalization technique. This method of equalization is mainly carried out to make the binaural recordings compatible with stereophonic (conventional microphone) recordings.

If the recording is binaural, then a reference-field equalized binaural recording (HRTF/REF) achieves a sound quality equivalent to a conventional microphone recording. When the equalized recording is played from a reference field equalized headphone (HPT-

CHAPTER 2. BACKGROUND

Figure 2.7: Decoupled Equalization. The recording and the reproduction systems are equalized separately with a common reference field. Thus, it allows elimination of the contribution of the transducers from the recording system. This method is a common choice when the measurement/recording equipments are different from the listening session. The different types of reference fields used are given in Table 2.2. This figure has been adapted from Larcher *et al.* [7].

F/REF), the spatialized sound is reproduced with its most natural timbre [7, 42]. The most common digital media content is typically mixed in stereo format and the reference field for the recording is in general not known. Headphones have to be designed to have a reference field target response in order to have an optimal sound quality for traditional stereo recordings. The reference fields are as follows:

CHAPTER 2. BACKGROUND

Table 2.4: Equalization techniques for different playback modes (Binaural and Stereophony)

Mode of equalization	Aim	Types of equalization and Target response	Characteristics
Non-decoupled (Binaural)	Spectrum at eardrum is the individual HRTF features	Conventional equalization (flat target response)	For conventional headphones. The spectrum at the eardrum has individual features (if individualized HRTF is used). Dependent on the individual's unique pinna features
		Type-2 equalization [144]	For frontal projection headphones. The spectrum at eardrum automatically models the individual pinna spectral cues. Removes only the distortion due to the headphone emitter. Independent of the idiosyncratic features of the ear
Decoupled (Binaural, stereophony)	Emulate the most natural reproduction closer to perception in a reference field	Free-field equalization (FF) [47]	Target response is the free-field response corresponding to the frontal incidence
		Diffuse-field equalization (DF) [47])	Target response is the diffuse-field response Lesser inter-individual variability
		Diffuse-field target response based on Møller [47]	Target response based on average of HRTFs between ± 45 degrees azimuth and elevation with unequal weighting
		Diffuse-field target response based on Lorho [139, 141]	Reduced a 3 kHz peak from about 12 dB to 3dB of diffuse-field response
		RR.G and RR1.G [139]	RR.G: Based on the impulse response of Harman Reference Listening Room RR1.G has lesser bass and treble

Free-field equalization (FF): The design goal of free-field equalization is identical to the HRTF of frontal incidence and aims to replicate the ear signals produced by the loudspeaker in front of the listener. Thus, this equalization reproduces sound sources coming from frontal directions in free-field anechoic conditions with its true timbre. However, if the sound does not come from the front, a free-field equalized dummy head microphone will modify the spectrum of the sound. An important application is the use of dummy heads as recording microphones for loudspeaker reproduction. When a FF equalized dummy head is used, frontal sounds will sound

CHAPTER 2. BACKGROUND

as if they were recorded using a conventional stereophonic microphone technique. However, the non-frontal sounds might sound altered. Theile [137] reported that listeners prefer a diffused-field equalized system if the recorded sound contains significant amount of non-frontal sounds. Hammershoi *et al.* [138] proposed an FF equalization curve, which has additional high-frequency energy above 3 kHz to approximate listening to stereo loudspeakers in the free-field [139]. A free-field equalized headphone can reproduce a frontal sound with natural sound quality but colors the sound spectrum when the sound does not originate from the front. Moreover, it is important to note that there is a large inter-individual variation in the free-field equalization filter [140].

Diffuse-field equalization (DF): In this case, the target response for the equalization of the headphone is the diffused-field response, which is computed by taking the average of HRTFs of all measured directions in horizontal plane. The effect of inter-individual features is reduced drastically due to the averaging effect [140]. Thus, the DF target response can be achieved universally over a large number of individuals. Møller [47] identified certain headphones that were diffused field equalized and recommended such type of headphones for stereo listening.

Other target responses: A typical listening room is not entirely a diffuse-field but the acoustic space of a listening room can be considered somewhere between a free-field and a diffused field. Møller [140] also illustrated other alternative target responses, which are partially diffuse by giving unequal weighting to different directions within $\pm 45^\circ$ azimuth and elevation. Lorho [139, 141] modified the equalization filters with the help of certain parametric filters and found that the subjects generally preferred the target response with a 3 kHz peak lower in amplitude than in the diffuse field response for both music and speech. Other recent experiments [140–143] also showed that listeners prefer other alternative target responses than the conventional

FF and DF equalization.

Decoupled equalization technique also helps to improve the fidelity when the recording is carried out on a dummy head, which is different from the listener's head. The decoupled method of equalization can also be seen as means of adapting the equalization of the headphones to the listener's head, irrespective of the head used for recording. Assuming that there are two binaural recordings BIR_1 and BIR_2 recorded on different heads, the goal here is to approximate the signal BIR_2 that would be captured in the listener's ears as follows:

$$BIR_2 \simeq BIR_1 * REF_2 \% REF_1, \quad (2.14)$$

where REF_1 and REF_2 are the two reference fields, and $\%$ corresponds to the convolution and deconvolution operators, respectively. Diffused field equalization technique is a robust technique for improving the consistency of the binaural recording for the whole population. The inter-individual differences are reduced due to diffused-field equalization. The equalization process removes all the individual characteristics that do not depend on the incidence direction like the effect of the position of the probe microphone placed at the ear canal.

In the next few sections, a brief overview of the types of headphones and their theoretical modelling is presented.

2.7 Circumaural headphones : A theoretical model

In this section, circumaural headphones are modeled using a lumped-element technique [301] [302] [303]. Headphones differ from loudspeakers in the manner they produce sound at the ear. In the case of loudspeaker playback, the ear is immersed in a propagating sound field, while in headphone playback, it registers the SPL in a leaky pressure

CHAPTER 2. BACKGROUND

Table 2.5: Analogies between Electrical, Mechanical, and Acoustical domains

Energy type	Electrical (V,I)	Mechanical (F,v)	Acoustical (P,U)
Dissipative	RI	R_mv	R_AU
Potential (Spring)	$\frac{1}{j\omega L}I$	$\frac{k}{j\omega}v$	$\frac{C_A}{j\omega}U$
Kinetic (Inertia)	$j\omega L$	$j\omega Mv$	$j\omega mU$

Figure 2.8: Examples of closed-back, open-back and the open headphones.

chamber. The sound field of headphones is confined to a relatively small volume of approximately 30 cm^3 .

In a typical headphone working principle, the voltage excites the voice coil (Electrical) which in turn generates the membrane vibrations (Mechanical). These vibrations displace the air giving a perception of audible sound (Acoustical). The analogies between different domains is shown in Table 2.5. Headphone consists of different components such as a cavity, acoustical bottle necks, compliant membrane, Radiation impedances,mechanical-acoustical interface, electrical-mechanical interface. Each of these building block can be represented in electrical or acoustical domains (Appendix D).

One of the deciding factors for the choice of headphones (Fig. 2.8) is the extent of headphone leakage and its effects. Headphones do not rigidly attach to the side of the head leading to leakage of sound. The leakage can be uncontrolled as sound can leak through hair/ or through porous cushions or through uncontrolled porosity in squashed

Figure 2.9: Simple model for an open-back headphone. Figure adapted from [302].

Figure 2.10: Acoustic circuit analogue for an open-back headphone.

foam [301]. The main advantage of an open-back circumaural headphone is that the leakage can be controlled using the rear vents.

Figure. 2.9 shows a simple model for an open-back headphone. V_F and V_B represent the front cavity volume and the rear vent cavity volume, respectively. $Z_{A,0}$ and $Z_{A,T}$ indicates the radiation impedances at the entrance of the earcanal of length l and at the eardrum, respectively. The vent with surface area S_2 represents the variable acoustic leakage. The open-back design lacks low-frequency emphasis unlike a closed-back design

but the key feature about these open-back designs is the lesser significant effect on the pressure at the eardrum. The acoustic circuit analogue for the open-back headphone is shown in Figure.2.10. C_{AB} and C_{AF} is acoustic backload and acoustic frontload, respectively. M_{A1} and R_{A1} represents the fixed pre-determined rare vent leakage path in open-back headphones. The uncontrolled acoustic leakage can be modeled as a lumped mass M_{AL} . As a result of controlled leaks, the sound pressure level (SPL) in the cavity becomes more stable with respect to additional chance leaks. The two are like resistances in parallel, whichever is lower will dominate.

The lumped-element model for the circumaural headphones is explained in detail in Appendix D.

2.8 Open headphones

One of the essential conditions for binaural playback is that the headphone should be an “open headphone”. Møller [134] mentioned that the main requirement for a headphone to be called “open” is that the radiation impedance, as seen looking out from the ear-entrance while wearing headphones ($Z_{headphone}$), matches the impedance seen in the free-field listening condition ($Z_{radiation}$). Møller [134] also observed that $Z_{headphone}$ is much smaller than the ear canal impedance at low frequencies below 1 kHz. In such a case, using an open headphone can be beneficial, since it does not need to cancel the radiation impedance mismatch. This has been explained in detail in Section 4.5 of Chapter 4. Møller measured the pressure division ratio (PDR), which compares the sound pressure at the ear-entrance with and without the headphones [47]. A headphone is called a free-air coupling equivalent (FEC) headphone if the PDR equals one. For the rendering of binaural signals, it is preferred that the headphone is a FEC headphone. It is important to note that the term “open headphones” here is different from the terminology of the “open headphones” in the commercial headphone market, which primarily refers to the

“open-back headphones”.

2.9 Drawbacks of using non-individualized HRTFs

As mentioned earlier, measurement of individualized HRTFs is a highly tedious process. Hence, it is extremely convenient to use the publicly available non-individualized HRTF databases. The price to be paid for the use of these HRTFs is that the perceived spatial image is highly distorted. This is especially true when the binaural playback is over headphones. Numerous experimental results indicate that several subject-dependent errors exist. Front-back confusions, elevation localization errors, in-head localization, and timbral coloration are some of the most common problems associated with non-individualized binaural playback.

The interaural cues (ITD and ILD) determine only the cone of confusion rather than a well-defined spatial position of the sound sources. The interaural cues remain essentially the same for symmetrically frontal and rear directions. The high-frequency spectral cues introduced by the pinnae, and the dynamic cues introduced by head-movement, are important cues that disambiguate the front-back confusions and improve vertical localization. Lack of dynamic head movement cues makes the role of pinna-relevant high-frequency spectral cues extremely critical. These high-frequency spectral cues are highly idiosyncratic and are prone to impairment either by the use of non-individualized binaural recordings or even non-individualized headphone equalization. Additionally, high-frequency errors in HRTF and HPTF measurements along with the errors in every step of the signal processing chain may lead to incorrect high-frequency spectral cues.

Wightman and Kistler [55] studied the errors in headphone reproduction using both individualized HRTFs and individualized HPTFs and compared it to that with the real sources. They observed in this study that the front-back confusion rates almost doubled (11%) for the virtual sources compared to that with real sources (6%). Bronkhost [56]

CHAPTER 2. BACKGROUND

also investigated the localization of real and virtual sound sources using two different experiments. Subjects were asked to turn their heads and respond to the direction when they faced the source in the first experiment. In the second experiment, subjects had to indicate the quadrant of the source. Harmonic signals with a fundamental frequency of 250 Hz and with an upper frequency ranging from 4 kHz to 15 kHz were used as the stimuli in their experiment. Bronkhost [56] found that the localization performance for real and virtual sources were similar when the stimuli did not contain frequencies above 7 kHz. However, it was observed that when the frequencies up to 15 kHz were included, the performance for the virtual sources was in general much poorer than that with real sources. This difference in the performance was associated with the high-frequency distortion of the HRTF spectral cues introduced by the microphone measurement in the ear canal. However, there are other studies that reported equivalent localization performance of virtual sources as in free-field [145]. In the study conducted by Martin *et al* [145], the measurements were carried out using the blocked-ear canal technique, while a probe microphone was used in the studies conducted by Wightman and Kistler [55] and Bronkhost [56].

Blauert [11] mentioned that the correct frontal perception is the result of a combination of a proper loudness balance between the different frequency bands. Researchers have also tried to amplify the spectral difference between the front and rear directions to enhance the perceptual differences between the sounds coming from these directions [146].

Wenzel *et al.* [106] conducted a detailed study on the localization using non-individual HRTFs. Their studies suggested that localization performance with non-individual HRTFs depended highly on the individual, but the localization was degraded for most of the subjects. The azimuth judgment was very much accurate, while the elevation and frontal localization was the most affected in this case. The rates of front-back and up-down confusion in headphone-reproduction using non-individualized HRTFs increased from 19%

CHAPTER 2. BACKGROUND

and 6% to 31% and 18%, respectively. The externalization of the frontal sound source was poor and was most likely perceived inside the head. In addition, the sound sources appeared to be blurred when non-individual HRTFs were used. Use of dynamic HRTFs with the help of head tracking, which is computationally very expensive, improved the localization performance but the confusions were still not completely resolved. This is because the pinna is one of the most individualized elements of the morphology of the listener's ear and thus, the individual spectral cues are highly disturbed by the use of the non-individual HRTFs. Similar results were obtained in the experiments conducted by Begault and Wenzel [108]. Pralong and Carlile [135] argued that the localization errors observed by Wenzel *et al.* [106] are mainly due to non-individualized headphone equalization. Møller *et al.* [133] compared the performance of individual and non-individual binaural recordings and found that non-individualized binaural recordings increases the errors for sound sources in the median plane.

Moreover, with non-individualized recordings and equalization, the spatial image of the sound collapses to inside of the head, thereby lacking externalization [98, 146, 147]. IHL is also known by other names, such as intracranial lateralization or lateralization. Weinrich [95] reported that the optimal externalization is often dependent on the correct ratio between ITD and ILD that is frequency dependent, location dependent, as well as idiosyncratic. Hartmann *et al.* [148] carried out psychophysical experiments to study externalization. Externalization was found to depend on the interaural phases of low-frequency components with a boundary near 1 kHz. However, ILDs in all frequencies were found to be important for externalization perception. Most importantly, accurate spectral cues in the two ears are a necessary condition to avoid IHL of sound. Plenge [149] argued that lateralization is caused by incorrect spatial information on both ears and that externalization does not depend on any kind of electroacoustic transmission. Blauert [11] attributed externalization perception to several reasons:

CHAPTER 2. BACKGROUND

1. Differences in sound transmission between real life listening.
2. Unnatural resonances of the headphone cavity.
3. Unnatural loadings of the eardrum.
4. Invariability of the headphone signals when the head is moved.
5. Bone transmitted sounds.

Kim *et al.* [28] conducted subjective experiments studying the effect of individualization of HRTFs and equalization of headphones on the degree of externalization in terms of distance of auditory perception. Their results indicated that individualized equalization of the headphones is important for externalization and individual synthesis is important for consistent distance perception. Durlach *et al.* [150] reported that non-individualized HRTFs and incorrect headphone equalization often accounts for IHL perception. Wightman and Kistler [55] found that individualized HRTFs and equalization could result in a very realistic auditory experience similar to the free-field real sources. Other researchers [150–153] found that dynamic cues caused by head movements are considered essential for externalization. Environmental reflections, such as the presence of reverberation, are extremely important for externalization of the perceived virtual source. Begault [154] verified that incorporating room reflections into non-individualized binaural recording can reduce IHL. In another study, Begault *et al.* [108] carried out a direct comparison of the dynamic cues, reverberation and individualized binaural synthesis on the spatial perception of a speech source. They observed lower azimuthal errors and higher rate of externalization (a ratio of 2:1) with reverberation. However, contrary to other studies, it was observed by Begault *et al.* that use of individualized HRTFs or head tracking did not have a prominent influence on externalization. In an interesting study by Brookes and Treble [155], it was shown that the externalization can be tremendously improved by using asymmetrical pinna for the left and right ears.

2.10 Summary

In this chapter, we have introduced the basic principles governing the synthesis and playback of binaural audio. Binaural audio can either be synthesized by convolving the mono source with the HRTFs, or recorded at the ears (eardrum or the blocked ear-entrance) of the dummy head or human subject. HRTF encapsulates the important interaural cues, and the spectral cues that are necessary to localize the source accurately. The spectral cues that are generated due to the interaction of the incident acoustic waves with the pinna are highly idiosyncratic, and thus, the HRTFs are also unique.

The main challenge in rendering binaural audio is the degradation of the spatial image either due to the distortion of the individualized HRTFs or by the use of non-individualized HRTFs. Front-back confusions, IHL, localization blur, and timbral coloration are common, especially with headphone playback of non-individualized binaural audio. In order to ensure an accurate spatial perception, individualized HRTFs and individualized equalization should be carried out. However, measurement of individualized HRTFs and HPTEs is extremely tedious and time consuming. Thus, there is a strong need for individualizing the non-individualized HRTFs in order to improve the accuracy of the rendered spatial image. In the next chapter, procedures for the measurement of HRTFs and the measurement and compilation of the DSPLab HRTF database are explained in detail.

Chapter 3

Measurement of Head Related Transfer Functions

In this chapter, the procedure for the measurement of HRTFs and computation of a set of HRTFs at varied distances is explained in detail. Though there are several publicly available databases for HRTFs measured in the far-field, there are very few databases that contain HRTFs measured in the near-field [76] [77]. Moreover, the density of measurements in the near-field in these databases is low. The DSPLab HRTF database contains HRTFs measured at 8 distances in the near-field. The measured HRTFs are stored in the DSPLab HRTF database, which is made available to the public for free. The HRTFs are measured in this study for the following reasons:

1. HRTFs contain the information of the acoustic sound propagation between the sound source and the two ears, and thus, form an essential part of 3D sound synthesis.
2. The variation of HRTFs among different individuals and their perception can be studied.
3. The modeled HRTFs using frontal projection headphones in this thesis can be

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

compared with the measured HRTFs using objective and a subjective analysis.

4. The distance-dependent HRTFs database for 75 azimuthal directions, 13 elevation angles, and 8 distances (7,800 directions) measured on the B & K HATS dummy head is released online in spatially oriented format for acoustics (SOFA) format. The DSPLab HRTF database can be downloaded at <http://eeeweba.ntu.edu.sg/DSPLab/DspLabHRTF/>.

3.1 System identification methods

Given a continuous or a discrete-time LTI system, the impulse response $h(t)$ or $h(n)$ or the transfer function $H(f)$ of the system has to be determined in order to completely identify the system. Here, the variables t , n , and f represent time as a continuous variable, time in samples, and frequency, respectively. There are three commonly used system identification methods:

1. **Impulse method:** This method is the most straightforward method to obtain the system transfer function. Here, the system is directly excited by a Dirac delta impulse function $\delta(n)$, and the impulse response function of the system $h(n)$ can be obtained as the output.
2. **Fourier analysis method:** In this method, the transfer function of the system $H(f)$ is obtained by calculating the ratio between the output and input frequency responses. However, this method is generally suitable only for deterministic input signals with finite energy:

$$H(f) = \frac{Y(f)}{X(f)}. \quad (3.1)$$

where $Y(f)$ and $X(f)$ are the output and input signal spectrum, respectively.

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

3. Correlation analysis method: This method is suitable for both deterministic and random stationary input signals (White gaussian noise, Maximum length sequence, Chirp). In this method, an input signal, whose autocorrelation function is a Dirac delta function $\delta(n)$, is used.

The normalized N -point autocorrelation function $R_{xx}(n)$ and cross-correlation function $R_{xy}(n)$ are defined as follows:

$$R_{xx}(n) = \frac{1}{N+1} \sum_{q=0}^{N-1} x(q) x(q+n), \text{ and} \quad (3.2)$$

$$R_{xy}(n) = \frac{1}{N+1} \sum_{q=0}^{N-1} x(q) y(q+n), \quad (3.3)$$

where $x(n)$ and $y(n)$ are N -point finite real sequences and the variable n ranges from 0 to $(N-1)$. The cross-correlation function $R_{xy}(n)$ can be represented in terms of the autocorrelation function $R_{xx}(n)$ as follows:

$$R_{xy}(n) = \sum_{q=0}^{N-1} h(q) R_{xx}(n-q). \quad (3.4)$$

Here, if we choose an input signal such that its N -point autocorrelation function is a Dirac-delta function $\delta(n)$, i.e ,

$$R_{xx}(n) = \delta(n). \quad (3.5)$$

Then, the N -point cross-correlation value results in the impulse response function of the system $h(n)$ in question:

$$R_{xy}(n) = h(n). \quad (3.6)$$

The impulse response sequence $h(n)$ can be calculated even if the autocorrelation of the input signals does not correspond to the Dirac-delta function. In such a case, the

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

auto and cross-spectral densities (S_{xx} and S_{xy}) can be calculated by applying N -point DFT to $R_{xx}(n)$ and $R_{xy}(n)$:

$$S_{xx}(k) = \sum_{n=0}^{N-1} R_{xx}(n) \exp(-j \frac{2\pi}{N} kn), \text{ and} \quad (3.7)$$

$$S_{xy}(k) = \sum_{n=0}^{N-1} R_{xy}(n) \exp(-j \frac{2\pi}{N} kn), \quad (3.8)$$

where the variable k ranges from 0 to $N - 1$. The system transfer function in frequency domain is then obtained by dividing the spectral density functions $S_{xy}(n)$ by $S_{xx}(n)$:

$$H(k) = \frac{S_{xy}(k)}{S_{xx}(k)}. \quad (3.9)$$

The impulse response function $h(n)$ is then obtained by taking the IDFT of the system transfer function $H(k)$.

3.2 Measurements of HRTFs

In this section, the principle and design of experiments for the measurement of HRTFs for both dummy head and human subjects is presented. A large variability can be found in the experimental methods to measure the HRTFs in terms of the excitation signal, microphones used, speakers used, dummy head/human subjects, and listening environment, etc. [48, 64, 65, 78, 79, 81–84, 86, 156–159]. The measurement procedures can have an impact on the quality of the HRTFs, thereby affecting the subjective perception of the synthesized binaural audio. Several widely used public databases are available, where the researchers have documented precise details and procedures in measuring the HRTFs that ensure the correctness of the measurement of HRTFs [48, 57, 57, 70, 70]. A simple schematic diagram of the HRTF measurement system employed in this study is shown

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.1: HRTF measurement system employed in this study.

in Fig. 3.1.

In this thesis, HRTF is defined similar to that in [48] as the ratio of complex sound pressures arriving at the left and right ears ($P_{L,R}(\omega)$) to the sound pressure measured at the position of the center of the head (in the absence of the head) in the free-field ($P_0(\omega)$). Note that these pressures are a complex function of the angular frequencies ω , thus $HRTF(\omega)$ is also a complex function of frequencies.

$$HRTF(\omega) = \frac{P_{L,R}(\omega)}{P_0(\omega)}. \quad (3.10)$$

Since $P_{L,R}(\omega)$ is direction dependent and $P_0(\omega)$ being direction independent, $HRTF(\omega)$ also varies with directions.

In a practical measurement scenario, there are several components that do not have a flat transfer function, which might affect the overall measurement of HRTFs. Let us consider these components:

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

- The transfer function of the dummy head microphone, $H_{mic}(\omega)$.
- Free-field microphone, $H_{refmic}(\omega)$.
- Loudspeaker transfer function, $H_{LS}(\omega)$.
- A/D, D/A converter or sound card inside Brüel & Kjær (B & K) PULSE system, $H_{SC}(\omega)$.
- Power amplifier inside B & K PULSE system, $H_{PA}(\omega)$.
- Transfer functions of two B & K measuring amplifiers, $H_{MA}(\omega)$.

The effect of the transfer functions mentioned above can be compensated if the same setup is used for both the measurement of $P_{L,R}(\omega)$ and $P_0(\omega)$. For example:

$$HRTF = \frac{\text{Pressure at the two ears}}{\text{Pressure in free-field}} = \frac{H_{SC} \cdot H_{PA} \cdot H_{MA} \cdot H_{LS} \cdot H_{mic} \cdot H_{L,R}}{H_{SC} \cdot H_{PA} \cdot H_{MA} \cdot H_{LS} \cdot H_{refmic} \cdot H_{reffield}}, \quad (3.11)$$

which simplifies to

$$HRTF = \frac{H_{mic} \cdot H_{L,R}}{H_{refmic} \cdot H_{reffield}}. \quad (3.12)$$

It should be noted that the variable ω has been omitted from the above equations for the sake of brevity. The microphones $H_{mic}(\omega)$ and $H_{refmic}(\omega)$ have a transfer function close to a flat response, thus, the HRTF can be directly calculated as the ratio of pressures at the two ears to the free-field pressure at the position of the center of the head with the head absent.

3.3 Excitation signal

The choice of the input excitation signal is extremely important in the accurate measurement of HRTFs. In general, the excitation signal should satisfy the following condi-

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

tions:

- Excitation signal should contain all the frequency components of interest.
- Uniform energy or power spectrum across all the frequencies.
- The measured impulse response should have a high signal-to-noise ratio (approximately > 80 dB).
- Input signal should have a low-crest factor (ratio of the peak value to the effective value of the signal).
- The excitation signal and the deconvolution technique must aid in eliminating the non-linear artefacts in the deconvolved impulse response.

In general, the signal-to-noise ratio can be improved by taking multiple averages of the measured output signal before start of the impulse response deconvolution process. Some of the most commonly used excitation signals are maximum length sequences (MLS), inverse repeated sequence (IRS), and time stretched pulses and sine sweep.

The MLS and the IRS techniques use pseudorandom noise, while the time stretched pulse and sine sweep use time varying signals. The MLS technique was first introduced in a room acoustic measurement system by Schroeder [160]. The distortion artefacts in MLS technique was reduced theoretically thereby developing the IRS technique [161–163]. Aoshima later introduced a new technique for the measurement of impulse responses that led to the time-stretched pulses technique [164]. More recently, the logarithmic sine sweep technique using a new deconvolution technique was developed by Farina and Ugolotti [165, 166]. In this thesis, logarithmic sine sweep is used as an excitation signal to measure the HRTFs.

One of the main problems of the MLS technique is the time-aliasing error. The use of circular cross-correlation to obtain the impulse response, results in a periodic impulse

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

response. This error becomes significant if the length of one period is shorter than the length of the impulse response to be measured. Another issue with the MLS technique is the appearance of the distortion artefacts also known as the distortion peaks. These distortion artefacts introduce crackling noise when the impulse response is convolved with the signal in binaural synthesis. Optimization of certain measurement parameters are required to reduce the distortion in this technique. The distortion peaks in the MLS technique can be attenuated even by using the IRS method. The time-stretched pulses technique expands the excitation signal in order to increase the signal-to-noise ratio. The distortion peaks in the impulse response are also reduced compared to the MLS technique. The MLS, IRS, and the time-stretched pulse techniques suffer from distortion artefacts when their assumption of being an LTI system fails. The logarithmic sine sweep method however avoids these limitations.

3.3.1 Exponential sine sweep

The exponential sine sweep method has the following advantages over other techniques:

- Better noise rejection.
- Better separation of non-linear effects.
- Avoids time variance effects due to the use of single long sweep.

$$x(t) = \sin \left[\frac{\omega_1 \cdot T}{\ln \left(\frac{\omega_2}{\omega_1} \right)} \cdot \left(e^{\frac{t}{T} \cdot \ln \left(\frac{\omega_2}{\omega_1} \right)} - 1 \right) \right], \quad (3.13)$$

where $x(t)$ is the exponential sine sweep signal, ω_1 and ω_2 are the start and the end angular frequencies of the sine sweep, which last for T seconds.

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

The theory of system identification can be developed by considering a black box with a known input signal $x(t)$ (Fig. 3.2). The output signal $y(t)$ can be represented as a sum of generated white Gaussian noise (WGN) and a deterministic function of the input signal:

$$y(t) = WGN + F[x(t)] . \quad (3.14)$$

If the system is assumed linear and time-invariant, then the function F is the convolution of the excitation signal with the impulse response:

$$y(t) = WGN + [x(t) \otimes h(t)] . \quad (3.15)$$

However, the system is not entirely linear. The non-linear effects due to the loudspeaker transducer often come into picture at the very beginning and the process is substantially memoryless. However, as the sound is radiated into air, the propagation is linear, including multiple reflections, echoes, and reverberation. A more complete representation of the system including the non-linear effects of the loudspeaker is shown in Fig. 3.3. To model the non-linear effects, the input signal is assumed to pass through a memoryless non-linear device characterised by an N^{th} order Volterra kernels ($k_N(t)$):

$$w(t) = x(t) \otimes k_1(t) + x^2(t) \otimes k_2(t) + x^3(t) \otimes k_3(t) + x^4(t) \otimes k_4(t) + \dots + x^N(t) \otimes k_N(t) , \quad (3.16)$$

where $x(t)$ is the exponential sine sweep, $w(t)$ is the distorted signal and ($k_i(t)$) represents

Figure 3.2: A basic input/output system.

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.3: A more complete representation of the system. The loudspeaker contributes to the non-linear effects and the propagation in the air represents a linear system.

the i^{th} component of the Volterra kernel [166]. The response of the global system can be considered as a summation of the additive WGN with the output of the linear system having an impulse response $h'(t)$:

$$y(t) = WGN + x(t) \otimes k_1(t) \otimes h'(t) + x^2(t) \otimes k_2(t) \otimes h'(t) + x^3(t) \otimes k_3(t) \otimes h'(t) + x^4(t) \otimes k_4(t) \otimes h'(t) + \dots + x^N(t) \otimes k_N(t) \otimes h'(t). \quad (3.17)$$

In general, the linear part and the non-linear distortion effects is difficult to be separated from each other, and thus, the deterministic part of the transfer function is described by a set of impulse responses, each being convolved with a different power of the input signal:

$$y(t) = WGN + x(t) \otimes h_1(t) + x^2(t) \otimes h_2(t) + x^3(t) \otimes h_3(t) + x^4(t) \otimes h_4(t) + \dots + x^N(t) \otimes h_N(t), \quad (3.18)$$

where $h_i(t) = k_i(t) \otimes h(t)$ and $y(t)$ is the output of the system that is affected by non-linear distortions.

The distortion peaks can be avoided by substituting the circular deconvolution with a linear deconvolution method implemented directly in the time-domain. This is feasible if an inverse filter function $f(t)$ can be obtained such that its convolution with the input

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.4: An example of an exponential sine sweep. Parameters : $F_s = 44,100 \text{ Hz}$, Start frequency = 10 Hz , Final frequency = $1,000 \text{ Hz}$, Duration = 1 s .

signal $x(t)$ results in a delayed dirac delta functions $\delta(t)$:

$$x(t) \otimes f(t) = \delta(t - k). \tag{3.19}$$

Once $f(t)$ is obtained, the system's impulse response $h(t)$ can be found out by convolving the output signal $y(t)$ and the inverse filter $f(t)$:

$$h(t) = y(t) \otimes f(t). \tag{3.20}$$

The inverse filter can be generated as follows :

1. The logarithmic sine sweep is time reversed and delayed to obtain a causal signal.

This time-reversed signal has a sign reversal in its phase spectrum. The convolution

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.5: Magnitude Spectrum of the exponential sine sweep. Parameters : $F_s = 44,100 \text{ Hz}$, Start Frequency = 10 Hz , Final Frequency = $1,000 \text{ Hz}$, Duration = 1 s .

of the time-reversed signal with the exponential sine sweep results in a purely delayed signal but with the magnitude squared.

2. The magnitude spectrum of the resulting signal is then normalized by dividing it by the square of the magnitude spectrum of the initial sweep signal.

With this approach of calculating the inverse filter that employs circular convolution, the distortion peaks are generated in the output at harmonics frequencies that are higher than the instantaneous input frequency. The main advantage of linear convolution over circular convolution is that the use of linear convolution pushes all the distortion effects much beyond the linear response. This property enables the measurement of the system's linear impulse response even if the loudspeaker operates in the non-linear re-

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

gion. Moreover, linear convolution avoids time aliasing problems. In general, even if the time analysis window has the same length as the sine sweep signal (and is shorter than the impulse response to be measured), the tail of the system response may be lost. But, unlike the MLS technique, the late part of the tail will not fold back at the beginning of the deconvolved output signal thereby avoiding time aliasing. An important advantage of using exponential sine-sweep over linear sweep is the high signal-to-noise ratio with the exponential sine-sweep at low frequencies. Another important result using the exponential sine-sweep method is that if the sweep is slow enough, each of the harmonic distortions packs into a separate impulse response without overlapping with the previous one.

An example of an exponential sine sweep and its magnitude response is shown in Fig. 3.4 and Fig. 3.5, respectively. The magnitude response of its inverse is also shown in Fig. 3.6. There are several parameters that have to be fixed to carry out the measurement of HRTFs using the exponential sine sweep method:

1. Sampling Frequency: 44,100 Hz.
2. Number of averages: 3 (Number of times the sine sweep signal is sent to the loudspeaker and recorded).
3. Recording Mode: 2 Channels (left ear and right ear).
4. Initial frequency (f_1): 100 Hz.
5. Final frequency (f_2): 20,000 Hz.
6. Sweep duration(t_{sweep}): 10 seconds.
7. Sweep rate $s^{-1} (\frac{\log_2(f_2/f_1)}{t_{sweep}})$: 1.066.
8. Silence duration: 0.05 seconds (Duration of silences introduced after each sweep).

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.6: Magnitude Spectrum of the inverse exponential sine sweep. Parameters : $F_s = 44,100 \text{ Hz}$, Start Frequency = 10 Hz, Final Frequency = 1,000 Hz, Duration = 1 s.

3.4 Measurement/Recording point

HRTFs can be measured at any point along the line from the ear canal entrance to the eardrum, since the ear canal can be approximated as a one-dimensional transmission line. Typically, a point closer to the ear canal entrance or the eardrum is chosen as the measuring point for the purpose of standardization. HRTFs of artificial heads are often measured at the location of their eardrums using the fixed pressure field microphones located at the end of the occluded-ear simulator near the eardrum [48]. However, this is extremely tedious especially for human subjects, as tiny movements of probe in the ear canal can cause large inaccuracies at higher frequencies. A probe microphone is often

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

used for the measurement of HRTFs on human heads placed at a position of about 1 to 2 mm from the eardrum, in order to avoid uncertainties from standing wave interference. The measurement position is often difficult to fix in this technique and might result in severe errors in the case of incorrect operation. A probe microphone also suffers from poor frequency response and low-sensitivity. Hiipaka [167] presented a method to obtain individually correct magnitude response of the HRTF at the eardrum from the pressure velocity (PU) measurements at the ear canal entrance with a miniature PU sensor. Other researchers have also tried to model the ear-canal to eardrum response to complete the acoustical path upto the eardrum in the case of blocked entrance measurements [138, 168].

Measurement of HRTFs at the blocked ear canal (blocked-entrance) is a very common technique for the measurement of HRTFs. In the blocked ear-canal position, a pair of miniature microphones with relatively high sensitivity and wide frequency range is positioned near the entrance of the ear inserted in a rigid moulding of the ear canal. Hammershoi *et al.* [138] measured the sound transmission to the eardrum from various points in the external ear using a probe microphone technique. They calculated the standard deviation between subjects for HRTFs measured at the eardrum, the open entrance, and the blocked-entrance and the lowest values were found for the blocked-entrance HRTFs. Researchers found that binaural recordings carried out at the blocked-entrance captured all the spatial information and displays minimum amount of direction independent individualistic variation. The blocked-entrance pressure is more likely to be representative of a population. Furthermore, Møller *et al.* [133] conducted detailed localization experiments comparing the individualized binaural recordings recorded at the blocked-entrance and the real life scenario with loudspeakers. It was found that listeners were able to perceive an authentic reproduction of spatial sound for which the localization performance was equal to that in real life.

The main advantage in the blocked entrance method is the accurate repeatability

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

of the microphone positioning and good signal-to-noise ratio. However, the transfer function does not include the ear-canal response. It was found that the transmission from the free-field to the eardrum can be divided into a directional dependent part and two directional independent parts by recording the response at different points within the human ear canal [138]. As the ear-canal response is independent of the sound direction, blocked entrance HRTFs capture all the spatially relevant features. In this thesis, the measurement of HRTFs for human subjects is carried out in the blocked entrance of the ear.

3.5 Near-field HRTFs : Measurement constraints

HRTFs are most commonly measured in the far-field at a distance greater than 1 m. However, a great amount of recent research has focussed on the near-field HRTFs due to its distance-dependent characteristics. Unlike the measurement of HRTFs in the distal region, the near-field HRTF measurements require the use of omni-directional acoustic point sources to avoid the scattering effects from the source during measurement. Researchers have developed various types of acoustic point sources, like a combination of electro-dynamic driver and tygon tubing [43], micro-dodecahedral loudspeaker [169], micro-dynamic- type loudspeakers [170], and a spark plug [171] to measure the near-field HRTFs. Other researchers, however, used a Bose Acoustimass loudspeaker [172] and a Tannoy loudspeaker [173], instead of a point source to measure the near-field HRTFs. Due to the requirement of an acoustic point source and the difficulty in measuring individualized HRTFs, only few researchers have measured individualized proximal region HRTFs.

For the measurement of near-field HRTFs, ideally a point sound source is required in order to avoid multiple diffractions back and forth between the head and the source. The low-frequency features of the near-field HRTFs are extremely important. If the

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

sound source used in a near-field HRTF measurement has a low-frequency response that drops from approximately 1 kHz, then additional signal processing may be required. However, if the drop is very high, then it might be difficult to compensate for the low-frequency loss. To obtain a better low-frequency response, a speaker of larger volume is required, however, this would lead to multiple scattering between the head and the source. Guang-Zheng *et al.* [174] studied the influence of scattering of the sound source on the measurement of near-field HRTFs. They found that the effect of source scattering in the HRTF is negligible for measurements beyond 0.2 m if the source radius does not exceed 0.03 m.

In our measurements, due to unavailability of a point sound source, the near-field HRTF measurements were carried out using a Cambridge Soundworks mini cube speaker of dimensions less than 0.03 m in radius (Fig. 3.7). The HRTFs were precisely measured at distances of 0.35 m, 0.45 m, 0.50 m, 0.60 m, 0.75 m, 0.80 m, 1 m and 1.4 m from the center of the head. Since all the HRTFs were measured for distances beyond 0.2 m, and the source radius being less than 0.03 m; the effect of source scattering on the measurements were assumed to be negligible [174]. The speakers used have a lower frequency limit of approximately 400 Hz that can be corrected by theoretical calculation.

3.6 Measurement procedure

The non-individualized HRTFs were measured at the eardrum of a Brüel and Kjær (B & K) 4128 head and torso simulator (HATS), which is equipped with the B & K 4158-C (Left) and 4159-C ear simulators. The ear simulator contains a 1/2" microphone connected to a microphone amplifier with an adaptor. In addition, individualized HRTFs were measured at the blocked-entrance of the ear canal [138] using a B & K 4101 binaural microphone for three subjects. The measurements were controlled by a computer located in a control room next to the anechoic chamber. The computer was connected via a LAN

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.7: Setup for the measurement of near-field HRTFs in the anechoic chamber. The speaker used is a Cambridge Soundworks speaker and the dummy head is the B & K HATS dummy head.

cable to the B & K PULSE analyser system, which generates the excitation signal and simultaneously records the responses of the microphone. An exponential sine sweep was the preferred choice of excitation signal in these measurements, since they provided better SNR at lower frequencies [175]. The signals were recorded at the microphones at the recording point (eardrum or blocked-entrance) at a sampling rate of 65,536 Hz. The recorded signal was first down-sampled to 44,100 Hz and then processed using a custom-made MATLAB function to extract the impulse response. The ITDs were obtained by first calculating the lag that maximizes the cross-correlation function between the impulse responses of the two ears within the range ± 1 ms. The interaural cross-correlation coefficient $\zeta(\tau)$ is defined as:

$$\zeta(\tau) = \frac{\int_{-\infty}^{\infty} HRTF_L(\omega) HRTF_R^*(\omega) e^{-i\omega\tau} d\omega}{\left\{ \left[\int_{-\infty}^{\infty} |HRTF_L(\omega)|^2 d\omega \right] \left[\int_{-\infty}^{\infty} |HRTF_R(\omega)|^2 d\omega \right] \right\}^{\frac{1}{2}}}, \quad (3.21)$$

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.8: Free-field response of the speaker for various distances.

where $HRTF_L$ and $HRTF_R$ denotes the HRTFs for the left and right ear, respectively; ζ is the cross correlation function, with $-1 \leq \zeta(\tau) \leq 1$. Prior to the computation of cross-correlation, the impulse responses for both the ears are low-pass filtered with a cut-off frequency of 2 kHz, since ITD is predominantly a low-frequency cue.

HRTFs were measured for a total of 7,800 speaker locations. The measurements were carried out for all locations at 5° resolution in azimuthal plane and at elevation angles -40° below the head to 80° (with a resolution of 10°) above the head for every azimuth. These HRTFs were precisely measured at distances of 0.35 m, 0.45 m, 0.50 m, 0.60 m, 0.75 m, 0.80 m, 1 m, and 1.4 m from the center of the head.

For every azimuth location, a set of 10 measurement trials were taken to ensure consistency and repeatability of the measurements. Reference measurements were first carried out in the free-field with a 1/2 " B & K 4134 pressure microphone kept at the

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

position of the center of the dummy head in the absence of the head (Fig. 3.8). The reference measurements were calculated for all the eight distances for which the HRTFs were measured. The magnitude of the HRTF for each location was obtained by the frequency-wise division of the amplitude spectrum measured on the dummy head/human subject at the recording point (eardrum or blocked-entrance) and the free-field reference measurement in order to compensate for the speaker response as well as the SPL drop due to distance. Figure. 3.9 shows the measured HRTFs for the frontal ipsilateral, frontal contralateral and rear ipsilateral locations measured on three different human subjects.

3.7 Calibration

A series of calibration techniques were carried out prior to each set of experiments. The dummy head or human head was first adjusted to point directly to the center of interaural axis with respect to the speaker. Correct placement is extremely important when the source is closer to the head, as small angular deviations in the near-field leads to larger deviations with distance and thus, affects the measurements. With the help of acoustical measurements, it was ensured that the source is placed directly in front of the head (0° azimuth and 0° elevation). The adjustment of the head (dummy/human) position is carried out until the magnitude of the ITD was reduced below $2\mu s$.

3.8 Minimum phase representation of HRTF

The measured raw HRTFs can also be represented and stored as minimum-phase filters. According to principles of signal processing [176], HRTFs, which is a transfer function of an LTI system, can be represented by a product of its minimum-phase function $H_{min}(r, \theta, \phi, \omega)$, all-pass function $\exp[j\psi_{all}(r, \theta, \phi, \omega)]$, and linear-phase function

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.9: Measured HRTFs of Subject K, A and R for the frontal ipsilateral, frontal contralateral and rear ipsilateral locations.

$$\exp [-j2\pi fT(r, \theta, \phi)]:$$

$$HRTF(r, \theta, \phi, \omega) = H_{min}(r, \theta, \phi, \omega) \cdot \exp [j\psi_{all}(r, \theta, \phi, \omega)] \cdot \exp [-j2\pi fT(r, \theta, \phi)] \quad (3.22)$$

$$= H_{min}(r, \theta, \phi, \omega) \cdot \exp [j\psi_{cess}(r, \theta, \phi, \omega)], \quad (3.23)$$

where the function $\psi_{cess}(r, \theta, \phi, \omega)$ includes the linear-phase components, as well as the all-pass components; $T(r, \theta, \phi)$ is the frequency independent pure time delay caused by

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

propagation from sound source to ear corresponding to the initial delay of the HRIRs. The variables r, θ, ϕ in the equations indicates the dependence on distance, azimuth, and elevation angles, respectively. The minimum phase HRTF representation $H_{min}(r, \theta, \phi, \omega)$ can be expressed as:

$$H_{min}(r, \theta, \phi, \omega) = |HRTF(r, \theta, \phi, \omega)| \exp [j\psi_{min}(r, \theta, \phi, \omega)] . \quad (3.24)$$

We can immediately note that the magnitude spectrum of the minimum-phase $HRTF$ is the same as the original HRTF. The minimum phase component contains all the spectral and frequency dependent information that helps in sound localization, while the excess phase component encodes the frequency-independent time information relevant for spatial localization. The minimum phase representation is the shortest FIR filter, which has the same magnitude spectrum as the original filter, and thus results in lower computation and memory usage. Damera-Venkata *et al.* [177] reported that given optimal minimum phase and linear phase filters for the same magnitude response, the minimum phase filters have a reduced length that is one-half or one-quarter times of the length of the linear phase filter. Moreover, minimum-phase HRTFs have most of the energy concentrated at the beginning of the impulse response (Fig. 3.10).

The idea of using minimum-phase HRTF filters in binaural studies is not new and has been used by several researchers [45, 57, 86, 101, 106, 113, 178, 179]. Huopaniemi *et al.* [180] found that the HRTFs were almost minimum phase and used them in the synthesis of virtual sounds due to their ease of implementation. Kulkarni *et al.* [181] conducted theoretical and psychophysical experiments validating that minimum phase representation of the phase spectra, with the frequency-independent, position-dependent ITD, is an adequate representation of the HRTF phase. Minimum phase systems are both causal and stable, and thus all the zeros of the filter are inside the unit circle in the z -plane. The zeros outside the unit circle correspond to the excess phase part of the

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

Figure 3.10: Raw HRIRs and their minimum-phase representations. (a) HRIR of contralateral ear of azimuth 60° and 0° elevation (b) HRIR of contralateral ear of azimuth 80° and 0° elevation

filter. Toledo [114] analysed the number of zeros outside the unit circle for a database of 3880 pairs of HRTFs. It was found that HRTFs showed minimum phase characteristics for most of the directions, except for contralateral HRTFs with angles at 90°, 270° azimuth to the side of the ear. Other researchers [181,182] also found that the contralateral ear HRTFs deviated from minimum phase characteristics at angles near the sides. The validity of minimum-phase representation for distance-dependent HRTFs has not been directly investigated in the literature. However, since distance-dependent HRTFs can be

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

obtained from far-field HRTFs, minimum-phase representation can accurately represent the magnitude characteristics of distance-dependent HRTFs [241] [243]. Moreover, ITDs do not vary much with distance, and thus, the same frequency independent time delay can be used as ITD for varying distances in the minimum-phase representation [43].

There are mainly two methods by which minimum phase HRTF filters are computed: Hilbert transform and Homomorphic filtering. In this thesis, minimum phase HRTF filters are derived using the homomorphic filtering method, which is implemented in the MATLAB function `rceps.m`.

3.9 Creation of HRTF database

The measured range-dependent HRTFs are packed and organized in a special format named the spatially oriented format for acoustics (SOFA). The SOFA aims at representing the spatial data in a standardised manner that allows more flexibility in usage [32]. These files have `.sofa` extension, where the data is serialized into a binary stream with special libraries that ensures data compression, self-describing data and open access. The database is organized as a single file that typically contains the impulse response and all the metadata in object based format. All the measurements are usually stored in a single data structure. We use the SOFA convention “SimpleFreeFieldHRIR”, which defines all the attributes to be used for a typical HRTF database. Matlab files are provided to extract the HRIR files from the SOFA format file. The database provides HRIRs with both 256 samples and 512 samples. More details about SOFA and the database can be found in Appendix A. The entire database is available for download in this link [30,31] : <http://eeeweba.ntu.edu.sg/DSPLab/DspLabHRTF/>

CHAPTER 3. MEASUREMENT OF HEAD RELATED TRANSFER FUNCTIONS

3.10 Summary

In this chapter, the procedures for the measurement of HRTFs were discussed. Various issues related to the accurate measurement of HRTFs were investigated. In this study, an exhaustive HRTF database was created containing HRTFs for source locations at 75 azimuth directions, and 13 elevation angles for each of the 8 distances. The database is then compiled in SOFA format and is made publicly available for the benefit of the entire research community. All the HRTF measurements reported in the thesis are based on the procedure presented in this chapter.

Chapter 4

Individualization of

Non-Individualized Binaural

Synthesis using Frontal Projection

Headphones

Binaural signals are highly idiosyncratic. Lack of the individualized pinna spectral cues in the binaural audio leads to degradation of spatial perception. The errors in non-individualized binaural playback over headphones like the front-back, up-down reversals, in-head localization, timbral coloration have already been discussed in Chapter 2. This chapter explains the main idea of this PhD thesis. In this chapter, a new individualization technique for binaural synthesis using the frontal projection headphones is introduced (Fig. 4.1).

The frontal projection headphone is a special type of headphone that projects the sound from the front of the pinna unlike the conventional side headphones, which projects the sound from the side of the pinna. The Frontal projection headphone due to its

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.1: Comparison of a frontal projection and a conventional side projection headphone.

projection of sound from the front, inherently embeds the idiosyncratic pinna cues at the listener's eardrum during its playback, which contributes to the individualization process. The work presented in this chapter has also been published in the Journal of Audio Engineering Society [144].

4.1 Individualization techniques available in literature

Sound synthesis using individualized HRTF filters are required in order to provide a high quality virtual auditory space. Ideally, to achieve this, we would need to measure the HRTF of every individual, which is an impractical solution. Non-individualized HRTF databases are available, but it can be used only at the cost of a degraded spatial image [12, 28, 54, 106, 107]. Thus, the need to individualize the non-individualized HRTFs arises. The individualized HRTFs can either be jointly modeled or can be separately modeled using the combination of ITD and SC. Though individualization of both ITD and SC are important, the non-individualized SC has a more prominent role in degrading the spatial perception [106]. In this thesis, the individualized HRTF model is mainly

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

based on the SC. Some of the techniques to individualize the ITDs are presented next.

4.1.1 Individualization using ITD

ITD is an important directional cue that relates to the phase of the HRTFs and helps especially in azimuthal localization. The ITDs display a strong correlation with anthropometric parameters, and thus, it is most suitable to customize the ITD using an anthropometric based model.

Woodworth *et al.* [103] first proposed a very simple ITD model, from which the far-field and individualized ITDs could be predicted from the head radius a . This model is based on a spherical head with two diametrically opposite ears. Woodworth's model was later extended to a three-dimensional space, which defines the ITD in terms of azimuth and elevation in the far-field (Eq. 2.9). When the head-radius a is taken as 0.0875 m, the Woodworth's formula is found to be consistent with the measured mean for western population.

Savioja *et al.* [183] derived ITD in the horizontal plane HRTFs from human subjects using a linear approximation of interaural excess phase differences. The ITD measurement data was found to fit well to a spherical-head-based ITD model. The dependence on elevation was taken into account by adding a scaling term to the ITD equation:

$$ITD(\theta, \phi) = \frac{a(\sin \theta + \theta)}{2c} \cos \phi, \quad (4.1)$$

where a is the radius of the head, c is the speed of sound, θ and ϕ corresponds to the azimuth and elevation angle, respectively.

Algazhi *et al.* [126] then derived an optimal spherical head model based on the high frequency ITD from a population of 25 subjects. To do this, the optimal head radius a_{opt} is obtained by carrying out a least mean-square error minimization procedure between

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

the measured ITDs and the ITD estimated from the Woodworth model:

$$\hat{\tau} = \frac{a}{c} (\sin \theta + \theta). \quad -90^\circ \leq \theta < 90^\circ, \quad (4.2)$$

where $\hat{\tau}$ is modeled ITD (Woodworth's formula). For each direction, the ITD measurements at M directions can be expressed as an $M \times 1$ column vector $\bar{\tau} = [\tau_0, \tau_1, \tau_2, \dots, \tau_{M-1}]^T$. Let $\hat{\tau} = \frac{a}{c} r$, where $r = \sin \theta + \theta$ and let $\bar{r} = [r_0, r_1, r_2, \dots, r_{M-1}]^T$ be the corresponding vector of values of r for all measurement locations. The optimal radius a_{opt} could be obtained by minimizing the error vector length $e = \bar{\tau} - \frac{a}{c} \bar{r}$, which results in:

$$a_{opt} = c \frac{\bar{r}^T \bar{\tau}}{\bar{r}^T \bar{r}}. \quad (4.3)$$

The distribution of optimal head radii for the 25 subjects showed a range of values from 79 to 95 mm, with a mean of 87 mm. The angular errors were calculated by mapping the measured ITD to a lateral angle based on the optimal spherical head-model for each subject. It was found that for most part, the angular error was less than 5° except for high lateral and medium polar angles, where larger errors were found. The mean result obtained was close to the commonly used value of $a = 0.0875$ m.

In this method [126], the optimal radius a_{opt} is derived from the ITDs measured from the individualized HRTFs that are usually unknown for every subject. Several other models were developed like the spherical head model and the spherical head model with offset ears in order to overcome this. Algazi *et al.* [126] estimated the individualized radius as a linear summation of 3 anthropometrical parameters namely the half head-width (X_1), the half head-height (X_2) and the half head-depth (X_3) of the individual's head [126]. One of the techniques is to take the mean of the three head dimensions, but this led to an overestimation of the optimal radius for all the subjects. A regression on the anthropometric data for all the subjects is then used to compute the weights w_1, w_1 ,

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

w_3 corresponding to the three anthropometric parameters. The optimal radius and their corresponding weights were found to be:

$$a_{opt} = 0.51X_1 + 0.019X_2 + 0.18X_3 + 32mm. \quad (4.4)$$

This result shows that the contribution of height of the head is not very significant in determining the optimal radius.

ITDs vary with elevation and these variations are audible in terms of just noticeable difference (jnd) of the ITD. To incorporate the ITD variation in elevation, an additional degree of freedom was introduced by shifting the listener's ears around the head [126,184]. This led to the development of the spherical head model with offset ears. It was shown and later confirmed that the ITD variations matched the measured data with a small shift of a few centimetres with respect to the two diametrically opposite ears [185].

Li *et al.* [186] developed a statistical model to obtain the individualized ITD based on spatial fourier analysis and experimental data fit. The model reflects the spatial left-right symmetry and front-back asymmetry of ITD. Multiple linear regression analysis on the 52 Chinese subjects and their 17 anthropometric parameters resulted in a novel ITD formula given by:

$$\begin{aligned} ITD' = & 128 \sin \theta + 21 \sin 2\theta - 2 \sin 3\theta - 42 \sin 4\theta - 21 \sin 5\theta \\ & + (3.02 \sin \theta + 0.25 \sin 5\theta) \times x_3 - 1.33 \sin 2\theta \times x_{12} \\ & + 0.15 \times (-\sin 3\theta + \sin 4\theta) \times x_{15}, \end{aligned} \quad (4.5)$$

where ITD' is in terms of μs ; x_3 , x_{12} , and x_{15} denote the dimensions of the head, pinna flaring distance, and ear position, respectively in units of mm . The equation indicated that the individualized ITDs could be predicted from three anthropometric parameters. Performance of the four subjects outside the database demonstrated that the mean of

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

the total error is less than $20 \mu s$, with the lateral performance being inferior to that at other directions. In a similar manner, an individualized ILD model was established by Watanabe *et al.* [187] by performing an azimuth fourier series expansion at any given azimuth plane.

Although the individualization of ITDs is important, it cannot resolve the disambiguity in the cone of confusion region. Individualization of SC is required to uniquely resolve the confusions and to achieve accurate localization of sound.

4.1.2 Individualization of spectral cues

The modeling of spectral cues mainly involves the identification of the appropriate parameters that completely define the idiosyncratic features in the spectral cues. There are several individualization techniques to obtain the individualized HRTFs from acoustical measurements, anthropometric features of the listener, customizing generic HRTFs with some perceptual feedback or frontal projection of sound, as summarized in Table 4.1.

Acoustical measurements: The most straightforward individualization technique is to measure the individualized HRTFs for every listener at different sound positions [86, 188]. This is the ideal solution but it is extremely tedious and involves highly precise measurements. These measurements also require the subjects to remain seated with fixed orientation and remain relatively motionless for long periods, which may cause fatigue to the subjects. Zotkin *et al.* [64] developed a fast HRTF measurement system using the technique of reciprocity, where a micro-speaker is placed into the earcup and several microphones are placed around the listener to capture the impulse responses. A faster and more accurate method was designed at the Air Force Research Laboratory [189], which allows the entire collection of HRTFs in just 6-10 minutes.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Table 4.1: Comparison of the various HRTF individualization techniques.

How to obtain individual features	Techniques	Pros	Cons	Performance and remarks
Acoustical Measurements	Individual measurement [86], IRCAM, France [69], CIPIC [70], Uni. of Maryland, Tohoku Uni [71], Nagoya Uni [190], Austrian Academy of Sciences [74, 188]	Ideal, accurate	Requires high precision and time consuming; impractical for every listener	Usually a reference to compare for individualization techniques
Listening/ Training	Selection from non-individualized HRTF [191, 192], Frequency scaling [193] Tune magnitude spectrum [194–196], Active Sensory Tuning [197], PCA weight tuning [198], Critical Band Tuning [199]	Easy to implement; directly relates to perception	Takes time; like a training, might require regular training sessions, can cause fatigue	Making use of the individual listening to obtain the best HRTF, perceptually
	Select cepstrum parameters [200, 201], Optical Descriptors : 3D mesh [42, 202]			
Anthropometric data	Analytical or Numerical Solutions: PCA + multiple linear regression [203], Finite element method, boundary element method [42], Multiple linear regression [204], Multiway array analysis [205], Support vector machine [206], Artificial neural network [207]	Based on acoustic principles; Can study effects due to independent elements of the morphology	Need a large database; Tedious, Requirement : High resolution of imaging; Expensive equipments; Qualified users	Making use of the correlation between individual HRTF and anthropometric data
	Structural model of HRTFs [42] and HRTF database matching [208]			
Playback Mode	Frontal projection headphone [144, 209]	No additional measurement, listening training	A new structure, not applicable to normal headphones. Special equalization required	Automatic customization. More than 50% reduction on front-back confusion
Non-individualized HRTF	Generalized HRTF [12]	Easy to implement	Not accurate; bad localization	Not an individualization technique

Anthropometric data: Individualized HRTFs can also be modeled as weighted sums of basis functions, which can be performed either in the frequency or space domain.

The basis functions are usually common to all individuals and the individualization information is often conveyed by the weights. The HRTFs are essentially expressed as a weighted sum of a set of eigen vectors, which can be derived from PCA or

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

ICA [203, 210, 211]. The individual weights are derived from the anthropometric parameters that are captured by optical descriptors either by directly measuring, or derived from pictures, or by obtaining a 3D mesh of the morphology [42, 212]. The solution to the problem of diffraction of an acoustic wave with the listener's body results in individual HRTFs. This solution may be obtained by analytical or numerical methods, such as the boundary element method (BEM) or the finite element method (FEM) [42, 188]. Other methods of individualization include techniques such as multiple linear regressions [204], multiway array analysis [205], support vector regression [206], and even artificial neural networks [207]. The inputs in these methods can be a simple geometrical primitive [213] (a sphere, cylinder or an ellipsoid), a 3D mesh obtained from MRI or laser scanner or a set of 2D images [202]. An important advantage of these techniques is that the relative effects of a particular morphological element (pinna, torso, head) and their variation with size, location and shape can be independently investigated [42]. Individualization can also be carried out using a simple customization technique, where a HRTF is selected by matching certain anthropometric parameters [208]. One of the major challenges today to numerically model the HRTF is the very high resolution of imaging techniques required for accurate prediction of HRTFs at high frequencies. The resolution of the mesh imaging required, depends on the shortest wavelength, which is around 17 mm at 20 kHz [42]. Moreover, obtaining these optical descriptors demands for the use of extremely expensive laser, MRI scanners, and requires qualified users.

Perceptual feedback: Several attempts have been made to individualize HRTFs from a generic HRTF database using perceptual feedback. Subjects go through listening tests, where they choose the HRTFs based on the correct perception of frontal sources and reduced front-back reversals [191]. Listeners can also adapt to the non-

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

individual HRTF by modifying the HRTFs in order to suit his or her perception. Middlebrooks observed that the peaks and notches of HRTFs are frequency shifted for different individuals and that the extent of the shift is related to the size of pinna [193]. Tan *et al.* [194] developed a method for customizing non-individualized HRTFs by tuning the magnitude spectrum in certain frequency bands. Listeners tune the spectrum until they achieve a good and realistic spatialization. Other techniques involve active sensory tuning [197], tuning the PCA weights [198] and critical bands to individualize the HRTFs [199]. Martens [192] proposed a method where the subject is asked to identify the HRTF that he or she can localize the virtual sound in a particular reference direction in the vertical plane. These perceptual based methods are much simpler in terms of the required resources, and effort as compared to the individualization methods that are either obtained by acoustical measurements or derived from the morphology. However, these training sessions can be quite long and can result in listener fatigue.

4.2 Individualization using Frontal Projection Headphones

The individualization technique proposed here is the “frontal projection headphones playback” method that can help in individualizing the non-individual HRTFs. The advantage of the proposed technique is that this type of individualization technique does not require any measurement, training or the anthropometric data of the listener. In this chapter, the individualization of non-individualized binaural audio using the frontal projection headphones is first investigated with the help of detailed objective as well as subjective analysis. The frontal projection headphones playback is also compared with playback using conventional side-emitter headphones. The frontal projection headphones

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

due to its unique frontal positioning, inherently embed the idiosyncratic frontal pinna cues at the eardrum spectrum, thereby mimicking playback from a physical set of loudspeakers. It is hypothesized that this approach should improve the frontal image by reducing the front-back confusions. To illustrate this, the headphone transfer functions (HPTFs) of both the conventional as well as frontal projection headphones are measured on both dummy heads as well as human subjects (Section 4.4). In the next section, the design of frontal projection headphones and how they help in providing 3D cues is discussed.

4.3 Frontal projection headphones : Design

The choice of headphones type can greatly affect the transparency of the binaural rendering even with the correct headphone equalization. The external ear is un-hindered in the natural hearing conditions, where the sound pressures at the ear are governed by free-field characteristics. With headphones placed over the ear (circumaural headphones), the pressure characteristics of the sound arriving at the eardrum are greatly affected compared to the free-field characteristics due to the interaction between the external ear and the headphone enclosure. The closer the coupling characteristic of the headphones with that of the free-field, the more accurate and transparent is the reproduced sound. Thus, the frontal-projection headphone chosen is an “open-type” headphone. The driver along with its enclosure in the frontal projection headphone is placed at 7 cm away from the listeners ears at the front. This can be considered as a tiny loudspeaker brought near the ears. Neglecting the near-field effects, which is a topic of future work, the frontal-projection headphones can be modeled using the lumped-element technique (Appendix D). As the diaphragm of the speaker vibrates, the sound waves hits the pinna from the frontal direction, thereby capturing the listeners highly unique pinna spectral cues. In the following sections, the free-field and headphone listening conditions are compared.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.2: Sound transmission through the external ear in free-field a) Sketch of the anatomy b) an analogue model. Adapted from [140]

4.3.1 Transmission in free-field

Consider the transmission of the acoustic wave in the free-field to the eardrum. The related pressures at various points and its electrical equivalent can be shown in an analogue model as in Fig. 4.2. The sound pressure created at the entrance of the ear canal is denoted by $P_3(\omega)$ and the resulting pressure at the ear-drum is denoted by $P_4(\omega)$. The free-field pressure at the position of the center of the head in the absence of the head is often taken as the reference and is denoted by $P_1(\omega)$.

The ear canal is treated as a single acoustical transmission line in a certain range of frequencies. The transmission line analogy is typically valid for frequencies whose wavelength λ is much larger than the diameter of the ear canal. If the diameter of the ear-canal is assumed to be 8 mm, then

$$\lambda \gg 8 \times 10^{-3} \text{ m} , \quad \text{and} \quad (4.6)$$

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

$$f \ll \frac{340 \frac{m}{s}}{8 \times 10^{-3} m} = 42.5 \text{ kHz}. \quad (4.7)$$

An upper limit is often used where the diameter is a quarter of the wavelength and hence, the model is adequate upto 10 kHz. Thevenin's theorem can be used at the entrance of the ear-canal to split the source into the "open circuit" pressure $P_2(\omega)$ and the radiation impedance $Z_{radiation}$. "Open circuit" pressure $P_2(\omega)$ is the pressure that would exist at the position of $P_3(\omega)$, when there is no volume velocity running through $Z_{radiation}$, which is possible only when the ear-canal is physically blocked. It is possible to obtain $P_3(\omega)$ from a pressure division of $P_2(\omega)$ between the radiation impedance and the impedance of the ear canal. Note that both sound pressures and impedances are a complex function of pressure. In the following equations, the variable ω is omitted from the pressure terms for the sake of brevity.

$$\left[\frac{P_3}{P_2} \right] = \frac{Z_{earcanal}}{Z_{radiation} + Z_{earcanal}}. \quad (4.8)$$

The "open circuit" pressure does not exist physically, but can be obtained by measurement of sound pressure at the entrance of the ear canal with the ear canal blocked. The different transfer functions that need attention here are:

$$\left[\frac{P_4}{P_1} \right] = \frac{\text{Sound pressure at the ear drum}}{\text{Sound pressure at the center of the head (free-field)}}, \quad (4.9)$$

$$\left[\frac{P_3}{P_1} \right] = \frac{\text{Sound pressure at the entrance of the ear canal(Open)}}{\text{Sound pressure at the center of the head (free-field)}}, \text{ and} \quad (4.10)$$

$$\left[\frac{P_2}{P_1} \right] = \frac{\text{Sound pressure at the entrance of the blocked ear canal}}{\text{Sound pressure at the center of the head (free-field)}}. \quad (4.11)$$

The transmission from the sound source to the eardrum can actually be split up into

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.3: Sound transmission through the external ear in presence of headphone a) Sketch of the anatomy b) an analogue model. Adapted from [140]

different intermediate transfer functions:

$$\left[\frac{P_4}{P_1} \right] = \left[\frac{P_4}{P_3} \right] \cdot \left[\frac{P_3}{P_2} \right] \cdot \left[\frac{P_2}{P_1} \right]. \quad (4.12)$$

Møller [134] found that the transfer functions $\left[\frac{P_4}{P_3} \right]$ and $\left[\frac{P_3}{P_2} \right]$ are one-dimensional thus independent of direction and distance to the sound source. However, it should be noted that Møller [134] limited the analysis upto 10 kHz. The transfer functions $\left[\frac{P_2}{P_1} \right]$ measured at the blocked-ear canal thus encapsulates all the directional-variations in the HRTF. Therefore, it can be concluded that the physical location of a point where full directional information is present, can be chosen anywhere from the eardrum to the entrance of the ear canal and even at a few millimetres outside the ear canal. Thus, blocked-ear canal measurements are often preferred for both human subjects and dummy head.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

4.3.2 Frontal Projection headphone listening

In this chapter, HPTFs for both conventional and frontal headphones were measured at the blocked ear canal position. Consider a case where the free-field listening medium is disturbed in the presence of a headphone (Fig. 4.3). The pressure at the eardrum denoted by $P_6(\omega)$, $P_7(\omega)$ represents the pressure at the entrance to the ear canal, and P_5 represents the “open circuit” pressure at the entrance. $Z_{headphone}$ represents the impedance being looked out into from the entrance that includes the volume enclosed by the headphone, and the mechanical and electrical side transferred to the acoustical side. The headphone transfer functions that are important here are :

$$\left[P_7/E_{HP} \right] = \frac{\text{Sound pressure at the ear drum}}{\text{Voltage between the headphone terminals}}, \quad (4.13)$$

$$\left[P_6/E_{HP} \right] = \frac{\text{Sound pressure at the entrance to the open ear canal}}{\text{Voltage between the headphone terminals}}, \quad \text{and} \quad (4.14)$$

$$\left[P_5/E_{HP} \right] = \frac{\text{Sound pressure at the entrance to the blocked open ear canal}}{\text{Voltage between the headphone terminals}}. \quad (4.15)$$

The transmission from the sound source to the eardrum can be split up as follows:

$$\left[P_7/E_{HP} \right] = \left[P_7/P_6 \right] \cdot \left[P_6/P_5 \right] \cdot \left[P_5/E_{HP} \right], \quad (4.16)$$

where $\left[P_5/E_{HP} \right]$ and $\left[P_6/P_5 \right]$ are both headphone and subject dependent, while $\left[P_7/P_6 \right]$ is only subject dependent. The ratio $\left[P_7/P_6 \right]$ is the same as $\left[P_4/P_3 \right]$ since both these transfer functions represents the ear canal transmission. The transmission from $P_5(\omega)$ to

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.4: Transmission in a binaural system with frontal projection headphones.

$P_6(\omega)$ consists of pressure divisions and can be expressed as:

$$\left[\frac{P_6}{P_5} \right] = \frac{Z_{earcanal}}{Z_{headphone} + Z_{earcanal}}. \quad (4.17)$$

Thus,

$$\frac{\left[\frac{P_3}{P_2} \right]}{\left[\frac{P_6}{P_5} \right]} = \frac{Z_{headphone} + Z_{earcanal}}{Z_{radiation} + Z_{earcanal}}. \quad (4.18)$$

If the headphone listening condition is desired to be similar to the free-field listening condition, then,

$$\left[\frac{P_3}{P_2} \right] = \left[\frac{P_6}{P_5} \right]. \quad (4.19)$$

i.e. the headphone now does not have any impeding effect on the free-field listening characteristics. This is possible only when:

$$Z_{headphone}, Z_{radiation} \ll Z_{earcanal}. \quad (4.20)$$

or

$$Z_{headphone} = Z_{radiation}. \quad (4.21)$$

It is observed that there is considerable loading of $Z_{radiation}$ at certain frequencies, thus, $Z_{radiation}$ is not small compared to $Z_{earcanal}$. Thus, to emulate a free-field listening con-

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

dition, $Z_{headphone}$ has to be equal or atleast close to $Z_{radiation}$.

Headphones for which $Z_{headphone}$ is equal to $Z_{radiation}$ are termed as “open headphones” or free-air equivalent characteristic (*FEC*) headphone. The frontal projection headphone has a small emitter positioned at some distance from the ear, due to which the radiation impedance as from the ear is not disturbed. Due to this reason, frontal projection headphones can be considered as an open headphone with *FEC* characteristics.

Let us now look at the transfer functions involved in the binaural synthesis and playback using the frontal projection headphones (Fig. 4.4). The binaural synthesis is carried out using non-individualized HRTFs measured on a dummy head. If the microphone has a transfer function $M_1(\omega)$ that includes the sensitivity as well as the frequency response, and the electrical transfer function from the microphone to the frontal projection headphone is $G(\omega)$, then the total transfer function from the source ($P_1(\omega)$) to the sound pressure at the eardrum can be described as:

$$H = \left[P_2/P_1 \right] \cdot M_1 \cdot G \cdot \left[P_7/E_{HP} \right]. \quad (4.22)$$

The transfer function H should be equal to $\left[P_4/P_1 \right]$ to emulate the free-field transmission from the source to the eardrum. Thus,

$$\left[P_2/P_1 \right] \cdot M_1 \cdot G \cdot \left[P_7/E_{HP} \right] = \left[P_4/P_1 \right], \quad (4.23)$$

$$G = \frac{\left[P_4/P_1 \right]}{\left[P_2/P_1 \right] \cdot M_1 \cdot \left[P_7/E_{HP} \right]}, \quad (4.24)$$

$$G = \frac{\left[P_4/P_3 \right] \cdot \left[P_3/P_2 \right]}{M_1 \cdot \left[P_7/P_6 \right] \cdot \left[P_6/P_5 \right] \cdot \left[P_5/E_{HP} \right]}. \quad (4.25)$$

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL
SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

As, $\left[P_4/P_3\right] = \left[P_7/P_6\right]$, the first term in Eq.4.25 becomes unity.

$$G = \frac{\left[P_3/P_2\right]}{M_1 \cdot \left[P_6/P_5\right] \cdot \left[P_5/E_{HP}\right]}, \quad (4.26)$$

$$G = \frac{Z_{headphone} + Z_{earcanal}}{Z_{radiation} + Z_{earcanal}} \cdot \frac{1}{M_1 \cdot \left[P_5/E_{HP}\right]}. \quad (4.27)$$

The correct transfer function required for accurate binaural playback consists of a compensation for (a) the microphone sensitivity $M_1(\omega)$ (b) the headphone transfer function from the terminals to the blocked ear canal (c) transmission difference between free-field and the headphone listening condition. Since the frontal projection headphones is an “open headphone”, $\frac{Z_{headphone} + Z_{earcanal}}{Z_{radiation} + Z_{earcanal}}$ can be considered as unity. Thus, the expression $G(\omega)$ has to compensate only for the microphone sensitivity and the headphone transfer function.

$$G = \frac{1}{M_1 \cdot \left[P_5/E_{HP}\right]}. \quad (4.28)$$

The headphone equalization in general is described as the measurement of the transfer function from the voltage at the headphone terminals (E_{HP}) to the sound pressure at the position of the blocked ear canal ($P_5(\omega)$). The measurement of the pressure $P_5(\omega)$ also involves the microphone transfer function with a sensitivity and transfer function $M_2(\omega)$ and an output voltage $E_{microphone}$. Thus, $G(\omega)$ becomes:

$$G = \frac{1}{M_1 \cdot \left[\left(\frac{E_{microphone}}{M_2} \right) / E_{HP} \right]} \quad (4.29)$$

$$G = \frac{M_2}{M_1} \cdot \frac{E_{HP}}{E_{microphone}}. \quad (4.30)$$

If the microphones used to measure the headphone transfer function and the pressure at

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.5: How frontal projection provides the 3D cues. Figure adapted from [303]

the blocked ear canal were the same, then $M_1(\omega) = M_2(\omega)$, and thus,

$$G = \frac{E_{HP}}{E_{microphone}}. \tag{4.31}$$

4.3.3 How the frontal projection headphones provide 3D cues

To understand this, let us first look at the ear spectrum for frontal sound source. Some of the characteristic features of the frontal ear spectrum is shown in Fig.4.5. The dips at 1 kHz and 10 kHz arise from comb filtering effects of the shoulder and concha, respectively. It should be noted it is the pinna spectral cues that is highly individualized. The 10 dB peak around 3 kHz is due to the concave nature of the pinna region around the earcanal entrance. The diffraction with multiple reflection in craters always give rise to a global amplification of the incoming sound. The pinna including the concha gives approximately a 10 dB amplification, in addition to the 10 dB amplification due to the

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

$\lambda/4$ resonance in the earcanal. The frontal projection headphone projects the sound front and provides these unique comb filtering effects due to the pinna.

4.4 Measurement of headphones transfer functions (HPTF)

The HPTFs for the conventional and frontal projection headphones were measured in a semi-anechoic chamber at Nanyang Technological University, Singapore. The excitation signal used is an exponential sine sweep signal is similar to the HRTF measurements explained in Chapter 3.

The experimental procedures were all similar to the measurement of HRTFs (Chapter 3). Measurements were carried out using a Windows PC equipped with a B & K PULSE data acquisition system (Fig. 4.6). The HPTF measurements were first carried out using Bruel & Kjaer (B & K) 4128-C head and torso simulator (HATS) [1], which is equipped with the B & K 4158-C (Left) and 4159-C (Right) ear simulators. Additionally, the HPTFs were also measured at the blocked ear canal entrance of human subjects using the B & K 4101 binaural microphones. Both side and frontal projection headphone responses were measured.

The microphones captured the signals at the recording point (eardrum or blocked ear canal entrance) at a sampling rate of 65,536 Hz. The recorded signal was first downsampled to 44,100 Hz and processed using a MATLAB function to extract the impulse response. Kulkarni *et al.* [27] reported the variability of headphone responses at the high frequencies due to repositioning of the headphone on the head. To reduce the effect of this variability, 10 trials were taken and their average was taken as its representative headphone response. The headphones were removed and repositioned on the head after each trial. Subjects were asked to keep their heads still during the

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.6: Measurement of HPTFs using a HATS dummy head.

measurement process. “Open” headphones generally do have FEC characteristics. Since the frontal projection headphone used in this study is a “Open” headphone, the use of blocked-entrance technique is appropriate here (Section 2.7).

The variability of the responses for the side emitter and the frontal emitter response is shown in Figs.4.7 and 4.8, respectively. It can be noted that as Kulkarni *et al.* [27] pointed out, the variability at the high frequencies is much higher compared to that in the low-frequencies due to the strong headphone-pinna coupling. The average magnitude headphone responses of the conventional side projection and the frontal projection headphones based on the 10 trials for both human subjects and dummy head are shown in Fig. 4.14 .

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.7: Top : Variability in the side emitter headphone response based on 10 trials.
Bottom : Standard Deviation within the 10 trials.

4.5 Equalization of headphones: Type-1 and Type-2

Headphones are not acoustically transparent as they not only color the input sound spectrum but also affect the free-field characteristics at the ear. Thus, the headphone response has to be equalized before binaural playback over the same pair of headphones. It is observed that, in the absence of headphone equalization, the front-back reversals are increased and the elevation localization is distorted [42]. Thus, headphone equalization is critical to create a convincing perception of virtual sound sources. Furthermore, the main difficulty of headphone equalization is that the HPTF depends on individual morphology (headphone-ear coupling). Researchers have also reported that use of non-individualized equalization can reduce the externalization [28] and the effect can be as critical as the

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.8: Top : Variability in the frontal emitter headphone response based on 10 trials. Bottom: Standard Deviation within the 10 trials.

use of non-individualized HRTFs [42]. Thus, equalization using individualized HPTFs is strongly recommended.

Typically, the headphone transfer function (HPTF) comprises of the headphones transducer response and the acoustic coupling between the headphones and the listener's ears, as shown in Fig. 4.9. The headphones transducer response is characteristic of the type of transducer used in a given headphone, while the acoustic headphone-ear coupling is subject-dependent and depends on the positioning of the headphones on the listener's ear. For this reason, the headphone-ear coupling is highly idiosyncratic. Due to the nature of the headphones, different types of equalization have to be applied to the side emitter and frontal emitter headphones. Equalization involves spectral inversion that often leads to excessive boosting of certain frequencies rendering the equalized audio

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.9: Collapsed view of headphone response

perceptually unpleasant. To alleviate this problem, inverse filters were created based on Kirkeby’s fast deconvolution method using frequency dependent regularization to obtain acceptable accuracy of equalization [214].

4.5.1 Type-1 equalization (conventional headphones)

The Type-1 equalization aims at obtaining a flat target response as it tries to compensate for both the headphone transducer response as well as the unwanted headphone-ear coupling. Such a type of equalization is the most commonly used equalization method and is best suited for conventional side emitter headphones. Type-1 equalization is highly dependent on the headphone-ear coupling and thus, varies with the subject’s morphology. Perfect equalization of the headphone is very difficult to achieve as the response also varies with headphone repositioning [27].

$$Y(\omega) = S(\omega) \cdot HRTF(\omega) \cdot \frac{1}{HPTF(\omega)} \cdot HPTF(\omega), \tag{4.32}$$

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.10: Type-1 Equalization process.

where $Y(\omega)$ = Equalized Binaural Signal

$X(\omega)$ = Source Signal Spectrum

$HRTF(\omega)$ = Head Related Transfer Function (Left/Right)

$HPTF(\omega)$ = Headphone Transfer Function (Left/Right), and

$\frac{1}{HPTF(\omega)}$ = Type-1 Equalization Filter

The binaural signal is convolved with the Type-1 equalization filter to obtain the Type-1 equalized binaural signal (Fig. 4.10). When this signal is played back from the same headphone, the distortion created by the headphone is eliminated so that the spectrum at the eardrum contains only the HRTF features. Some of the drawbacks of Type-1 Equalization are:

1. The Type-1 equalization filter is highly idiosyncratic, since the headphone transfer functions display large variations with individuals.
2. The HRTFs and the HPTFs should be measured on the same head.

4.5.2 Type-2 Equalization (frontal emitter headphones)

Equalization plays a greater role in frontal projection headphone playback, since the frontal projection headphone transfer function also includes the embedded personal pinna cues. It is intended to preserve these individualized spectral cues in the equalization process, since these are vital frontal localization cues. To cater to this need, we propose

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.11: Example of a measurement of free-field response of the frontal projection headphone.

a robust equalization technique for the frontal emitter, which is referred as the Type-2 equalization [144, 209].

In the case of using frontal projection headphones, it is desired to preserve the frontal personal pinna cues at the eardrum, which is inherently created during playback. Thus, using conventional Type-1 headphone equalization is not advisable as such a type of equalization would equalize even the individualized pinna spectral features generated during playback. For the frontal projection headphones, only the distortion created by the headphone transducer along with the resonant modes of the earcup has to be compensated in order to ensure that the personal pinna cues (*pinnacues*) due to frontal projection are preserved after equalization. Such a type of equalization improves the performance of binaural audio playback using frontal projection headphones [144, 209].

To obtain the Type-2 equalization filter, the free-field response (*FFR*) of the frontal projection headphone is first measured. *FFR* is obtained by measuring the response of the headphone in the absence of dummy head to capture only the spectral effects of the transducer and the resonant modes of the earcup. Here, a B & K 4961 Multi-Field 1/4" microphone is placed at the center of the earcup just outside its peripheral surface as

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.12: Type-2 Equalization process.

shown in Fig. 4.11. The whole setup was placed in a B & K 4232 anechoic test box to provide some isolation from the external environment. The distance of the microphone from the transducer is approximately equal to the distance (roughly 5 cm) at which the ears would have been present if the headphone is put on. The FFR response (shown in Fig. 4.13) obtained approximately captures the spectral effects of the transducer as well as the resonant modes in the earcup and does not include the personalised frontal pinna cues. The Type-2 equalization is carried out by convolving the non-individualized binaural signal with the inverse of the free-field response (FFR) of the frontal projection headphone (Fig. 4.12). The Type-2 equalization can also be described in mathematical terms as follows:

$$\text{FER}(\omega) = \text{FFR}(\omega) \cdot \text{pinnacues}(\omega), \quad (4.33)$$

$$Y(\omega) = X(\omega) \cdot \text{HRTF}_{\text{non-ind}}(\omega) \cdot \frac{1}{\text{FFR}(\omega)} \cdot \text{FER}(\omega), \quad (4.34)$$

$$Y(\omega) = X(\omega) \cdot \text{HRTF}_{\text{non-ind}}(\omega) \cdot \text{pinnacues}(\omega), \quad (4.35)$$

where $\text{pinnacues}(\omega)$ corresponds to the personalized cues due to frontal projection. $\text{FER}(\omega)$ and $\text{FFR}(\omega)$ is the frontal projection headphone response and the free-field response, respectively. $\text{HRTF}_{\text{non-ind}}(\omega)$ is the non-individual HRTF used and $\frac{1}{\text{FFR}(\omega)}$ is the Type-2 equalization filter. $X(\omega)$ and $Y(\omega)$ are the input mono signal and the output binaural signal, respectively recorded at the eardrum.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.13: The free-field response based on 10 trials and their average FFR taken as the representative transfer function.

The frontal projection headphone transfer function, which is measured at the eardrum can be approximated as a combination of the personalised pinna cues (*pinnacues*) generated and the free-field response (*FFR*) of the headphone as shown in Eq. 4.5. On equalizing the binaural signal with the Type-2 equalization filter, the response of the transducer and the spectral effects of the earcup are removed, which results in a superposition of the personalised cues and the non-individualized HRTF cues at the eardrum as shown in Eqs. 4.6 and 4.7. Some of the main advantages of Type-2 equalization are:

1. Type-2 equalization is subject-independent. Thus, it is required to measure the Type-2 equalization filter only once.
2. Type-2 equalization filter also encounters lesser variation between different trials of

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

measurement.

3. It preserves the unique idiosyncratic pinna cues.

4.6 Objective assessment

The frontal emitter response and side emitter response, which were measured on the HATS and a human subject 'K', are plotted in Fig. 4.14. For both the HATS and subject 'K', Frontal projection headphone response (FER) exhibits a deep notch, which is inconspicuous in the conventional side projection headphone response (SER). This notch has been identified as the frontal perception notch (FPN); a potential cue for localizing frontal directions as shown in Fig. 4.14. FPN is highly idiosyncratic and is formed due to the interaction of the frontally incident acoustic wave with the external ear. Frontal projection of sound automatically creates personalized reflections, diffractions with the external ear, required to localize frontal sources. On the other hand, this cue is not prominent in the SER leading to poor frontal perception unless individualized HRTFs and headphone equalization are used.

Figure 4.15 shows the spectral difference (SD) between FER and SER (i.e., $SD = FER - SER$). We note that the SD displays a similar boost/attenuation pattern as Blauert's [109] observation of the sound pressure level differences at the ear canal entrance between the front and rear directions. However, they differ in the amount of gain in each of the directional bands. This kind of complex spectral shaping is thought to be an important monaural cue for disambiguating the cones of confusion. These findings confirm that the frontal emitter naturally aides in frontal perception and the side emitter tend to bias the perception towards the rear due to imperfect equalization.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.14: Comparison of Frontal emitter response and Side emitter response for Subject ‘K’ and HATS dummy head (Bottom). Note that the bottom curves have been shifted by -40 dB for better clarity.

4.6.1 Comparison of HPTFs with the measured HRTFs

In this section, we investigate the similarity in the pinna cues in FER and SER with the experimentally measured HRTFs of the HATS dummy head. Mean square error (MSE) [10] is considered as the objective criterion to compare the high-frequency pinna cues (5 kHz to 16 kHz).

The HPTF and the HRTF curves are normalized first before calculating the MSE. The normalization is carried out by subtracting each of the HRTF (in dB) and the HPTF (in dB) from their respective average values calculated across the whole spectrum. The addition of one in the MSE expression is to obtain a value of 0 dB error for a perfect match model:

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.15: Solid Line: SD between FER and SER; Dotted Line: Blauert's observed pattern of SPL differences between front and rear directions.

$$\text{MSE} = 10\log_{10} \left\{ \left[\frac{1}{N} \sum_{f=f_l}^{f_h} (HRTF_{norm}(f) - HPTF_{norm}(f))^2 \right] + 1 \right\}, \quad (4.36)$$

where f_l is 5 kHz and f_h is taken as 16 kHz. N is equal to the number of frequency bins between f_l and f_h . $HRTF_{norm}$ and $HPTF_{norm}$ are the normalized HRTF and HPTF, respectively.

Figure. 4.16(a) shows the comparison of FER with the frontal HRTFs. It is observed that HRTFs in the frontal directions have similar pinna cues as the FER. SER also shows a high degree of similarity in the pinna cues with the rear azimuthal HRTFs as shown in Fig.4.16(b). The objective analysis shows that MSE between the FER and the frontal HRTFs is least ($MSE_{avg} = 1\text{dB}$) and increases as the azimuth angle increases towards the rear ($MSE_{avg} = 6\text{dB}$). On the other hand, SER has an MSE_{avg} of 6 dB with respect

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.16: (a) Comparison of FER and Frontal HRTFs (0° to 50°) (b) Comparison of SER and Rear HRTFs (100° to 140°).

to the frontal HRTFs and an MSE_{avg} of 1 dB with respect to the rear HRTFs.

These observations establish the fact that the frontal emitter projection generates

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

frontal spectral cues as found in the frontal azimuthal HRTFs, and the side emitter projection creates similar pinna cues as in the rear azimuthal HRTFs.

4.7 Subjective experiments

Subjective experiments were carried out to validate the effectiveness of the frontal emitter playback over the side emitter playback. The three subjective experiments conducted are as follows:

1. Effect of equalization on frontal emitter playback (15 subjects).
2. Localization experiment (40 subjects).
3. Superposition experiment to study the superposition of the frontal projection spectral cues and the non-individual HRTF cues (40 subjects).

The subjects were paid for their participation in the listening tests. All the subjects were between 18 and 30 years. Most of the listeners had prior experience as participants in similar listening tests. An audiometry test was first conducted to eliminate the subjects with any hearing disorder. Double blind test procedure was adopted in all the three tests, where the evaluator as well as the subject is unaware of the order of the test signal set.

4.7.1 Effect of equalization on frontal projection headphone playback

The purpose of this experiment was to explore the effectiveness of the frontal emitter playback and validate the proposed Type-2 equalization method. Both the side and frontal emitter headphones were equalized before playback. The side projection headphone was equalized using the Type-1 equalization, while the frontal projection headphone was tested for both Type-1 and Type-2 equalization. The test stimuli used were

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

WGN filtered with HRTFs of azimuth angles of 0° , 30° , 60° , and 75° across the frontal horizontal plane (0 elevation) taken from the DSPLab HRTF database measured on a dummy head [31]. In this experiment, the term binaural spatialization refers to the HRTF filtered WGN. Thus, there were three experimental conditions in this experiment:

1. Binaural spatialization with Type-1 equalization - Side emitter playback.
2. Binaural spatialization with Type-1 equalization - Frontal emitter playback.
3. Binaural spatialization with Type-2 equalization - Frontal emitter playback.

All the subjects attended to three trials for each azimuth angle, which makes a total of 36 sound tracks ($3 \text{ experimental conditions} \times 3 \text{ trials} \times 4 \text{ azimuth angles}$) in this test. The subjects responded to the sound and recorded the relative position of the perceived virtual auditory image from the head (front/behind) on a subjective scale of 0 to 5.

4.7.2 Results

The rate of front-back reversals experienced by the subjects for all the azimuth angles tested is shown in Fig. 4.17. The responses from the subjects, who localized the sound as coming from behind, are considered to be confused since only frontal sounds were presented. It can be seen that the confusions are highest, when the playback is from the side emitter headphones with the conventional Type-1 equalization. Playback from the frontal emitters with Type-1 equalization reduces the rate of front-back confusions approximately by 50% of those obtained for the side emitter playback. Use of Type-2 equalization on the frontal emitter further reduces the front-back confusion to below 20% for 30° , 60° , 75° azimuth and 23.8% for 0° azimuth. This finding validates the effectiveness of the Type-2 equalization employed for the frontal emitters.

An ANOVA test [215] was carried out in-order to validate the results for the whole population. The effect of the frontal emitter on the rate of front-back confusions is

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.17: Front-back confusions (%) for the side emitter (Type-1 equalization) and frontal emitter (Type-1 and Type-2 equalization).

first considered. The type of emitter has a significant effect on the rate of reversals [$F(1, 14) = 52.8, p \ll 0.01$] (Note that F is the standard F -distribution and p refers to the probability value, which when compared to the critical value determines the significance of the factor). The subjects are seen to be a significant main factor [$F(1, 14) = 5.96, p \ll 0.01$] and the two-way interaction effect between subject and type of playback is also significant [$F(1, 14) = 6.6, p \ll 0.01$]. In the playback using the frontal emitter, the type of equalization (Type-2) also has a significant main effect in reducing the front-back reversals [$F(1, 14) = 10.6, p < 0.05$] for a 95% confidence interval. The two-way interaction effect between the subjects and the type of equalization is also significant [$F(1, 14) = 9.89, p < 0.0001$]. We also note that the performance of the

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.18: Relative distance from the head for different equalization conditions.

frontal emitter is consistent among all the frontal azimuth angles tested ($p > 0.01$).

Figure. 4.18 shows the subjective score indicating the relative distances of the auditory image perceived for all frontal azimuthal angles tested. We observe that the auditory image is most likely perceived at the back of the head when the playback is from the side emitter. On the contrary, the frontal projection of sound results in better frontal perception. It can also be seen that the proposed Type-2 equalization method results in an enhanced frontal image relative to the conventional Type-1 equalization for the frontal emitter. The ANOVA analysis returned significant main effects for the emitter type [$F(1, 14) = 47.8, p < 0.01$], equalization type [$F(1, 14) = 11.06, p = 0.01$], and the subjects [$F(1, 14) = 9.04, p < 0.0001$].

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.19: Localization Response Chart.

4.8 Localization experiment

In this experiment, the localization accuracy of the subjects using the frontal emitter and the side emitter headphones were tested. A clock test procedure was adopted for testing the azimuth localization as suggested by Evans [216], as shown in Fig. 4.19. The stimulus in this experiment is a HRTF filtered WGN (binaural spatialized signal) with Type-1 equalization for the side emitter and Type-2 equalization for the frontal emitter playback. Each subject listened to a total of 96 sounds (12 azimuthal directions $\times$ 4 trials $\times$ 2 types of emitters) from both the emitters (frontal and side). Subjects responded by choosing the position of the direction of the virtual sound in an on-screen GUI presented to them.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.20: Percentage localization accuracy in Frontal and Rear directions using both frontal and side projection emitters.

4.8.1 Results

In this experiment, the localization response is considered accurate, if the subject identified the correct clock position. The average localization performance for frontal and rear azimuthal directions is shown in Fig. 4.20. With frontal emitter playback, the localization performance in the frontal and rear directions is found to be 74.3% and 77%, respectively. On the contrary, the side emitter playback, result in slightly better performance in localizing rear directions, (79.5 %) while the frontal direction localization is poor (27%) as expected. The localization responses of the subjects for both side and frontal emitters are shown in Fig. 4.21(a) and Fig. 4.21(b), respectively. It is clear that, higher number of front-back confusions exist in the side emitter playback compared to the frontal emitter playback. However, there were very few back-front confusions observed in the frontal emitter playback hinting its frontal perceptual bias.

For the ANOVA analysis, the type of headphone playback and directions are con-

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

sidered as the main factors of interest. The ANOVA procedures demonstrated that the difference between the frontal and rear directional localization accuracy using frontal emitter playback is not significant [$F(1, 4) = 2.34, p = 0.2$]. On the other hand, the difference in the accuracy of localizing frontal and rear directions using the side emitter is significant [$F(1, 4) = 588.84, p < 0.0001$].

The difference in the rear localization performance between the frontal and side emitter is also not significant [$F(1, 4) = 4.32, p = 0.1$], while the difference in frontal localization performance between the two types of playback is highly significant [$F(1, 4) = 804.71, p \ll 0.0001$]. Furthermore, for the frontal emitter playback, a Tukey Honest Significant Difference test post hoc test (Montgomery, 2008) showed that the localization performance was consistent ($p > 0.01$) for all the directions in the front.

4.9 Superposition experiment

The purpose of this experiment is to study the prominence of the frontal projection spectral cues over the non-individual HRTF cues in disambiguating the front-back directions during playback. To investigate the superposition of these cues, the spectral cues in 1/2, 1, 2 octave BW above 4 kHz are removed from the non-individual HRTF during the synthesis of the binaural signal (Fig. 4.22). The spectral cues were removed by replacing the HRTF spectrum in different frequency bands with the average value of the HRTF across the whole spectrum.

A total of 10 experimental conditions were considered for the design of this test, as shown in Table 4.2. The frequency bands to be flattened were chosen similar to that used by Langedijk *et al.* [113]. In each experimental condition, subjects were tested for azimuth angles of 30° , 45° , 60° in the frontal azimuthal plane. Each subject listened to a total of 180 sound tracks (10 experimental conditions $\times$ 3 azimuth angles $\times$ 6 repetitions). The binaural signal was equalized before playback using Type-1 equalizer for side emitter

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.21: Localization test responses for (a) Side Emitter (b) Frontal Emitter.

playback and Type-2 equalizer for frontal emitter playback.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.22: Frequency response of the modified HRTF spectrum after spectrum flattening.

Table 4.2: Experimental Conditions in Superposition Test

S/n	Experimental Conditions	Frequency range removed	Playback Mode
1	2-Octave BW	4 kHz - 16 kHz	Frontal Emitter
2	1-Octave BW high	8 kHz - 16 kHz	
3	1-Octave BW middle	5.7 kHz - 11.3 kHz	
4	1-Octave BW Low	4 kHz - 8 kHz	
5	1/2-Octave BW High	11.3 kHz - 16 kHz	
6	1/2-Octave BW middle high	8 kHz - 11.3 kHz	
7	1/2-Octave BW middle low	5.7 kHz - 8 kHz	
8	1/2-Octave BW Low	4 kHz - 5.7 kHz	
9	Baseline	No Flattening	Side Emitter
10	Baseline	No Flattening	

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

4.9.1 Results

The rate of reversals for each experimental condition is examined and is plotted in Fig. 4.23. It can be noticed that the rate of reversals is highest (70%) for the baseline condition (Experimental condition 10), when played back using the side emitter. Playback from the frontal emitter for the baseline condition (Experimental condition 9) leads to much lesser front-back confusions (around 20%) than the side emitter playback.

For the experimental condition 1, where the spectral cues from 4 kHz to 16 kHz were removed, the frontal emitter projection still resulted in lesser front-back confusions compared to the baseline condition for side emitter playback. Rate of confusions for the experimental conditions 1 to 3 (2-octave, 1-octave high and 1-octave middle BW) are slightly higher because the non-individual pinna spectral cues are removed completely in these cases.

ANOVA procedure indicates that there is a significant main effect of using the frontal emitter playback against the side emitter playback system [$F(9, 39) = 198.3, p = 0.01$]. A Tukey HSD post hoc analysis (Montgomery, 2008) showed that the 2-octave BW flattened test condition is significantly different ($p < 0.001$) from the experimental conditions 5, 6, 7, 8 and 9, where the high-frequency pinna spectral cues are preserved. The subjects were also identified as a significant main factor [$F(9, 39) = 38.51, p < 0.01$] in the experiment. Furthermore, a two-way interaction between the subjects and the various experimental conditions is also found to be significant [$F(9, 351) = 4.2, p < 0.01$].

4.10 Additional processing blocks

Additional processing blocks like the decorrelation and the front-back biasing can be applied to enhance the spatial perception. However, this comes at a cost of timbral coloration and thus, it is suggested that these processing blocks be used in applications

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.23: Percentage rate of reversals for various experimental conditions.

where the timbral quality is not that critical. The frontal projection of the sound exaggerates the frontal spectral cues and resolves the ambiguity in the cone of confusion region. The non-individualized HRTFs are partially individualized by the superposition of the auxiliary individual cues over the generic HRTF cues. To further enhance the externalization perception using non-individualized synthesis as in this chapter, a decorrelation filter can be additionally applied [217]. Furthermore, front-back bias filter [218] can be used to enhance the spectral differences between the front and back, thereby enhancing the frontal spatial imagery. This can also be crudely used as a technique to control the auditory depth of the virtual image by adjusting the low-frequencies.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.24: Process of obtaining decorrelated signals for both left and right channels.

4.10.1 Decorrelation

Decorrelation is a technique used to externalise the stereo sound source and produces a diffused sound field preserving the localization of the auditory image [217]. Decorrelation filters can be easily created by generating an all pass filter with a phase response constructed from combinations of random number sequences (Fig. 4.24). Filtering the stereo sound source with a decorrelating filter reduces the correlation between the two channels rendering an auditory image that is more diffused and out of head.

4.10.2 Front-back biasing

Blauert [11] found that certain bands (directional bands) in the frequency spectrum were boosted/attenuated, characterising the direction of arrival for frontal and rear directions. He also noticed that these directional frequency bands had alternated boost and attenuated patterns for the sounds coming from the front and rear directions (Fig. 4.25). Thus, a frontal bias filter can be applied to enhance the frontal imagery, while a rear bias filter can be applied to enhance the rear imagery [209, 218].

In addition, a crude technique to control the auditory depth can be achieved by adjusting the low-frequency band (225 Hz to 680 Hz) in the frontal bias filter. A frontal bias filter is adopted here that spans up to 16 kHz, based on Blauerts results for directional

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.25: Frontal bias filter specifications. Bands (1, 3, 5) are frontal boosted bands; Bands (2, 4) are rear boosted bands.

bands [11]. In the next section, we describe some of the subjective experiments carried out with the two additional processing blocks.

4.11 Subjective test: Additional processing blocks

This test investigated the effect of decorrelation and the frontal bias filtering to enhance the externalization and frontal imagery. Subjects listened to four different stimuli (white noise) filtered with HRTF of 30° azimuth (0° Elevation) and judged the relative position of the virtual image from the head in the horizontal plane. All the test signals were played through the frontal emitters in this test. Figure. 4.26 depicts the subjective response chart. The different stimuli used in this test were:

1. HRTF filtered signal with Type-2 equalization (Binaural spatialization).
2. Binaural spatialization + decorrelation.
3. Binaural spatialization + decorrelation + frontal bias (All bands present).
4. Binaural spatialization + decorrelation + adjusted frontal bias (Boost in band 1 attenuated to 0 dB).

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.26: Subjective test response chart. IH = In-head localization, NF = Near front, F = Front, FF = Far front. (1, 1'), (2, 2'), (3, 3'), (4, 4') are all symmetrical pairs of same ITD and ILD values.

Double blind test procedure was adopted in both the tests, where the evaluator as well as the subject is unaware of the order of the test signal set. In our subjective tests, we presented all the virtual sounds from the frontal azimuthal plane and thus, the subject's responses were based on only the absolute cues provided by frontal sound with no additional cues for front-back discrimination.

4.11.1 Results

Decorrelation has the effect of externalizing the auditory image when the source is perceived to be inside the head. Frontal projection, with Type-2 equalization does externalize the sound source well in front of the head. Additionally, applying decorrelation has the effect of further increasing the perceived auditory distance in the front away from the head. This can be seen from Fig. 4.27, where the relative distance from the head increases (stimulus 1 to stimulus 2) when decorrelation is applied.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

Figure 4.27: Relative distance from the head (Subjective test 2).

From the preceding experiment that showed that the decorrelation block improves the externalization of the virtual auditory image, the frontal bias filter is applied to modify the auditory depth by adjusting different frequency bands (stimuli 3 and 4). From Fig. 4.27, it is noted that as the frontal bias is applied (stimulus 3), the auditory image approaches closer to the head mainly due to a boost at the lower frequencies (Band 1 in Fig. 4.25). Further, it is observed that the perceived auditory distance increased away from the head as the gain from band 1 is attenuated (stimulus 4).

However, it should be noted that the front-back bias filter leads to severe timbral coloration and thus, it is not recommended where individualization of the spatial sound is the main focus.

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

4.12 Summary

The localization performance with frontal projection headphones and the conventional side projection headphones were compared for non-individualized binaural synthesis. The mode of playback and the equalization technique is identified as the two main discerning factors for better spatial audio reproduction. For the frontal projection headphone playback, Type-2 equalization technique (in Section 4.4) is employed, which preserves the personal pinna spectral cues important for frontal perception. Furthermore, the frontal projection of sound creates personal pinna cues similar to those present in the measured frontal azimuthal HRTFs. On the other hand, an unequalized side projection headphone response displayed spectral features similar to that present in the measured rear directional HRTFs. The side projection headphone response cannot be perfectly compensated as the Type-1 equalization is highly subject-dependent. The headphone responses also vary with repositioning of the headphone on the head. Since the spectrum at eardrum does not have any individualized pinna spectral cues due to the use of non-individual HRTFs, the spatial image is distorted even if the Type-1 equalization for the side projection headphones achieved is close to perfect. On the other hand, the Type-2 equalization technique does not encounter these issues as the equalization filter is subject-independent.

Subjective experiments showed that there are lesser front-back reversals using the frontal projection headphones compared to the side projection headphones. The Type-2 equalization technique further improves the binaural playback using the frontal projection headphones. Frontal projection headphones are found to deliver superior frontal localization performance when compared to side projection headphones. However, the side projection headphones performed marginally better in localizing the rear directions compared to the frontal headphones. However, this difference was found to be statistically

CHAPTER 4. INDIVIDUALIZATION OF NON-INDIVIDUALIZED BINAURAL SYNTHESIS USING FRONTAL PROJECTION HEADPHONES

insignificant from the ANOVA tests. Some back-front confusions were also observed in frontal emitter playback, as some of the subjects were probably influenced by the frontal perceptual bias of the frontal emitter. It is important to note that in this experiment, the signal spectrum at the eardrum is the superposition of the non-individualized spectral cues (from non-individualized HRTF) and the individual pinna cues due to frontal projection.

The superposition experiment revealed that non-individualized high-frequency pinna cues are not necessary for front-back discrimination using frontal projection headphone playback. This result suggested that the frontal projection cues are self-sufficient in providing all the necessary pinna spectral cues for good frontal perception. In fact, the presence of non-individual HRTF pinna spectral cues only marginally improves the frontal perception and thus, individualized pinna cues seem to be the most prominent.

In this chapter, we have validated the frontal projection playback technique, which can aid the individualization process by adding personal pinna cues. In the subsequent chapters, individualized HRTFs in three dimensions are directly modeled using the frontal projection headphones. The distance-dependent individualized HRTFs in the horizontal plane are modeled in Chapter 6 using the frontal projection headphones. As we have seen before, headphones due to its non-uniform transfer function can adversely affect the perceptual image. In order to model the distance-dependent individualized HRTFs, it is important to first understand the effect of headphone equalization on distance perception. The effect of headphone equalization on directional perception is well studied, however, the effect of equalization on distance perception is still unclear. In the next chapter, the effect of equalization on distance perception is investigated for different types of headphones based on detailed subjective experiments.

Chapter 5

Effect of headphone equalization on distance perception

Headphones are not acoustically transparent and thus, it affects both the timbral as well as the spatial quality of the perceived sound. The effect of headphones has to be compensated by calculating an equalization filter and convolving it with the synthesized binaural audio. HPTF depends on headphone-ear coupling, and thus, displaying high spectral variation between individuals. It has been found that the type of equalization (individualized or non-individualized) affects the directional perception of the virtual audio reproduced using headphones. Before we consider the modeling of distance-dependent individualized HRTFs (Chapter 6) using frontal projection headphones, it is important to understand how the headphones affect distance perception. In this chapter, we study in detail the perceptual effects of equalization on the auditory distance perception in the proximal region in anechoic conditions. The effect of repositioning of the headphone on auditory depth perception is also studied in this chapter.

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

5.1 Introduction

Auditory depth perception forms a critical component in the veridical reproduction of virtual audio. The detailed literature related to auditory distance perception is presented in Chapter 6. Since headphones are the most preferred and convenient medium of sound reproduction, it is important to understand the effect of headphones on the perception of distance. Headphones reproduction itself is not acoustically transparent as it modifies the input sound spectrum and thus has to be compensated [12]. Absence of headphone equalization results in increased front-back reversals and elevation errors [12, 42, 188].

Researchers have also found that the effect of non-individualized equalization can be as critical as using non-individualized HRTFs or recording [135]. Thus, headphone equalization is extremely essential for accurate directional localization [12]. Moreover, headphone response also varies with repositioning over the ears every time the listener wears the headphone [27]. All these idiosyncratic factors affect the directional localization perception but little has been studied on its effect on distance perception.

In this chapter, the effect of equalization on auditory distance perception especially in the proximal region is studied with the help of perceptual experiments. We try to answer the following questions in this chapter:

1. What is the effect of equalization on the distance perception?
2. Is individualized equalization necessary for accurate distance perception?
3. Does repositioning of headphones effect the distance perception?

These set of research questions have not been investigated in detail before. The extent of equalization and individualization on externalization was studied by Kim *et al.* [28] They found that individualized equalization is important for externalization and individualized synthesis is necessary for consistent distance perception. Durlach *et al.*

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

Figure 5.1: The closed-back, open-back and the open headphones used in this study.

[150] also reported that non-individualized HRTFs and incorrect equalization results in the localization of sound inside the head. In this chapter, we study in detail the effects of equalization on distance. This chapter is structured as follows : The measurement of HPTFs and equalization are explained in Section 5.2. Procedures of the subjective experiments are presented in Section 5.3. The perceptual experiments and the results for evaluating the effect of equalization using non-individualized and individualized HRTFs are explained in Section 5.4 and 5.5 respectively. The effect of repositioning of the headphone is investigated in Section 5.6. ANOVA analysis and a discussion of the work is presented in Sections 5.7 and 5.8, respectively.

5.2 Measurements of HPTFs and equalization

HPTFs were measured at the blocked ear canal entrance with the same settings as for the measurement of the HRTFs explained in Chapter 3. In this experiment, the HPTFs of three different types of (closed-back, open-back, open) headphones were measured. AudioTechnica ATH-M30 (closed-back), Sennheiser HD-600 (open-back), and Sony PFR-V1 (open) were chosen as the representative headphones for each type as shown in Fig. 5.1.

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

Figure 5.2: Headphone transfer function and equalization filters for various headphone types. a) Closed-back headphones b) Open back headphones c) Open headphones.

HPTFs were measured at the blocked ear canal of three subjects and the HATS dummy head. A total of 10 different trials were carried out to measure the HPTFs, with repo-

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

sitioning of the headphone over the ears after each trial. An average response of all the trials is considered as the representative HPTF [27]. Headphone equalization filters for each of the closed-back(Fig.5.2a), open-back(Fig.5.2b), and open-type (Fig.5.2c) headphones were then obtained. Figure.5.2 shows the calculated equalization filters along with the average HPTFs for the dummy head and the three subjects for the closed-back, open-back and open headphones. We note that the equalized transfer functions obtained by convolving the headphone response and the equalization filter are very close to a flat response.

5.3 Subjective experiments

Subjective experiments were carried out to understand the effect of equalization on auditory distance perception. All the fifteen subjects who participated in the experiments were volunteers and had prior experience as participants in similar listening experiments. Subjects had to first undergo an audiometry test and all were reported to have normal hearing. The procedure employed was double blind, where the subject as well as the evaluator was unaware of the test signal set. Subjects were instructed to identify the “apparent distance” of the perceived sound from the head and respond (in cm) on a custom GUI developed in MATLAB. Before the experiment started, subjects were asked to stretch their hands and the distance between the head and the tip of the hand was given as a distance reference to the subjects. The raw stimuli used for the experiment was a burst of 4 different WGN sequences, each of duration 300 ms, and separated by an interval of 30 ms. A 30 ms cosine-square gating window is applied to the stimuli in order to avoid the transient effects. The measured distance-dependent HRTFs (Chapter 3) were used to synthesize the virtual sounds. The distances simulated in this experiment were 0.35 m, 0.45 m, 0.50 m, 0.60 m, 0.75 m, 0.80 m, 1 m, and 1.4 m. In this experiment, the virtual sounds were synthesized for three lateral directions 40° , 60° , and 80° in

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

order to include the presence of sufficient interaural cues to aid distance perception. This is reasonable, as it was not intended to study the variation of distance perception with azimuth in this experiment. In all the experiments, three stimulus trials for every distance and angle were presented, making a total of 216 stimuli presentations (8 distances $\times$ 3 directions $\times$ 3 trials $\times$ 3 headphones) for each subject. It should be noted that the stimulus spatialized with distance-dependent HRTFs might display certain distance-dependent level differences that can bias the subjects. Thus, the stimulus was normalized to compensate for the sound pressure level variation with distance. The normalization factor (N) scales the stimulus based on the distance of the source with respect to the left and right ears. The correction is such that the distance at 1 m has a scaling factor of 1 [219]. The normalization factor used in this experiment was:

$$N = \frac{1}{\frac{50}{\text{Distance to left ear (cm)}} + \frac{50}{\text{Distance to right ear (cm)}}}. \quad (5.1)$$

Zahorik [220] also found that a compressive power function is a good fit to individual listener data under a variety of stimulus conditions:

$$r_{res} = k \cdot (r_{actual})^m, \quad (5.2)$$

where r_{res} is the estimated response distance, r_{actual} is the actual or the target distance, k and m are the fit parameters of the power function. In addition, the effect of repositioning of the headphone is studied. The three experiments and the results are presented in Sections 5.4, 5.5, and 5.6.

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

Figure 5.3: Stimulus-Response pairs for distance (in centimeters) using Non-individualized HRTFs for the two cases a) without and b) with equalization. The data points are shown on a log-log scale. The solid line in each panel of the figure represents the best linear fit of the stimulus location to the response location. Also shown is the correlation (ρ) between the log-stimulus and the log-response distances and the compressive function fit parameters k and m .

5.4 Experiments : Phase 1

In this experiment, the distance perception was evaluated for 15 subjects based on the virtual sounds synthesized using the B & K HATS dummy head (non-individualized) HRTFs. There were two sets of stimuli presented to the listeners in this experiment, which were as follows:

1. SET 1: Non-Individual HRTF synthesized sound without headphone equalization.
2. SET 2: Non-Individual HRTF synthesized sound with non-individualized head-

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

phone equalization.

The overall experiment lasted for 30 minutes for each subject. The responses of the subjects were collected and were plotted in log-log scale for analysis Fig. 5.3. It was found that the distance perception is poor when the headphone is not equalized (Fig. 5.4a). The correlation ρ between the log-stimulus and log-response distance was found to be 0.4, 0.46, 0.49 for the closed-back, open-back and open headphones, respectively. Moreover, it was observed that the subjects under-estimated the distance. There were some cases where the subjects localized the virtual sound inside the head. However, the distance perception improved when the headphone was equalized (Fig. 5.4b). The correlation ρ improved to 0.61, 0.62, 0.65 for the closed-back, open-back and open headphones, respectively. Subjects, however, over-estimated the distance in the near-field and under-estimated the distances in the far-field. Externalization improved with the equalization of the headphones, but there were still cases of IHL. There were some improvements in the distance perception using the open-back and open headphones, but the improvement was only marginal. ANOVA analysis is carried out in Section 5.7 to study the influence of the type of headphones used.

5.5 Experiments : Phase 2

In this experiment, auditory distance perception was evaluated for the three subjects using virtual sounds synthesized from individualized HRTFs. In this experiment, three sets of stimuli were presented to the listeners, which consist of:

1. SET 1: Individualized HRTF synthesized sound without any headphone equalization.
2. SET 2: Individualized HRTF synthesized sound with non-individualized headphone equalization.

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

Figure 5.4: Stimulus-Response pairs for distance (in centimeters) using Individualized HRTFs for the three cases: a) No equalization, b) Non-individual equalization, c) Individual equalization. The data points are shown on a log-log scale. The solid line in each panel of the figure represents the best linear fit of the stimulus location to the response location. Also shown is the correlation (ρ) between the log-stimulus and the log-response distances and the compressive function fit parameters k and m .

3. SET 3: Individualized HRTF synthesized sound with individualized headphone equalization.

The three sets differ in the type of equalization applied. In the first set, equalization

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

of the headphone was not carried out, while the other two sets have non-individualized and individualized equalization, respectively. The responses were plotted in log-log scale as in the previous experiment. We find that the distance perception is poor even in the case of using individualized HRTFs, when equalization is not applied (Fig. 5.4a). The correlation ρ between the log-stimulus and log-response distance is found to be 0.51, 0.54, 0.56 for closed-back, open-back and open headphones respectively, which is similar to that obtained with non-individualized HRTFs (Fig. 5.4b). Similar to the previous experiment, subjects under-estimated the distance perception and experienced IHL. For Set 2 with non-individualized equalization, the distance perception improved with a correlation ρ of 0.69, 0.71, 0.74 for the closed-back, open-back, and open headphones, respectively. With the use of individualized equalization (Fig. 5.4c), the distance perception of the subjects remained almost the same with a correlation of 0.73 (closed-back), 0.74 (open-back) and 0.78 (open). In the presence of equalization, subjects over-estimated the distance in the near-field and under-estimated the distance in the far-field. This kind of bias was also observed by other researchers, who studied distance perception in the free-field anechoic conditions [220]. The improvement in the distance perception using an open-back or an open headphone is only marginal; however, the in-head localization of the sound is lesser compared to the closed-back headphones.

5.6 Effect of headphone repositioning

In this study, only individualized HRTFs with individualized equalization were investigated to avoid the effect of other factors like IHL and timbral coloration. It was observed that the repositioning of the headphones did not have any noticeable effect on the auditory distance perception (Fig. 5.5). The correlation ρ values without repositioning of the headphone (Fig. 5.5a) were found to be 0.72 (closed-back), 0.73 (open-back) and 0.74 (open). With repositioning, the correlation values were of the similar order 0.74

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

Figure 5.5: Stimulus-Response pairs for distance (in centimeters) using Individualized HRTFs and Individualized equalization to investigate the effect of headphone repositioning. a) Without repositioning b) With repositioning. The data points are shown on a log-log scale. The solid line in each panel of the figure represents the best linear fit of the stimulus location to the response location. Also shown is the correlation (ρ) between the log-stimulus and the log-response distances and the compressive function fit parameters k and m .

(closed-back), 0.72 (open-back) and 0.79 (open) headphones (Fig. 5.5b). Thus, it seems that the repositioning of the headphone does not affect the auditory distance perception of the sound.

5.7 Analysis of variance

An ANOVA analysis [215] is carried out to investigate and extend the perceptual results based on the sample population to the whole population. For the first experiment,

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

the effect of equalization on the distance perception is first considered. In the experiment using non-individualized HRTFs, the effect of equalization is found to have a significant effect [$F(1, 14) = 49.33, p \ll 0.01$] on the auditory distance perception. For the case of individualized HRTFs, the effect of applying non-individualized equalization is significant [$F(1, 7) = 90.35, p = 0$] compared to the case when no equalization is applied. Similarly, the effect of applying individualized equalization with respect to no-equalization is significant [$F(1, 7) = 268, p \ll 0.01$]. However, the effect of type of equalization (non-individualized or individualized) does not have any significant effect on the distance perception. The ANOVA test also suggests that the difference in means for the case of non-individualized and individualized equalization is insignificant [$F(1, 7) = 0, p = 0.95$].

The influence of the type of headphone (closed-back, open-back and open) is also investigated with ANOVA tests. For the case of using non-individualized HRTFs without any equalization, the mean differences between the closed-back and open headphones are found to be significant [$F(1, 7) = 17.84, p = 0.003$]. However, the mean differences for both the pairs (closed-back, open-back) headphones and (open-back, open) headphones are found to be insignificant ($p > 0.01$). When non-individualized equalization is used, the ANOVA analysis returned insignificant mean differences between all the three different types of headphones ($p > 0.01$). In the case of individual HRTFs, the differences between the three types of headphones is found to be insignificant for all the cases: No equalization [$F(2, 7) = 2.91, p = 0.08$], Non-individualized equalization [$F(2, 7) = 0.05, p = 0.95$], Individualized equalization [$F(2, 7) = 0.8, p = 0.4$]. An ANOVA analysis is also carried out in-order to investigate the effect of repositioning of the headphone in distance perception. It was found that repositioning of the headphone does not have a significant main effect [$F(1, 7) = 6.27, p = 0.04$] on the perception of distance.

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

Table 5.1: Summary of results

Experiment	Experimental condition	Target - Response Correlation (ρ)			Remarks
		Closed- Back	Open- Back	Open	
PHASE 1 Non - Individualized HRTFs	No EQ	0.40	0.46	0.49	Non - Individual EQ significantly improves distance perception No EQ : Effect of open headphones is significant
	Non-Individualized EQ	0.61	0.62	0.65	Non-Individual EQ : Type of headphones have no significant effect
PHASE 2 Individualized HRTFs	No EQ	0.51	0.54	0.56	EQ (Individual or non-individual) significantly improves distance perception
	Non- Individualized EQ	0.69	0.71	0.74	No significant difference between Non-individual and Individual EQ
	Individualized EQ	0.73	0.74	0.78	Type of headphones : No significant effect
Headphone Repositioning (Individualized HRTF and equalization)	No Repositioning	0.72	0.73	0.74	Repositioning : No significant effect
	Repositioning	0.74	0.72	0.79	Type of headphones : No significant effect

5.8 Discussion

In this study, the effect of headphone equalization on the distance perception in anechoic conditions was investigated with the help of detailed psychoacoustical experiments. The major results reported in this study are reported in Table 5.1. Subjects were presented virtual sounds using both individualized and non-individualized distance dependent HRTFs with equalization as a variable factor. It was found that equalization of the headphone is critical for consistent distance perception.

Absence of equalization of the headphone led to in-head localization and thus, the subjects under-estimated the distance of virtual sound source in both the near-field as well as far-field. This was the case irrespective of whether the virtual sound was synthesized using non-individualized or individualized HRTFs. On equalizing the headphone,

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

the distance perception improved with externalization of the sound for both the individualized and non-individualized equalization. Subjects over-estimated the distance in the near-field and under-estimated the distance in the far-field when equalization was used. It was also seen that the type of equalization (individualized or non-individualized) had no significant effect on the distance perception. However, the use of non-individualized equalization led to in-head localization of sound, which was minimal, when individualized equalization was used. These results obtained are also consistent with the study by Kim *et al.* [28], who found that the subjects could perceive the sound well externalized, when individualized equalization was employed. The repositioning of the headphones also does not have any significant effect on the distance perception.

All these results suggest the lack of dependence of auditory distance perception on the high-frequency spectral cues as found in other experiments investigating the distance perception in anechoic conditions [43]. The highly idiosyncratic pinnae due to its small size affect the high frequencies, which are found to be critical for sound localization. This reasoning is also evident from the fact that the type of equalization (non-individual or individual) does not affect the distance perception. Furthermore, distance perception using virtual sounds synthesized by non-individualized HRTFs have similar distance localization performance compared to those using individualized HRTFs. These results are also consistent with Brungart's experiments with a 3 kHz center frequency low-pass filtered noise [43]. Brungart [43] noticed that the distance perception using the low-pass filtered noise is similar to that using broadband noise stimulus. This also hints that the low-frequency ILD cues are most critical for distance perception. An unequalised headphone would distort the low-frequency ILD and thus, would affect the auditory distance perception. Equalization (non-individual or Individual) ensures that the low-frequency features of the HRTF spectrum are not distorted at the eardrum thereby improving the distance perception. Since the low-frequency features of the HPTF/HRTF are not id-

CHAPTER 5. EFFECT OF HEADPHONE EQUALIZATION ON DISTANCE PERCEPTION

iosyncratic, any equalization at the low frequencies can be easily compensated either by non-individualized or individualized equalization. All these experiments point to the fact that the compensation of the headphone in the low frequencies is sufficient to maintain the distance perception; however, the high-frequency HRTF features are critical for localization perception.

These results are very useful and is used in the modeling of the distance-dependent individualized HRTFs using the frontal projection headphones in the next chapter. In particular, the role of interaural and auditory parallax cues for distance perception is investigated in this study. The distance-dependent individualized HRTFs are subsequently modeled using the frontal projection headphones along with a cascade of other signal processing techniques.

Chapter 6

Modeling distance-dependent individualized HRTF using frontal projection headphones

In Chapter 5, we studied the effect of headphone equalization on the distance perception. It was found that the equalization of the low-frequency response of the headphone is critical for good distance perception. Moreover, the type of equalization (individual or non-individual) did not have a significant effect on the auditory distance perception indicating that the distance perception does not depend on the idiosyncratic features. These results are important for the modeling of the distance-dependent individualized HRTFs using the frontal projection headphones.

The frontal projection headphones can effectively capture the individualized pinna cues during playback (Chapter 4). The individual spectral cues created by frontal projection superposed with the non-individualized spectral cues could reduce the front-back confusions, but will lead to spectral coloration. Spectral flattening experiments proved that the individualized spectral cues due to frontal projection are self-sufficient in re-

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

ducing the confusions. In this chapter, the individualized HRTFs are directly modeled using the frontal projection headphones. The modeling of the individualized HRTFs is extended to the near-field HRTFs in this chapter. This work on the modeling of distance-dependent individualized HRTFs has been accepted for publication in the Journal of Acoustical Society of America [221].

Measuring these individualized HRTFs is an immense task, which demands very precise and tedious experiments to be carried out for a number of speaker locations around the head in the 3D space. Moreover, unlike the far-field HRTFs [45–47, 51, 55, 126], the proximal region or the near-field ($<1\text{m}$) HRTFs vary significantly with distance [43]. Thus, the HRTFs have to be measured for a number of distances in the near-field for every azimuth and elevation.

Due to the requirement of an acoustic point source and the difficulty in measuring individualized HRTFs, only few researchers have measured individualized proximal region HRTFs. These issues call for simpler techniques to obtain the distance dependent individualized HRTFs. In order to model the distance-dependent individualized HRTFs in this study, the perceptual cues that affect distance perception are first identified and simulated. Subsequently, a suitable HRTF individualization technique is developed.

6.1 Literature survey

In order to develop the distance-dependent individualized HRTF model, a background on the various auditory cues, existing models of distance perception, and HRTF individualization techniques is necessary, which are outlined in the following sections.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

6.1.1 Auditory cues for distance perception

Auditory distance perception has been an area that has eluded scientists for several years. The ability of human beings to estimate distance is much more complex and displays lesser accuracy than directional perception [12]. It is found that the listeners often overestimated the distance in the near-field and underestimated the distance beyond 1 m in free-field anechoic conditions [222]. The perceived distance increases with the actual distance and then, saturates beyond a point in the far-field [220, 223]. This phenomenon was coined as “Acoustic Horizon” by Von Békésy [223]. Intensity, interaural level differences, spectral cues, and the direct-to-reverberant energy ratio are found to be the most important cues for distance perception [12].

Intensity cues: Intensity as a potential distance cue has long been investigated in various studies [222]. It has been found that the perceived distance is found to increase less rapidly than the physical distance for distances greater than 1m when intensity is the only prominent cue [223–225]. The 6-dB loss in sound pressure for every doubling of distance does not hold anymore in the near-field. It was also shown that a difference greater than 6 dB is required for doubling distance in perceived distance in the proximal region [220]. There have been other studies investigating loudness as a possible distance cue [226, 227]. Under conditions where intensity is the only prominent cue, loudness varies inversely with perceived distance and a 10 dB decrement is required for sensation of half loudness or doubling distance [226]. In fact, when there are multiple cues (diffused sound field) at the disposal of the listener, the loudness of the sound source has been found to remain constant under distance-varying conditions [227]. The existence of loudness constancy is attributed to the reverberant sound energy independent of the distance [227].

Binaural cues: Binaural cues play a critical role in the human distance perception in

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

anechoic conditions in the near-field. At low frequencies, ILD and the ITD values can be well approximated by the spherical head model [43, 50, 172, 173, 228–232]. It is noted that the ILD values rapidly increases with decrease in distance, while the ITD values increases only modestly [43]. This is typically because the ITD value depends on the difference in distance, while the ILD depends on the ratio of the distance between the source and the two ears [11]. It is also speculated that the listeners may make use of the distant invariant ITDs to first determine the azimuth location and then use the ILD cues to estimate the distance [43, 233]. Furthermore, the ILD of a lateral source increases by 15 or more jnds when the source distance decreases from 1 m to 12 cm, while the ITD increases only by 2-3 jnds [43]. Brungart [219] studied localization response for different stimulus conditions and found that the distance accuracy in the low pass condition with cut-off frequency at 3 kHz was similar to that in the broadband condition. This result along with the poor distance perception in monaural conditions indicates a strong dependence of the distance perception in the low-frequency ILD cues in anechoic listening conditions [219].

6.1.2 Models of distance-dependent HRTFs

Based on the perceptual cues identified for distance perception, several models for distance perception have been developed, which can be grouped as follows:

Using binaural cues: Stewart [228] first derived mathematical equations for ILD and ITD that originate in the near-field on a spherical head. Using Stewart's derivations, Hartley and Fry [229] also manually tabulated ILDs and ITDs at different distances and found that a combination of distance-invariant ITDs and distance varying ILDs could account for the binaural cues in near-field. Hirsch [234] presented a model on binaural distance perception allowing a listener to determine the distance of the

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

sound source from the ILD and ITD. However this model ignores the effects of sound diffraction by the head and the ears are represented by point receivers in free space. Molino [235] further modified the model by approximating the head by a rigid sphere, thereby eliminating the free-field condition as in Hirsch's model. Tahara and Sakurai [236] proposed a model similar to Hirsch's model that simulates the changes in the distance perception based on the simultaneity of the changes of the ITD, ILD as a function of distance as well as azimuth angle. A new model for binaural distance perception was later developed [237] based on the ILD measurements of a 500 Hz tone on a KEMAR manikin and the jnd for a 500 Hz tone measured by Hershkowitz and Durlach [238].

Using Auditory parallax: Kim *et al.* [225] modeled the distance perception up to 2 m using the parallax angle information available in the HRTFs as previously investigated by Brungart and Rabinowitz [43]. The “auditory parallax model” was simulated using a simple geometrical model defining the relation between the angle with respect to ears and the head [225]. It was found that the perceived distance of the virtual sound simulated by the auditory parallax model was similar to the free-field listening condition. However, in this study, the virtual sounds were reproduced using individualized HRTFs and played back from loudspeakers.

Simulating the spectral cues directly: Duda and Martens [172] developed an algorithm to compute the sound pressure at the surface of the spherical head for a source at various distances. Rabinowitz *et al.* [239] later studied the frequency scalability of HRTFs for an enlarged head, which maintained a fixed head size and varied the distance from the source. Ze-Wei *et al.* [240] modified the model developed by Rabinowitz *et al.* [239] and analysed the proximal region HRTFs by adding the neck and the torso. Interestingly, it was found that the human neck influences the near-field HRTFs but its perceptual effects are still unclear. Though

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

these models give an idea about the spectral variation in proximal region HRTFs, it cannot be used directly to synthesize binaural audio as the spectral effects of the external ear are missing. Shaw and Teranishi [116] first studied the pressure generated by a nearby point source using a rubber replica of the ear, which holds the pinna, concha and auditory meatus with dimensions comparable with those of real human ears. Shin and Park [232] used minimum realisation and Shanks' method to model the measured non-individualized HRTFs. It was found that the HRTF models obtained by the minimum realization method were able to match to more than 99% of the measured HRTF and was much more systematic and consistent compared to the Shanks' method [232].

Simulation from far field HRTFs: Duraiswami *et al.* [241] modeled the near-field HRTFs by range extrapolation of measurements at one distance. This was carried out by expressing the HRTF in terms of a series of multi-pole solutions of the Helmholtz equation. Menzies [242] also calculated the near-field HRTF from the far-field HRTFs using point source expansions into plane waves. Romblom and Cook [243] used difference filters between the near and far-field derived from the spherical head model and used a HRTF look up table to calculate the near-field HRTFs. However, perceptual results were not reported in this study. Kan *et al.* [244] used a similar method and simulated the near-field HRTFs by applying the distance variation function (DVF) to the individualized far-field HRTFs. The DVF for a particular location is calculated by taking the ratio of the HRTFs between the far-field and the near-field obtained using the spherical head model, which is the same as the difference filter used in the study by Romblom and Cook [243]. Kan *et al.* [244] showed that the DVF technique improved the distance perception compared to a simple intensity adjustment. The directional perception in this study was reported to be good using the modelled near-field HRTFs with less front-back

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

confusions as individualized HRTFs were used.

Furthermore, it is interesting to understand the effect of idiosyncrasy on the distance perception. Zahorik [245] noted that distance localization using non-individualized HRTFs is similar to those using individualized HRTFs in the presence of other distance cues. Brungart [219] observation of good distance perception using low-pass filtered stimuli further shows the lack of dependence of distance perception on the idiosyncratic spectral cues. On the contrary, the directional perception is degraded by the use of non-individualized HRTFs. The highly individual pinna reflections are associated with the high-frequency features of the HRTF and absence of these cues would lead to front-back and up-down reversals [9, 106]. Thus, even though distance perception is less dependent on the individual features, individualized characteristics in the HRTF are extremely important for accurate reproduction of 3D audio. Therefore, in this work, the range-dependent individualized HRTFs are modeled that simulates accurate localization and distance perception as with the listener's own HRTFs. In this chapter, we model the distance-dependent individualized HRTFs in two stages, as follows:

1. The most prominent cues for auditory distance perception in anechoic conditions are first identified, when played back over headphones. The role of auditory parallax and ILD cues are of particular interest. Using each of auditory parallax and ILD cues, distance-dependent HRTFs are synthesized from the measured individualized far-field and measured non-individualized far-field HRTFs. Perceptual experiments are carried out with each of these synthesized HRTFs and their results are compared with those obtained using the measured distance-dependent individualized HRTFs (Section 6.2.1).
2. Once the important cues that affect distance perception are identified (Section 6.2.1), the distance-dependent individualized HRTFs are modeled using a cascade of

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

signal processing blocks. The distance-dependent spectral cues due to head effects are simulated by the spherical head model [172]. Individualized pinna spectral cues are then modeled by playback through frontal projection headphones [144].

Unlike other individualization methods, the frontal projection headphones playback method does not require the measurement of individualized far-field measurements; nor does it require the anthropometric data of human subjects. Frontal projection headphones have a unique structure, as they project sound from the front directly onto the pinnae of the listener during playback, unlike the conventional side projection headphones. In Sunder *et al.* [144], the authors observed that the frontal projection headphones inherently create individualized pinna spectral features of HRTF during sound playback. The frontal projection headphones also help in reducing front-back confusions to a large extent, thereby enhancing the frontal perception of the virtual sound [144]. In addition, headphone equalization, low-frequency compensation, and rear-directional filters are required to model the distance-dependent individualized HRTFs.

6.2 Role of Auditory Parallax and ILD Cues in distance perception

Auditory parallax and ILD cues are often considered as relevant cues contributing to the near-field distance perception in the absence of reverberant and intensity cues. Though the effect of ILD cues on distance perception is well studied, the perceptual effects of auditory parallax cues are still unclear. In this section, we investigated the role of each of these cues-individually and combined, with the help of listening experiments. The distance-dependent HRTFs are first synthesized using ILD and auditory parallax cues from the measured far-field HRTFs.

HRTF synthesis using auditory parallax cues: In the near-field, the angle between

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

the source and the ears ($\theta_{L,R}$) deviates significantly from that between the source and the center of the head (θ). This is known as auditory parallax effect. The auditory parallax effect can be approximately modeled by a simple geometrical formulation that relates both these angles for any source to head distance [225]. The left and right ear HRTFs in the near-field for a source at distance d_n from the head and at an azimuth angle of θ can be obtained by remapping it from the far-field HRTFs at angles θ_L and θ_R , respectively.

$$\theta_R = \tan^{-1} \left(\frac{(d_n \sin \theta - a)}{d_n \cos \theta} \right), \quad (6.1)$$

$$\theta_L = \tan^{-1} \left(\frac{(d_n \sin \theta + a)}{d_n \cos \theta} \right), \quad (6.2)$$

$$HRTF_{AP,L}(d_n, \theta) = HRTF_L(d_f, \theta_L), \quad (6.3)$$

$$HRTF_{AP,R}(d_n, \theta) = HRTF_R(d_f, \theta_R), \quad (6.4)$$

where $HRTF_{AP}$ is the near-field HRTF synthesized using auditory parallax (AP) cues, a is the radius of the head, and d_f is the source to head distance in the far-field. θ_L and θ_R correspond to the angles of the sound source as seen from left and right ears in the near-field, respectively. All the angles (θ , $\theta_{L,R}$) are measured with respect to the mid-sagittal axis passing through the center of the head. It is important to note that Eqs. (6.4) and (6.5) model only the high-frequency distance-dependent features of the outer ear.

HRTF synthesis using ILD cues: The rapid rise of ILD cues with decreasing source to head distance in the near-field can be modeled using the spherical head model [172]. These distance-dependent ILD features in the near-field are obtained by computing the ratio of the pressure at the surface of the sphere arising from a sound source at distance d_n to the pressure arising from a sound source at a distance d_f in the far-field. This function, also known as the distance variation function

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

(DVF) [244], can be computed as follows:

$$DVF(\theta, n, f) = \frac{p_n(a, \omega, \theta, d_n)}{p_f(a, \omega, \theta, d_f)}, \quad (6.5)$$

$$HRTF_{ILD}(d_n, \theta) = DVF(\theta, n, f) \cdot HRTF(d_f, \theta), \quad (6.6)$$

where $HRTF_{ILD}$ is the near-field HRTF synthesized using ILD cues, a is the radius of the head, θ is the azimuth angle with respect to the center of the head, ω is the frequency in radians, p_n and p_f are the pressure in near and far-field as obtained from the spherical head model, respectively, d_n and d_f are the distances in the near-field and far-field, respectively.

HRTF synthesis using both ILD and auditory parallax cues: The synthesized HRTF containing both ILD and auditory parallax cues ($HRTF_{AP+ILD}$) can be obtained by cascading the auditory parallax and DVF models. The pinna spectral features are first remapped according to the parallax angles to obtain the HRTFs containing auditory parallax cues. The DVF model is then applied to $HRTF_{AP}$ to further incorporate the distance-dependent ILD cues, as given in the equation below:

$$HRTF_{AP+ILD}(d_n, \theta) = DVF(\theta, n, f) \cdot HRTF_{AP}(d_n, \theta). \quad (6.7)$$

The synthesized distance-dependent HRTFs were also compared with the measured HRTFs using objective and subjective measures. The low-frequency spectral cues (0.1 kHz to 1 kHz) and the high-frequency pinna spectral cues (5 kHz to 16 kHz) of the synthesized and measured HRTFs were compared by calculating the root mean

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

square spectral difference (RSD) as an objective criterion:

$$RSD_{low,high} = \sqrt{\frac{1}{N} \sum_{f=f_l}^{f_h} \left(20 \log_{10} \left| \frac{HRTF_{measured}(f)}{HRTF_{modeled}(f)} \right| \right)^2}, \quad (6.8)$$

where f_l, f_h indicate the frequency range of comparison, N represents the number of frequency bins between f_l and f_h . In this paper, RSD_{low} is computed for $f_l = 0.01kHz$ and $f_h = 1kHz$, while RSD_{high} is computed for $f_l = 5kHz$ and $f_h = 16kHz$.

The DVF mainly models the low-frequency distance-dependent spectral effects of the head. However, the DVF model does not simulate any of the distance-dependent spectral effects of the pinna. On the other hand, the auditory parallax effect only models the distance-dependent effects of the pinna but does not simulate the spectral effects due to head.

From Fig. 6.1, it can be seen that the HRTFs synthesized using auditory parallax cues fail to simulate the low-frequency near-field spectral cues well ($RSD_{low} > 6dB$). However, the HRTFs synthesized using auditory parallax cues model the high-frequency pinna cues reasonably well ($RSD_{high} = 2dB$). The low-frequency spectral cues in the measured near-field HRTFs are modeled accurately ($RSD_{low} = 2dB$) by the DVF model. However, there is a mismatch of small degree at high frequencies in the case of HRTF synthesis using DVF model ($RSD_{high} < 4dB$). We note that the HRTFs synthesized using both DVF and auditory parallax models match with the measured HRTFs much better ($RSD_{low} = 1dB$ and $RSD_{high} = 4dB$), as illustrated in Fig. 6.1. More importantly, it was observed that the auditory parallax contributes little to the synthesized distance-dependent HRTFs. This can be further analysed with the help of detailed subjective experiments, as shown in the next section.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

Figure 6.1: Comparison of measured and modeled proximal range HRTFs. Column (a) HRTF at a distance of 35cm and 50° azimuth; Column (b) HRTF at a distance of 45cm and 60° azimuth.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

6.2.1 Subjective experiments

The roles of auditory parallax and of ILD cues for distance perception using both individualized and non-individualized HRTFs were investigated with the help of perceptual experiments. A total of 15 subjects participated in the experiments carried out using non-individualized HRTFs. HRTFs measured on the B & K HATS dummy head were used as non-individualized HRTFs in this experiment. Three subjects, whose personal HRTFs were measured, participated in the listening experiments using individualized HRTFs. All subjects had normal hearing. They were aged between 18 and 30 years, and had past experience in similar tests. A double blind procedure was adopted whereby neither the evaluator nor the subjects were aware of the order and nature of the test stimuli. The stimuli were sequences of four different 300-ms bursts of WGN separated by 30-ms intervals of silence. Stimuli were gated with cosine-squared ramps of 30-ms onset and offset in order to avoid transient effects. The gated stimuli were then convoluted with the synthesized HRTFs obtained using both individualized and non-individualized far-field HRTFs. The four stimulus conditions used in this experiment are shown below:

1. Synthesized distance-dependent HRTFs containing auditory parallax cues ($HRTF_{AP}$).
2. Synthesized distance-dependent HRTFs containing binaural ILD cues ($HRTF_{ILD}$).
3. Synthesized distance-dependent HRTFs containing both auditory parallax and ILD cues ($HRTF_{AP+ILD}$).
4. Measured distance-dependent HRTFs ($HRTF_{measured}$).

For the first three stimulus conditions, virtual sounds were synthesized for 9 distances at 0.2 m, 0.3 m, 0.4 m, 0.5 m, 0.6 m, 0.7 m, 0.8 m, 1 m, and 1.4 m. Stimuli

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

synthesized from twenty directions (0° to 340° at a resolution of 20° including those at 90° and 270°) were presented in a random order for each of the 9 distances. For the experimentally measured distance-dependent individualized and non-individualized HRTFs, the simulated distances (8 distances) were 0.35 m, 0.45 m, 0.50 m, 0.60 m, 0.75 m, 0.80 m, 1 m, and 1.4 m, as the HRTFs were measured only for these distances. The synthesized stimuli were then normalized to eliminate amplitude-based distance cues. In the near-field, the distance between source and the ears deviate significantly from the distance between source and the center of the head. The scaling factor used was based on the distance of the source from the left and right ears of the subject, similar to that adapted by Brungart [43]. The correction was such that a distance of 1 m had a scaling factor of 1. The scaling factor (N) was:

$$N = \frac{1}{\frac{50}{\text{Distance to left ear (cm)}} + \frac{50}{\text{Distance to right ear (cm)}}}. \quad (6.9)$$

Amplitude scaling was carried out by multiplying the noise waveform with the scaling factor prior to playback. In addition to distance normalization, source amplitude was randomized in order to ensure that subjects' perception was based only on the spectral features of HRTF and not biased by source levels [43]. To achieve this, the source amplitude was roved randomly by an additional 15 dB (from 0 to 15 dB in steps of 1 dB). Since the maximum amplitude of stimulus was approximately 60 dBA at 1 m, the effective stimulus amplitude ranged from 45 dBA to 60 dBA after randomization and correction due to scaling factor.

To ensure consistency, three measurements were made for each of the 9 distances and 20 azimuth angles making it a total of 540 stimulus presentations ($20 \text{ directions} \times 9 \text{ distances} \times 3 \text{ trials}$) for modeled HRTFs, and 480 stimulus presentations ($20 \text{ directions} \times 8 \text{ distances} \times 3 \text{ trials}$) for measured HRTFs. All the listening experiments were conducted

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.2: Headphone responses of Sennheiser HD600 and the frontal projection headphone measured on the blocked ear-canal of a B & K HATS dummy head.

in a semi-anechoic chamber of dimension 5.5m (length) $\times$ 3.1 m (width) $\times$ 2.4 m (height). The level of the background noise in the semi-anechoic chamber was 19 dBA.

Synthesized virtual sounds were played back using Sennheiser HD600 headphones. Headphone equalization was carried out prior to start of experiment to compensate for the non-uniform headphone response. The headphone response of Sennheiser HD600 is shown along with the frontal projection headphone response in Fig. 6.2. A graphical user interface (GUI) was used to present the stimuli in a random order to the subject. Before the experiment started, subjects were asked to stretch their hands and the distance between the head and the tip of their hands was taken as reference. For each condition, they had to identify the “apparent distance” of the sound from the head. Once the stimulus was presented, subjects were prompted to record their judgement of distance

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

(in cm). They were allowed to remove and reposition the headphones on their head in between trials. However, they were instructed to minimize any head movement during the experiment. Subjects were not given feedback on their responses.

6.2.2 Results

The main objective of this experiment was to examine the role of auditory parallax and ILD cues on distance perception over headphones in anechoic conditions. Distance perception was separately investigated for lateral (40° to 140° and 220° to 320° clockwise) and near-medial 340° , 0° , 20° in the front and 160° , 180° , 200° in the rear) directions. Distance responses using non-individualized and individualized HRTFs for all the four stimulus conditions are shown in Figs. 6.3 and 6.4, respectively. The responses are plotted in Figs. 6.3 and 6.4 on a logarithmic scale because the magnitude of distance errors tended to increase with increasing source to head distance. Two different measures were used to evaluate the response accuracy:

1. Correlation coefficient between the logarithm of actual distances and the response distances.
2. Slope of line best fitting the log-stimulus data to the log-response data.

Zahorik [222] and Zahorik *et al.* [220] found that the following compressive function represents a good fit for listeners' distance responses under a variety of stimulus conditions:

$$r_{res} = k.(r_{actual})^m, \quad (6.10)$$

where r_{res} is the estimated response distance, r_{actual} is the actual distance, k and m are the fit parameters of the power function. Distance responses of the subjects were fitted with Eq. 6.10. The value of exponent m was typically less than one and the value

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

of constant k was close to one, thereby indicating a strong perceptual bias in distance estimation.

It was observed that there was a general tendency to overestimate distances in near-field and underestimate distances in far-field. The distance perception using HRTFs synthesized with the auditory parallax model was poor for both lateral and near-medial directions (Fig. 6.3a). Using non-individualized HRTFs (Fig. 6.3), the mean correlation coefficient values between the estimated and actual distances were 0.31 for lateral and 0.30 for near-medial directions, respectively. Similarly, for individualized HRTFs (Fig. 6.4a), the correlation coefficient values obtained for distance responses with auditory parallax modeled HRTFs were 0.34 for lateral and 0.31 for near-medial directions, respectively. Both these observations are consistent with the results of Brungart (1999a), who observed poor correlation coefficient values between the perceived and the simulated distance for monaural and high-pass filtered stimuli. This finding suggests that auditory parallax cues alone are not sufficient for good distance perception.

With the DVF-synthesized HRTFs obtained using non-individual HRTFs, distance perception improved with a correlation coefficient value of 0.59 for lateral directions (Fig. 6.3b). However, the correlation coefficient value was poor (0.24) for the near-medial directions (Fig. 6.3b). This is due to the absence of sufficient binaural cues for positions near the medial plane. The distance responses with the DVF synthesized HRTFs obtained using individualized HRTFs (Fig. 6.4b) was found to be similar to that using non-individual HRTFs. The corresponding correlation coefficient values for the DVF synthesized HRTFs using individualized HRTFs were found to be 0.61 and 0.25 for lateral and near-medial directions, respectively.

When both auditory parallax and distance-dependent ILD cues were present in HRTF, subjects were able to localize distances better compared to the case when only auditory parallax cues were present in HRTF. The correlation coefficient values for distance re-

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.3: Stimulus-response pairs for distance (in cm) using non-individualized HRTFs. Note the log-log scale. The solid line in each panel represents a linear regression fit. The dashed line in each panel corresponds to the “correct” responses. Also shown is the correlation coefficient (ρ) between the log-stimulus and the log-response distances and the compressive function fit parameters k and m (Eq. 6.10). Left column corresponds to the lateral angles and the right column represents the data from the near-medial angles. Note that only the distance responses greater than 15 cm are displayed for visual clarity.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

Figure 6.4: Stimulus-response pairs for distance (in cm) using individualized HRTFs. The layout is as in Fig. 6.3.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

sponses with the modeled HRTFs containing auditory parallax and ILD cues were found to be 0.63 (lateral), 0.34 (near-medial) using non-individualized HRTFs (Fig. 6.3c), and 0.64(lateral), 0.32(near-medial) using individualized HRTFs (Fig. 6.4c). With the measured HRTFs, the correlation coefficient values were found to be 0.69(lateral), 0.41(near-medial) using non-individualized HRTFs, and 0.72(lateral), 0.46(near-medial) using individualized HRTFs. These values indicate that the distance localization with the modeled HRTFs containing distance-dependent binaural ILD cues was comparable to that with measured HRTFs (Fig. 6.3d). However, some in-head localization of sound was observed in distance responses for all the stimulus cases using both non-individualized and individualized HRTFs, as indicated in Figs. 6.3 and 6.4. Distance responses reported by the subjects that are less than 15 cm are not shown in Figs. 6.3 and 6.4 for the sake of clarity.

Figure. 6.5 shows the correlation coefficient between the logarithm of actual and responded distances, and the slopes of the linear regression line fitting the log-stimulus data to the log-response data. The correlation coefficient values and slopes are plotted for all the azimuth directions between 0° and 180° . Correlation coefficient values were poor across all the azimuth angles with the auditory parallax modeled HRTFs using both non-individualized HRTFs and individualized HRTFs. For other stimulus cases, the correlation coefficient values increased for lateral angles due to the presence of ILD cues. The slope indicates the scaling bias in responses. A slope of less than one indicated that responses varied over a narrow range of distances than the actual locations, showing compression in the responses. Similar to the correlation coefficient, the slope was relatively high for lateral locations and low for near-medial locations. This was true for all the stimulus conditions except for the stimulus condition synthesized with auditory parallax modeled HRTFs. Due to the presence of sufficient ILD cues in the lateral directions, subjects tended to be more sensitive to the actual source location in these

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.5: Top: Correlation coefficient between the logarithm of actual and responded distances; Bottom: Slope of the linear regression line best fitting the log-stimulus data to the log-response data for azimuth angles (between 0° and 180° for both (a) Non-individualized HRTFs, and (b) Individualized HRTFs.

directions resulting in a high correlation coefficient value between the responded distance and the actual distance. The slope was less than one even for lateral directions indicating a tendency to compress distance responses for all azimuthal directions. However, slope of the linear regression line fitting the log-stimulus to the log-response data with synthesized HRTFs (non-individualized and individualized) containing only auditory parallax cues was low, indicating a poor fit for all azimuth angles. This key observation suggests that in the presence of binaural cues, auditory parallax cues are not critical for accurate

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

distance perception.

An ANOVA test was carried out to further investigate the significance of auditory parallax and ILD cues for distance perception. The effect of auditory parallax cues on distance perception in the presence of other distance cues was first studied. The mean values of the responded distances with measured HRTFs ($HRTF_{measured}$), and with the modeled HRTFs containing only binaural ILD cues ($HRTF_{ILD}$), auditory parallax and ILD cues ($HRTF_{AP+ILD}$), were all similar for both non-individualized and individualized synthesis cases. This was evident from the ANOVA analysis, which returned no significant differences in distance responses with $HRTF_{measured}$, $HRTF_{ILD}$, and $HRTF_{AP+ILD}$ for non-individualized [$F(2, 25) = 0.55, p = 0.58$] and individualized cases [$F(2, 24) = 0.5, p = 0.55$], respectively. Note that F is the standard F-distribution. The p value refers to the probability of significance, which is compared to the threshold critical value to determine the significance of the factor. Throughout this paper, the significance of a factor is determined for critical significance levels (α) of 0.01 and 0.05 corresponding to 99% and 95% confidence intervals, respectively. The auditory parallax cues did not have a significant effect on distance perception in the presence of ILD cues using both non-individualized [$F(1, 15) = 0.28, p = 0.6$] and individualized HRTFs [$F(1, 16) = 0.34, p = 0.56$]. However, the effect of ILD cues simulated using the DVF model had a significant effect in both the individualized, as well as the non-individualized [$p < 0.001$] synthesis cases.

All the above observations suggest that auditory parallax cues in the presence of ILD cues do not have any prominent effect in anechoic headphone playback conditions. The correlation coefficient values also indicate poor distance perception for all directions, when auditory parallax is the only available cue. Therefore, it can be concluded that auditory parallax cues can be neglected and it is sufficient to include ILD cues in the model for the distance-dependent individualized HRTFs. This also means that the angular position of

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

frontal projection headphone transducer with respect to the ear need not be physically changed to incorporate the parallax effects in the near-field HRTFs.

In the next section, the proposed technique to model the distance-dependent individualized HRTFs using frontal projection headphones is discussed. The model is subsequently validated with the help of objective and subjective analyses.

6.3 Modeling distance-dependent individualized HRTFs

The frontal projection headphones project sound directly onto the pinnae from the front, emulating a loudspeaker setup in the far-field. Individualized pinna spectral cues are automatically modeled at the eardrum during frontal projection playback as shown by Sunder *et al.* [144]. The frontal projection headphones used in this study were open-type with free-air equivalent coupling (FEC) characteristics, as shown in Fig. 6.2 [47, 144]. The role of frontal projection headphones in modeling the individualized pinna spectral cues has been explained in detail in Sunder *et al.* [144].

From the analysis carried out in Section 6.2, it is clear that ILD cues are critical for distance perception in anechoic conditions, and the effect of auditory parallax can be neglected. In the model described in Fig. 6.6(a), mono sound source is first filtered with ipsilateral and contralateral ear transfer functions derived from the spherical head model. The spherical head model simulates the distance-dependent near-field effects and head shadow effects in the contralateral ear well, but it does not simulate the pinna spectral cues. The frontal projection headphones, however, inherently model the highly idiosyncratic frontal ipsilateral pinna cues at the eardrum during sound playback. Thus, in effect, ipsilateral spherical head transfer function is filtered with the frontal projection headphone transfer function in order to obtain the frontal ipsilateral cues in

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

the HRTF. However, additional headphone equalization and low-frequency compensation filters are required to ensure accurate modeling of the HRTF. In this way, the frontal projection headphones can synthesize the frontal directional distance-dependent individualized HRTFs without the requirement of extensive measurements. To model the rear directional HRTFs, a rear directional filter is additionally required, as shown in Fig. 6.7(b). Each of the processing blocks (Figs. 6.6(a) and 6.6(b)) involved in the modeling process is explained below:

Distance-dependent head effects: In the near-field, the reflections, and scattering of the acoustic waves with the head have significant spectral effects. The spherical head model [172] is used to model these distance-dependent head effects. In frontal projection headphones playback, idiosyncratic frontal pinna cues are generated at both the contralateral ear and ipsilateral ear. However, this does not affect contralateral response much, because of the large attenuation at high frequencies due to the head shadow effect. Humanski and Butler (1988) studied in detail the contribution of near-ear and far-ear in sound localization in the sagittal plane. They found that listeners could localize well with just the ipsilateral ear open and contralateral ear occluded. As the sound source moves towards lateral positions, the energy at high frequencies (above 4 kHz) in the contralateral ear is attenuated due to head shadowing. As a result, the exact response at the shadowed ear becomes less critical [246]. Avendano *et al.* [246], thus, showed that the contralateral ear HRTF could be modeled by a simple transformation of the ipsilateral ear HRTF using the spherical head model. The spherical head model also simulates the low-frequency ILD variation, which is a very prominent cue for distance perception.

Rear directional HRTFs: The frontal projection headphones due to the frontal positioning of the emitters emulate only the ipsilateral frontal pinna spectral cues. Besides, the spherical head model is based on a perfect symmetrical sphere and

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.6: Approaches to model distance-dependent individualized frontal (a) and rear (b) HRTFs.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

thus, the spherical head model fails to simulate the rear directional spectral cues. Therefore, an additional rear directional filter that can modify the spectral properties of the frontal HRTF to model the rear directional HRTF is required (Fig. 6.7). The rear directional filter is based on Blauert's directional bands [11], which suggest that certain frequency bands are boosted or attenuated characterising frontal or rear directions. Rear directional filter is developed by computing the average spectral differences between the HRTFs in the frontal and its symmetrical direction in the rear, for all subjects selected from the CIPIC HRTF database [126]. The rear directional filter is computed as follows:

$$RDF(\theta, \varphi) = \frac{\sum_{i=1}^{n_subjects} [HRTF_{cipc_rear}(\theta, \varphi) - HRTF_{cipc_frontal}(180^\circ - \theta, \varphi)]}{N_s}, \quad (6.11)$$

where N_s is the number of subjects in the CIPIC database (43 subjects), θ is the rear azimuth angle, and φ is the elevation angle. All angles are measured with respect to the mid-sagittal axis. $HRTF_{cipc_rear}(\theta, \varphi)$ is the rear HRTF and $HRTF_{cipc_frontal}(180^\circ - \theta)$ corresponds to the symmetrical frontal HRTF, both of which are selected from CIPIC database. The frequency responses of the rear directional filters are shown in Fig. 6.7.

$$HRTF_{rear}(\theta, \varphi) = HRTF_{frontal}(180^\circ - \theta, \varphi) \cdot RDF(\theta, \varphi). \quad (6.12)$$

The rear directional HRTFs can be obtained by filtering the frontal modeled HRTFs with the rear directional filter, as shown in Eq. (6.13).

Type-2 equalization: Typically, a HPTF comprises of the headphones transducer response and the acoustic coupling between headphones and listener's ears. Conventional headphone equalization methods often aim for a flat target response, so that the spectrum of the HRTF used remains undistorted at the eardrum. The

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

Figure 6.7: Rear directional filter for all the azimuth angles for the ipsilateral ear. Black: rear directional filters of all the subjects of the CIPIC database. Grey (Bold): average rear directional filter of all the subjects.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

equalization of the frontal projection headphones plays a critical role in the modeling process. In case of using frontal projection headphones, it is desirable to preserve the highly individualized frontal pinna cues at the eardrum, which are created during playback. Thus, using conventional headphone equalization is not advisable; as such an equalization technique would equalize even the individualized pinna spectral features generated during playback. For the frontal projection headphones, only the distortion created by headphone transducer along with the resonant modes of the earcup has to be compensated. Thus, the individual pinna cues (pinnacues) due to frontal projection are preserved after equalization. This type of equalization is known as the Type-2 equalization. Type-2 equalization improves the performance of binaural audio playback using frontal projection headphones by reducing the front-back confusions and in-head localization [144]. Type-2 equalization filter (Fig. 6.8) is obtained by computing the inverse of the free-field response (FFR) of the frontal projection headphones [Eq. (6.15)]. The frontal projection headphone with the transfer function $HPTF_{frontal}$, when equalized using a Type-2 equalization filter, models the unique personal pinna cues (pinnacues), as shown in Eqs. (6.14) and (6.15). FFR is obtained by measuring the free-field response of the headphones in the absence of dummy head. FFR captures only the spectral effects of transducer and resonant modes of the earcup [144]. To measure the FFR of frontal projection headphones, a B & K 4961 Multi-Field 1/4" microphone was placed at the center of the earcup just outside its peripheral surface. The distance of the microphone from the transducer was 5 cm, which is approximately equal to the distance at which the ears would have been present if the headphones were in place. Type-2 equalization can be described in mathematical terms as follows:

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

Figure 6.8: Type-2 equalization filter calculated by taking the inverse of an average FFR based on 10 trials.

$$HPTF_{frontal}(\omega) = FFR(\omega) \cdot PC(\omega), \quad (6.13)$$

$$Type2EQ(\omega) = \frac{1}{FFR(\omega)}, \quad (6.14)$$

where $PC(\omega)$ corresponds to the individualized pinna cues due to frontal projection. $HPTF_{frontal}(\omega)$ and $FFR(\omega)$ are the frontal projection headphone response and its free-field response, respectively. $Type2EQ(\omega)$ refers to the Type-2 equalization filter.

Low-frequency compensation filter: The low-frequency spectral cues and ILD cues are key for distance perception and have to be modeled correctly. The frontal projection headphones have a poor low-frequency response due to their open structure.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

The inability of the Type-2 equalization filter to completely correct the poor low-frequency response results in a noticeable deviation in the modeled HRTFs from the measured HRTFs. Deviation in the low-frequency response can have adverse effects on distance perception [29]. Thus, the low-frequency deviation has to be corrected in order to ensure that the auditory distance perception with modeled HRTFs is close to that with the true measured HRTFs [29]. The low-frequency response of the HRTF is non-idiosyncratic and is mainly due to head, shoulder, and torso effects. These distance-dependent low-frequency effects can be approximated using the spherical head model, as shown in Figs. 6.6(a) and 6.6(b). To begin with, the low-frequency deviation from the true measured HRTF has to be estimated. Low-frequency compensation filter is obtained by taking spectral inverse of the low-frequency response (up to 1 kHz) of the Type-2 equalized frontal projection headphone, which is measured on the B & K HATS dummy head. This filter model is valid as both dummy heads and human subjects display similar spectral properties in the HRTF in the low-frequency range. Note that the compensation filter is designed to have a flat response beyond 1 kHz, as shown in Fig. 6.9, to avoid any effect on the modeled HRTF spectrum at high frequencies.

Interaural time differences: HRTF is often approximated as a minimum-phase function with a position-dependent, frequency-independent time delay [45]. In this model, lateral azimuth perception is simulated by adding a frequency-independent ITD to the minimum-phase representation of the modeled HRTF. Wightman and Kistler [94] carried out localization experiments using minimum-phase description of HRTFs and reported that the sound localization with these models using headphone simulation was close to that obtained in free-field conditions. Kulkarni *et al.* [181] also validated the minimum-phase representation of HRTFs with the help of psychophysical experiments and modeled the ITD as frequency-independent time

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.9: Low-frequency compensation filter to compensate for the deviation in the low-frequency response of the modeled HRTF from the measured individualized HRTF.

delay. The ITD used in this study is obtained by taking the average ITD of all subjects selected from the CIPIC database for a particular location. The effect of using a non-individualized ITD is not as critical as the use of non-individualized pinna cues. This is evident from the fact that the azimuth localization performance with non-individualized HRTFs is comparable to that with the individualized HRTFs [106]. Moreover, ITD is primarily a low-frequency cue and is not affected much by the variations of individual’s pinna, which contributes mainly to the high-frequency cues. Thus, the use of an average ITD is a valid approximation in this model. As the variation of ITD values is minimal with distance, we consider a constant ITD value across all distances in this model. The variation in the

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

interaural cues with distance can also be measured in terms of their respective just noticeable differences (jnd). Hershkowitz and Durlach [238] found that listeners could discriminate changes in ILD on the order of 0.8 dB over a wide range of ILDs. Thus, the distance-dependent changes in ILD from 0.12 m to 1 m span a range of upto 15 jnd's at 500 Hz and 30 or more jnd's at 3 kHz. However, the jnd for ITD was approximately $15\mu s$ at ITDs below $400\mu s$ and increased rapidly for ITDs greater than $400\mu s$. The variation in ITD in the near-field spans only 2-3 jnds. Researchers have speculated that listeners may make use of distance-invariant ITDs to first determine the azimuth location and then use the ILD cues to estimate the distance [43, 233].

Figure. 6.10 shows a series of spectra that illustrate a step-by-step process for modeling the ipsilateral frontal, contralateral frontal, and ipsilateral rear HRTFs. The model can be mathematically described as follows:

$$Y_{frontal}(\omega) = X(\omega) \cdot HPTF_{frontal}(\omega) \cdot Type2EQ(\omega) \cdot SHM(\omega) \cdot LfCF(\omega), \quad (6.15)$$

$$Y_{frontal}(\omega) = X(\omega) \cdot PC(\omega) \cdot SHM(\omega) \cdot LfCF(\omega), \quad (6.16)$$

$$Y_{rear}(\omega) = X(\omega) \cdot PC(\omega) \cdot SHM(\omega) \cdot RDF(\omega) \cdot LfCF(\omega), \quad (6.17)$$

where ω is angular frequency in radians. $X(\omega)$, $Y_{frontal}(\omega)$, and $Y_{rear}(\omega)$ are the input mono signal, and the binaural signals synthesized for frontal and rear directions, respectively. The Type-2 equalization filter is denoted by $Type2EQ(\omega)$. The spherical head model that models the distance-dependent head effects is denoted by $SHM(\omega)$. The rear-directional filter and low-frequency compensation filter are represented by $RDF(\omega)$ and $LfCF(\omega)$, respectively. $PC(\omega)$ denotes the individualized pinna cues.

For ease of explanation, we start with the frontal projection headphone response $[HPTF_{frontal}(\omega)]$ that emulates the idiosyncratic frontal ipsilateral cues due to frontal

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

Figure 6.10: Different steps involved in the modeling process. (a) Subject K, source is placed at 40° azimuth angle and 70 cm from the center of the head (b) Subject A, source is placed at 60° azimuth, 50 cm from the center of the head (c) Subject R, source is placed at 100° azimuth, 60 cm from the center of the head. The azimuth angle is measured with respect to the sagittal plane.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

projection. The response $HPTF_{frontal}(\omega)$ apart from the pinna cues also contains the inherent spectral distortion of transducer and the resonant modes of the earcup that are compensated by the Type-2 equalization filter. A spherical head model filter [$SHM(\omega)$] is then applied to model the distance-dependent spectral effects of head in the ipsilateral ear [Fig. 6.10 (a)]. The spherical head model also simulates the head shadow effects in the contralateral ear [Fig. 6.10 (b)]. To compensate for the deviation at low frequencies, a low-frequency compensation filter [$LfCF(\omega)$] is applied to obtain the correct low-frequency response of the HRTF. However, to model the rear directional HRTFs, an additional rear directional filter [$RDF(\omega)$] is required that modifies the frontal HRTF response to a HRTF at a symmetrical rear position [Fig. 6.10 (c)]. The accuracy of the modeled and measured HRTFs are compared and analysed in the next section.

6.3.1 Objective analysis

To validate the proposed model, the modeled distance-dependent individualized HRTFs were compared with the measured distance-dependent individualized HRTFs for all the three subjects, whose HRTFs were measured. The modeled individualized frontal ipsilateral, frontal contralateral, and rear ipsilateral HRTFs are shown in Figs. 6.11, 6.12, and 6.13, respectively along with the measured individualized HRTFs. It should be noted that the modeled HRTFs using the frontal projection headphones exhibit a good match with measured HRTFs. The high-frequency individualized pinna spectral cues modeled by frontal projection headphones match accurately with that of the actual frontal pinna spectral cues of the listener (Fig. 6.11). The spherical head model simulates the distance-dependent low-frequency features and the head shadow effects accurately. The head shadow effects modeled by the spherical head model is clearly visible in the modeled contralateral ear HRTFs at high frequencies (Fig. 6.12). For the rear ipsilateral HRTFs, it is observed that the spectral pattern of the modeled HRTFs is similar to that of the

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.11: Comparison of modeled individualized frontal ipsilateral HRTFs and measured individualized frontal HRTFs of all the three subjects for different azimuthal angles and distances, as indicated in each panel.

measured individualized HRTFs (Fig. 6.13).

RSD between modeled and measured HRTFs was calculated as an objective measure for comparison (Eq. 6.9). Average RSD scores in dB for the modeled ipsilateral and contralateral ear HRTFs at all source distances and across all the subjects are plotted as a function of azimuth in Fig. 6.14. For the ipsilateral ear, RSD values were less than 4 dB in the frontal directions and increased for directions behind the head. The average RSD score for the ipsilateral HRTFs across all azimuth and all distances was found to be less than 4.5 dB (Fig. 6.14). The RSD errors for the contralateral ear (5 dB) were always greater than that for the ipsilateral ear HRTF for both frontal and rear directions

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

Figure 6.12: Comparison of modeled individualized frontal contralateral HRTFs and measured individualized HRTFs of all the three subjects for different azimuthal angles and distances, as indicated in each panel.

(Fig. 6.14). The high-frequency mismatch in the modeled contralateral ear HRTFs with the measured individualized HRTFs is due to the model generating ipsilateral pinna cues even in the contralateral ear HRTFs. The RSD scores for the modeled rear directional ipsilateral HRTFs (5 dB) were higher than that for the modeled frontal ipsilateral HRTFs (3 dB), as shown in Fig. 6.14. For the contralateral HRTFs in the rear directions, the RSD scores were again higher than that for the frontal directions due to the generation of ipsilateral cues during playback. The higher RSD values in the rear directions is

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.13: Comparison of modeled individualized rear ipsilateral HRTFs and measured individualized HRTFs of all the three subjects for different azimuthal angles and distances, as indicated in each panel.

due to the fact that the rear directional filter is only an approximate method to model the rear directional HRTFs. The modeled HRTFs were also validated using perceptual experiments. These experiments are discussed in the next section.

6.3.2 Subjective experiments

Detailed subjective experiments were carried out to validate the proposed model. Subjects judged the perceived direction and distance of the virtual source synthesized using the modeled distance-dependent individualized HRTFs that was played back from the frontal projection headphones. The measured distance-dependent individualized HRTFs

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.14: Root mean spectral difference (RSD)scores of the modeled HRTF with respect to the measured individualized HRTFs for the ipsilateral and contralateral ear.

were reproduced over an equalized (conventional individualized equalization) Sennheiser HD600 headphone. Four tests were carried out to validate the proposed model:

1. Localization experiment using modeled and measured distance-dependent individualized HRTFs.
2. Distance perception experiment using modeled distance-dependent individualized HRTFs.
3. Distance perception experiment using modeled distance-dependent individualized HRTFs with intensity scaling.
4. Distance perception experiment using real sound sources.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

A total of 25 subjects participated in experiments 1 to 3, and 5 subjects participated in experiment 4. All subjects had normal hearing. The basic stimulus used was a sequence of four bursts of cosine square-gated (30-ms) WGN of 300-ms each, separated by an interval of 30-ms of silence, as in the experiments in Section 6.2.1. The stimulus for experiments 1, 2, and 4 were normalized using a scaling factor N , where N is obtained from Eq.6.9. The scaling factor mainly compensates for the intensity cues that arise due to variation in distance between the source and head. In the intensity scaling condition, additional intensity cues were added to the un-normalized modeled HRTF to aid distance perception. A scaling factor of $(1/N)$ was used to account for the fact that the difference in distance between the sound source and the two ears can be substantially different, when the source is in the near-field and close to the interaural axis [244]. For the distance perception experiment with real sources, a Cambridge Soundworks mini cube speaker was used to playback the stimulus. Distance perception was tested for all the eight speaker distances for which the HRTFs were measured in Chapter 3. Apart from the loudspeaker from which the stimulus was played, additional dummy loudspeakers were placed at random positions in order to avoid visual bias.

All the subjects were blindfolded for the experiment, which was conducted with real sound sources. All the experiments were carried out in the semi-anechoic chamber that was used for the experiments described in Section 6.2.1. The highest stimulus level was approximately 60 dBA at 1 m distance. Subjects gave their responses for the perceived direction and distances on a GUI, similar to the one used in experiments in Section 6.2.1.

The analysis of results for directional localization was carried out by dividing the distances into three different categories as SET 1: near (20 cm, 30 cm, 40 cm), SET 2: medium (50 cm, 60 cm, 70 cm), and SET 3: far (80 cm, 90 cm, 140 cm). The front-back confusions in the localization responses with the modeled individualized HRTFs, measured individualized HRTFs, and non-individualized HRTFs were analysed for all

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.15: Localization responses for modeled individualized, measured individualized, and measured non-individualized HRTFs for all the three sets of distances. The distances are divided into three categories: a) near (20 cm, 30 cm, 40 cm), b) medium (50 cm, 60 cm, 70 cm), and c) far (80 cm, 90 cm, 140 cm). Modeled distance dependent individualized HRTFs were played back using frontal projection headphones, while the measured individualized and non-individualized HRTFs were played back using Sennheiser HD600 headphone.

the distance sets (Fig. 6.15).

An ANOVA analysis for front-back confusions revealed that the mean differences in front-back confusions between modeled and measured HRTFs were statistically significant for near (SET 1) [$F(1, 32) = 26, p \ll 0.01$] (Fig. 6.15a) and medium distances (SET 2) [$F(1, 32) = 7, p \ll 0.01$] (Fig. 6.15b). In the far-field (SET 3), front-back confusions in the localization responses with modeled and measured individualized HRTFs were similar

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

with no significant differences [$F(1, 32) = 7, p = 0.03$] for a critical significance level of 0.01 (Fig. 6.15c). However, their differences were statistically significant, if a critical significance level of 0.05 was considered. Overall, front-back confusions for localization with modeled individualized HRTFs were reduced by more than 50% as compared to those with non-individualized HRTFs. In fact, the front-back confusions with modeled individualized HRTFs tended to be closer to the measured individualized HRTFs (Fig. 6.15).

Figure. 6.15 shows the localization responses for modeled individualized HRTFs and measured (non-individualized and individualized) HRTFs. In this study, localization error is defined as the angular difference between the perceived angle reported by the subject and the actual azimuth angle of the source. Front-back confusions were corrected prior to the calculation of the localization error. An ANOVA test was carried out to investigate the directional localization performance using modeled and measured HRTFs. Comparing the modeled and measured individualized HRTFs, the localization errors (with front-back confusions removed) were not significantly different across all the three sets for a critical significance level of 0.01; SET 1 [$F(1, 26) = 0.39, p = 0.53$], SET 2 [$F(1, 26) = 0, p = 0.9644$], SET 3 [$F(1, 26) = 5.91, p = 0.02$]. The localization error differences between the modeled and measured individualized HRTFs in the far field (SET 3) were found to be significant for a higher critical significance level of 0.05.

The distance responses for modeled individualized HRTFs (with and without intensity scaling), and for the real sound sources are plotted in Fig. 6.16. The distance responses for the measured individualized HRTFs are shown in Fig. 6.4. The correlation coefficient values between the log-stimulus and log-response distance and the slopes of the linear regression line fitting the log-stimulus to the log-response data for the modeled individualized HRTFs without and with intensity cues are plotted across the azimuth in Fig. 6.16a and Fig. 6.16b, respectively. The correlation coefficient values for the distance

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF
USING FRONTAL PROJECTION HEADPHONES

Figure 6.16: Stimulus-Response pairs for distance (in cm) using modeled individual HRTFs (a) without and b) with intensity cues) and c) real sound sources (without intensity cues). The layout is as in Fig. 6.3.

responses using modeled HRTFs without intensity cues (0.68 lateral, 0.35 near-medial), were close to that using the measured individualized HRTFs, for both lateral (0.72) and near-medial (0.46) directions (Fig.6.4). When the modeled individualized HRTFs were

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

Figure 6.17: Correlation and Slope of the subject's responses for the modeled HRTFs without and with intensity cues

intensity scaled and played back, the mean correlation coefficient values increased to 0.83 and 0.54 in the lateral and near-medial directions, respectively. This result suggests that with intensity cues added, the distance perception could be improved for the modeled individualized HRTFs.

With real sources, the correlation coefficient values between log-stimulus distance and log-response distance for lateral and near-medial directions were found to be 0.76 and 0.50, respectively (Fig.6.16c). It was noted that the distance localization accuracy with the modeled HRTFs (without intensity cues) was close to that with real sound sources for lateral directions. However, for near-medial directions, localization with real sources was much better. One of the probable reasons for better performance with real sources could be the presence of some reverberant cues, as these experiments were carried out in the semi-anechoic chamber [56, 247–250]. Other researchers have also reported that the distance perception with real sources, is usually better than that in virtual auditory display [251].

The slope was less than one for both lateral and near-medial directions with and without the presence of intensity cues, indicating a general tendency to compress the dis-

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

tance responses for all azimuthal directions. The correlation coefficient values between actual and responded distance was poor for positions near the medial plane, and it improved for more lateral angles in both the cases, as shown in Fig. 6.17. An ANOVA test was carried out to compare the distance perception between the modeled individualized, measured individualized, and real sources. Critical significance levels of 0.01 and 0.05 were considered. For a critical significance level of 0.01, the distance perception with the modeled and measured individualized HRTFs were found to be similar with no significant differences [$F(1, 8) = 7.51, p = 0.02$]. However, these differences were significant for a higher critical significance level of 0.05. The distance responses with real sources for lateral directions were compared to that of modeled HRTFs. ANOVA results returned their differences [$F(1, 8) = 9, p = 0.015$] to be statistically not significant for a critical significance level of 0.01 and significant for a significance level of 0.05. In the median plane, significant differences were found between the distance responses with modeled and measured individualized HRTFs [$F(1, 8) = 16, p < 0.01$]. The effect of intensity scaling was found to be significant [$F(1, 8) = 11.36, p = 0.009$] and the distance perception improved for both lateral and near-medial locations in the presence of intensity cues.

6.4 Summary

In this chapter, a technique has been introduced to model distance-dependent individualized HRTFs in the horizontal plane. The main results are as follows:

- Auditory parallax cues at high frequencies displayed negligible perceptual effects for distance perception in the presence of ILD cues. Hence, the auditory parallax cues were not incorporated in the proposed model. The low-frequency ILD cues were found to be a prominent distance cue in anechoic headphone listening conditions.

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

- Distance-dependent individualized HRTFs in the horizontal plane were modeled using frontal projection headphones along with a cascade of signal processing blocks. Frontal projection of sound inherently modeled the frontal ipsilateral pinna cues at the eardrum spectra during sound playback. The head shadow effects in the contralateral ear were modeled by using the spherical head model. A rear-directional filter was developed to model rear-directional HRTFs from frontal HRTFs. In addition, an equalization filter (Type-2 equalization) and a low-frequency compensation filter were required to accurately model the distance-dependent individualized HRTFs.
- The proposed individualization technique modeled the frontal ipsilateral HRTFs well. Modeling errors were larger for the contralateral ear, as the ipsilateral pinna cues are generated in the modeled contralateral HRTF due to frontal projection. For the rear directions, the modeling errors were larger than in the frontal directions. This is because the rear-directional filter used to model the rear directional HRTFs is a generic filter, which is based on the average front-back spectral differences in the CIPIC database.
- Directional localization performances with the modeled and measured individualized HRTFs were similar with no significant differences for all the three sets of distances (near, medium, and far) for a significance level of 0.01. However, their differences were statistically significant for a significance level of 0.05. Larger errors in the modeled contralateral responses due to frontal projection did not have any perceptual effect in terms of localization and distance perception. Moreover, subjects could perceive the rear directions well, albeit with slight increase in error.
- Distance judgements for the modeled and measured individualized HRTFs were similar with no statistical differences for both lateral and near-medial directions for

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

a significance level of 0.01. However, these differences were statistically significant for a significance level of 0.05. Presence of intensity cues had a significant effect and improved the distance perception for both lateral and near-medial directions.

- In comparison with the real sources, distance judgements in lateral directions for both modeled and real sources were similar with no significant differences for a significance level of 0.01. However, their differences were statistically significant for a significance level of 0.05. For near-medial directions, distance localization with real sound sources were more accurate than for virtual sounds synthesized with modeled distance-dependent HRTFs. The improved performance with real sound sources was probably due to reverberant cues present in the semi-anechoic chamber and absent in our model.

The frontal projection headphones used in this study are of open-type with FEC characteristics. FEC characteristics are vital particularly when the HPTF measurements are carried out at the blocked ear canal. Use of FEC headphones is critical in making auralization realistic and to ensure that sound pressure levels with and without headphones are similar [47]. Since the position of the transducer in frontal projection headphone does not vary with changes in virtual source azimuth, the pinna spectral cues introduced due to frontal projection also remain the same with changes in azimuth angle. This assumption is valid, since the pinna-based spectral notch positions do not vary much with changes in azimuthal position in the horizontal plane [122]. The proposed method made use of a distance-invariant, frequency-independent, azimuth-dependent ITD value, taken as the average of ITDs of all subjects from the CIPIC database, to simulate lateral azimuthal perception in the horizontal plane. Other techniques can further be used to individualize the ITD, to achieve a more accurate azimuthal perception [126].

To conclude, a technique has been introduced in this work, to model distance-dependent individualized HRTFs in the horizontal plane. In the next chapter, this work is extended

CHAPTER 6. MODELING DISTANCE-DEPENDENT INDIVIDUALIZED HRTF USING FRONTAL PROJECTION HEADPHONES

to model the HRTFs at any elevation by manipulating the frontal notches created by frontal projection, which is an essential cue for elevation.

Chapter 7

Individualization of head-related transfer functions in the median plane using frontal projection headphones

The frontal projection headphones can be used to model the distance-dependent individualized HRTFs in the horizontal plane, as explained in Chapter 6. This approach has been extended to model the individualized HRTFs in the sagittal plane in this chapter. Frontal projection creates unique pinna spectral notches that are critical for accurate elevation perception. The spectral notches in the HRTFs vary monotonically in their center frequencies as the source elevation angle increases. In order to model the HRTFs in the median plane, the positions of the idiosyncratic pinna notches of the listener are first identified from a one-time measurement of the frontal projection headphones response at the listener's ears. The variation of the spectral notches with elevation is studied for all the subjects in the CIPIC database [70]. Using the information of variation of

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

notch frequencies with elevation, the spectral notches in the measured frontal projection response is shifted according to the elevation angle. The head diffraction and the torso effects provide important low-frequency cues for distance perception [126]. The head effects are modeled by the spherical head model [172]. The non-idiosyncratic torso effects are derived by extracting the contribution of the torso from the HRIRs of the CIPIC database. Furthermore, the rear-directional filters (Section 6.3) are used to model the rear-directional filters. In addition, Type-2 equalization and the low-frequency compensation filter are required to completely model the individualized head-related transfer functions at different elevations in the median plane. These results can be extended to any horizontal plane by using a azimuth-dependent frequency-independent ITD.

7.1 Spectral cues in the sagittal plane

The pinna spectral cues play a prominent role in the localization of elevation sources [9, 45, 99, 105, 129, 132]. Hebrank *et al.* [9] studied the spectral cues in the median plane and found that an increase in frontal elevation is signaled by an increase in the lower cut-off frequency of the 1-octave notch. Butler *et al.* [105] also found that a prominent notch in the HRTF curve moved towards the lower frequencies, as the source position was moved from above to below in the frontal half of median plane.

7.2 Models for simulating elevation effects

Chun *et al.* [252] used spectral notch filtering and directional band boosting to create a virtual source elevation for stereo loudspeaker reproduction. Bloom *et al.* [110] manipulated the spectral features of the external ear to create an illusion of a continuous variable source elevation. Bloom *et al.* [110] also measured the monaural hearing sensitivities for seven source elevations in the vertical plane to investigate whether listeners are sensitive

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

to spectral changes due to changes in elevation. They found that the monaural threshold curves exhibited definite threshold maxima or sensitivity minima, whose frequencies increased with source elevation. Martens [192] developed a technique to customize the source elevation perception with a rapid psychophysical calibration using bisection scaling. In this technique, the subjects first produce five settings indicating the point at which the perceived elevation of a virtual source matches their own internal standard for “ear-level” incidence. The median of these settings are then used to provide an anchoring stimulus to create an individualized psychophysical scale for controlling source elevation as perceived by the listener. Kendall and Rodgers [253] modeled the HRTFs using four different features (a peak, two notches and attenuation at high-frequencies) with a simple two-pole, four-zero digital filter. Avendano and Algazi [254] modeled the low-frequency elevation-dependent features in the HRTF spectrum that are due to torso, shoulder reflections, and head diffraction effects. Subjective experiments showed that the model produced significant elevation cues for virtual sound sources, whose spectra are limited to frequencies below 3 kHz. It was found that the low-frequency binaural elevation cues are perceptually significant away from the median plane, complementing the high-frequency monaural pinna cues. Algazi *et al.* [255] presented a simple geometrical model of the head and the torso that provide insight into the low-frequency elevation effects. The head and torso models were obtained by adding both spherical and ellipsoidal models of the torso to a classical spherical-head model. It was found that both torso reflections and torso shadow provide significant elevation cues. These geometric models are a good approximation of the HRTF for the KEMAR manikin with its pinnae removed. Gan *et al.* [256] proposed an algorithm for simulating elevation perception for digital home entertainment system by partitioning the system into low-frequency and high-frequency regions. At the high frequencies, the elevation perception is simulated by incorporating the elevation-dependent spectral notches of the HRTF. Shinn *et al.* [196]

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

developed a customisation technique to enhance the vertical elevation perception based on response tuning in the median plane. The pinna responses of 45 individual HRIRs in the CIPIC HRTF database were isolated and modeled as a linear combination of 4 or 5 basis functions for every elevation angle using PCA. Listeners tuned the weights of the each basis function to customize the HRTFs. Hwang *et al.* [257] developed a similar technique, where the subjects tuned the weights of 3 dominant basis functions corresponding to the 3 largest standard deviations at each elevation. Guldenschuh *et al.* [258] simulated elevation perception by investigating the elevation-dependent torso-related impulse responses (TRIR). Measured HRIRs for 10 different horizontal head-torso angles were averaged to obtain a basis HRIR that is independent of the torso position. In the study conducted by Guldenschuh *et al.*, TRIR was obtained by taking the difference between torso dependent HRIR and averaged basis HRIR. TRIRs revealed high energy for certain geometries between sound source and shoulders. An analytic system was presented in [258], which models the TRIR by linear regression, and the basis HRIRs by spherical harmonics.

Iida *et al.* [259] carried out several experiments to identify the peaks and notches in the HRTFs that play a role as important spectral cues in the median plane. They labelled the peaks and notches as $P1$, $P2$, ... and $N1$, $N2$ and so on, according to the increasing order of frequencies (Fig. 7.1). They first extracted the peaks and notches from the measured listener's HRTFs regarding the peak $P1$ around 4 kHz as the low-frequency reference. The peak $P1$ is found in general, to be constant with change in elevation [116]. Localization experiments were conducted using a parametric model of an HRTF that is composed of the extracted peaks and notches. It was found that the parametric HRTF composed of the first two notches $N1$, $N2$ and the first peak $P1$ provided almost the same localization similar with measured HRTFs. In another related work, Iida *et al.* [260] estimated the $N1$, $N2$ and $P1$ notches of the HRTF and the best matching HRTFs are

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.1: The spectral notches and peaks formed due to the interaction of the acoustic wave with the pinna. The peaks and notches are labelled as $P1$, $P2$, ... and $N1$, $N2$ and so on, according to the increasing order of frequencies.

selected from the database for which the estimated notches $N1$, $N2$ were closest to. They found that the best-matched HRTFs provided approximately the same performance with the listener's own HRTFs for both frontal and rear directions. Iida *et al.* [260] obtained the position of the notches and peaks from HRTFs of 54 ears. The peak $P1$ is found to be mainly distributed in the low frequencies (3.4 kHz to 4.3 kHz), while $N1$ and $N2$ notches were distributed at the higher frequencies from 5.7 kHz to 9.5 kHz and 8.2 kHz to 13.5 kHz, respectively. This observation suggests that the individual variation in $P1$ is much smaller than $N1$ and $N2$. Nishioka *et al.* [261] measured the jnds of $P1$ frequency for vertical localization with the help of localization experiments. The $P1$ frequencies in

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

the HRTFs (recomposed from $P1$, $N1$, $N2$) were shifted by 0.05, 0.1, 0.2, and 0.5 octaves compared with the reference HRTFs. The subjects were asked to indicate whether the vertical angles perceived for both the stimuli were the same. The jnds for $P1$ for the lower and upper sides were found to be 0.47 and 0.35 octaves, respectively. The jnds of $N1$ and $N2$ were found to be in the range of 0.1 to 0.2 octaves. Comparing the jnds with the individual variations of $P1$, $N1$, and $N2$, the individual difference in $P1$ can be considered to have negligible effect on vertical localization. Since, the jnd of $N1$ and $N2$ are much lesser than their individual variation; the individual differences in the notches can have a large impact on the localization. Thus, to model the individualized HRTFs in the median plane, the positions of the notches ($N1$, $N2$) must be known.

The frontal projection headphone projects the sound from the front and automatically creates the unique pinna spectral features that include the $P1$, $N1$, $N2$ notches. The pinna notch positions are first extracted from the frontal projection headphone response measured on the subjects. The spectral notch position variation with elevation is analysed in detail from the HRTFs in the CIPIC database. Using this information, the notch positions extracted from the measured frontal projection headphone response is shifted according to the elevation angle. The spherical head model is then applied to model the spectral effects due to the head and the head shadow effect in the contralateral ear. In addition, the low-frequency compensation filter and the Type-2 equalization filters are required to model the HRTFs in the vertical plane correctly (Chapter 6). To model the HRTFs in the rear directions, the modeled frontal directional HRTFs are filtered with a rear-directional filter that is obtained from the spectral differences between the frontal and rear directions of a CIPIC database, as explained in Chapter 6.

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER
FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION
HEADPHONES

7.3 Measurements of HRTFs and HPTFs

HRTFs were measured for three human subjects and B & K HATS dummy head for a total of 12 elevation angles. The HRTFs were measured for elevation angles ranging from -40° below the head to 80° above the head at a resolution of 10° . Apart from the median plane measurements, HRTFs for 11 azimuthal angles at a resolution of 30° were measured for each elevation. The experimental procedures for the measurement of HRTFs have already been explained in Chapter 3.

In addition, frontal headphone responses were measured for 15 subjects using the B & K 4101 microphones at the blocked ear-canal entrance to capture the pinna spectral features. Figure 7.2 shows the frontal projection headphone responses measured on human subjects for five trials of measurements. The following features in the frontal projection headphone response are immediately noticeable:

1. The distinct pinna spectral notches ($N1$, $N2$) are clearly visible for all the subjects at the high frequencies.
2. The notch positions are highly idiosyncratic for different individuals.
3. The poor low-frequency response of the frontal projection headphone due to its open structure.

The free-field response (FFR) of the frontal projection headphone was also measured to obtain the Type-2 equalization filter (Chapter 4).

7.4 Estimation of the position of the notches $N1$, $N2$

The positions of the notches ($N1$, $N2$) due to pinna in the measured HRTF and the frontal projection headphone response have to be identified. The spectral features in the

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.2: Frontal projection headphone responses measured on human subjects. Black: Average frontal projection headphone response of the five trials of measurements. Blue: Frontal projection headphone response based on five different trials. Frontal projection headphone responses were measured using a B & K 4101 binaural microphone placed at the blocked ear canal of the human subject. The notches at the high frequencies are due to the interaction of the frontal projected sound with the pinna. Also note that the low-frequency response of the frontal projection headphone response is poor due to its open structure.

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

HRTF can be due to the pinna, head, torso, shoulder, and knees. Thus, the spectral features due to the torso, head, shoulder, and knees have to be removed during the extraction of the pinna cues. Some of the common methods to extract the spectral notches and peaks is through a pole-zero model [121, 262–265]. In these models, to approximate the spectrum envelope better, higher orders (> 30) of the model are required. Moreover, it is difficult to isolate the notches in the HRTF spectrum due to the pinna alone with the pole-zero or all-pole models.

Raykar *et al.* [8] proposed a method to robustly extract the deep spectral notches due to the pinna. Raykar’s method utilizes several signal-processing techniques that include windowing, linear prediction residual analysis, group delay function, and autocorrelation. In this chapter, we use Raykar’s method to extract the positions of the pinna notches from the measured HRTFs, as well as the frontal projection headphone response. Figure 7.3 shows a step-by-step process to extract the pinna notch frequencies using Raykar’s algorithm. The spectral notches in the HRTF are due to the interaction of the acoustic wave with the head, shoulder, torso, knees, and pinna cavities. Therefore, it is necessary to isolate the spectral effects due to the pinna cavities. The direct component of the HRIR signal can be isolated from the reflected components using windowing in the time domain. The torso reflections and the knee reflections can be removed by using a Hann window of size 1 ms. This is because the torso and the knee reflections occur at around 1.6 ms and 3.2 ms, respectively in the HRIR. Windowing the waveform reduces the spectral effects of reflections. However, the frontal projection response consists of only the pinna spectral cues and do not contain the effect due to head, shoulder, torso, knees. Thus, windowing the frontal projection impulse response is not specifically required.

The main advantage of using the group delay function in Raykar’s method [8] lies in its additive nature of phase spectra and high-frequency resolution property. The high-frequency resolution property of the group delay function helps in providing a better

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.3: Illustration of the pinna extraction algorithm developed by Raykar *et al* (2006) for the CIPIC HRIR of Subject 127. (a) : Signal, (b) : LP residual, (c) Windowed LP residual, (d) Autocorrelation, (e) Windowed autocorrelation. (f), (g), (h), (i), (j) are their corresponding magnitude spectrum. (k) corresponds to the group delay spectrum of the windowed autocorrelation. Note that the pinna spectral notches are prominent in the group delay spectrum due to its high-frequency resolution property.

resolution of peaks and valleys for a short time segment of the data [266,267]. The group delay function ($\tau(\omega)$) is computed as the negative of the derivative of the phase spectrum

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER
FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION
HEADPHONES

of a signal with reference to frequency:

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega}, \quad (7.1)$$

where $\tau(\omega)$ is the group delay function, $\theta(\omega)$ is the angular frequency, and is the phase spectrum of the frequency response of the signal. The group delay can also be obtained from the Fourier transform of $x(n)$ and $nx(n)$.

$$X(\omega) = \sum_{n=0}^{N-1} x(n)e^{-j\omega n} = X_R(\omega) + jX_I(\omega), \quad (7.2)$$

$$Y(\omega) = \sum_{n=0}^{N-1} nx(n)e^{-j\omega n} = Y_R(\omega) + jY_I(\omega). \quad (7.3)$$

Since

$$\log X(\omega) = \log |X(\omega)| + j\theta(\omega), \quad (7.4)$$

$$\tau(\omega) = -\frac{d}{d\omega} [\theta(\omega)] = -\text{Im} \left(\frac{d}{d\omega} [\log X(\omega)] \right), \quad (7.5)$$

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + Y_R^2(\omega)}.$$

where $\text{Im}(z)$ corresponds to the imaginary part of the complex number z .

Windowing reduces the resolution in frequency domain and introduces artefacts. The artefacts due to windowing can be reduced by removing the interdependence among adjacent signal samples. The interdependence is achieved by using the linear prediction (LP) residual of the original signal and then windowing the residual. Windowing the LP residual essentially corresponds to the removal of the resonances from the signal. In this study, a 12th order LP analysis is used to derive the residual [268]. In LP analysis, the signal is approximately predicted from the linearly weighted summation of the past

**CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER
FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION
HEADPHONES**

p samples:

$$\hat{x}(n) = - \sum_{k=1}^p a_k x(n-k). \quad (7.6)$$

The LP residual is obtained by calculating the error in prediction, which is given by:

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^p a_k x(n-k). \quad (7.7)$$

The total squared error is obtained as:

$$E = \sum_n e(n)^2 = \sum_n \left[x(n) + \sum_{k=1}^p a_k x(n-k) \right]^2. \quad (7.8)$$

The normal equations are obtained by minimizing the mean squared error with respect to the coefficients $\{a_k\}$ [268]. The normal equations can be solved to obtain the coefficients $\{a_k\}$.

$$\sum_{k=1}^p a_k R(n-k) = -R(k), k = 0, 1, \dots, p, \quad (7.9)$$

where $R(k) = \sum_n x(n)x(n-k)$ is called the autocorrelation function for a lag of k samples. The minimum total error or the LP residual can be obtained by substituting the coefficients $\{a_k\}$ corresponding to the minimum error into the equation of the total squared error.

The samples of the past are linearly weighted to predict the current sampling instant. The LP analysis can be understood as a technique that eliminates the redundancy in the signal samples by removing the predictable part from the signal. The LP residual looks like noise, as the correlation among samples is significantly reduced as compared to the original signal. The autocorrelation function of the LP residual is an impulse at the origin with smaller amplitudes at other lags. The direct windowing of the LP residual as compared to the direct windowing of the original signal results in the more prominent appearance of the spectral notches as shown in Figure. 7.3.

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

The effects due to truncation and noise can be reduced by first calculating the autocorrelation of the LP residual. The characteristic feature of the autocorrelation function is that it produces decreasing amplitudes away from its peak. Computing the group delay of this signal enhances the resolution of the spectral components, while preserving the details of the spectral envelope. Thus, the group delay function of the autocorrelation function of the windowed LP residual signal results in a prominent display of the spectral peaks and notches. The positions of the spectral notches are obtained by marking and noting the frequencies of all the valleys below the zero threshold value of the group delay. In this study, a slightly lower threshold of -1 dB is used to avoid any spurious notches.

The LP analysis can also be used to extract the spectral peaks in the HRTF spectrum. The poles extracted by LP analysis seem to be corresponding to the resonances of the pinna as reported by [116]. The first mode is a simple quarter-wavelength depth resonance with uniform sound pressure across the base of the concha. This mode corresponds to the peak $P1$ in the HRTF spectrum. The notches $N1$ and $N2$ are selected based on the position of the peak $P1$.

7.5 Modeling HRTFs in the median plane using frontal projection headphones

The frontal projection headphone response captures the idiosyncratic pinna notches that are present in the HRTFs. The steps involved in modeling the individualized HRTFs in the median plane are as follows (Fig. 7.4):

1. Obtain frontal projection headphone response of the subject (Figure 7.2).
2. Using the Raykar's extraction algorithm [8] described in the previous section, the notches, and the peaks of the frontal projection headphone response are identified.

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.4: The different steps in the pinna extraction algorithm developed by Raykar *et al.* [8].

3. The position of $P1$ (around 3-4 kHz) is first noted and the subsequent notches are identified as $N1$ and $N2$.
4. Similarly, the attributes of the notches $N1$, $N2$ are extracted for all the subjects in the CIPIC database. In particular, the variation of these notches with elevation is studied.
5. The average value of notch shifts for all the subjects in the CIPIC database is computed for every elevation angle. The extracted $N1$ and $N2$ notches in the frontal headphones response is shifted by the average shift corresponding to each elevation angle to simulate elevation perception.
6. The head effects are modeled by incorporating the spherical head model (Section 6.3).
7. Additionally, the torso effects can be extracted from the CIPIC database by win-

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.5: Variation of center frequencies of the peak $P1$ and notches $N1$, $N2$ for Subjects 10,11, 48 and 138 taken from the CIPIC database. The extraction of the peak and notch positions were carried out using Raykar’s algorithm [8]. Note that the notch positions increase monotonically with elevation. On the other hand, the center frequency of the peak $P1$ is more or less constant with respect to elevation.

dowing the HATS dummy head/human subject’s HRIR.

8. A rear-directional filter is necessary to model the rear-directional HRTFs in the median plane (Section 6.3).
9. Low-frequency compensation filter is required to correct for the low-frequency response of the frontal projection headphone (Section 6.3).

Figure.7.5 shows the variation of the notch positions with elevation. It can be noted

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER
FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION
HEADPHONES

Figure 7.6: Rear directional filter for different elevation angles. Blue: Rear directional filters of all the subjects. Red (Bold): Average rear directional filter of all the subjects.

that the center frequencies of $N1$ and $N2$ increase monotonically with elevation. However, the center frequency of the peak $P1$ is more or less constant with elevation. Thus, the

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

peak $P1$ is not shifted in the modeling process. In the next section, the modeled median-plane HRTFs are compared with the measured median plane HRTFs with the help of objective and subjective analyses.

7.6 Objective analysis

The modeled median plane HRTFs using the frontal projection headphones are compared with the measured HRTFs. Figures 7.7, 7.8, and 7.9 show the modeled and measured frontal HRTFs for different elevations in the median plane for each of the three subjects.

For all the three subjects A, K, and R, it can be noted that the modeled frontal HRTFs match the measured HRTFs closely as shown in Figures 7.7, 7.8, and 7.9, respectively. The pinna notches estimated by the proposed frontal projection headphones playback model are close to the actual positions of the notches in the measured HRTFs.

Root mean square spectral differences (RSD) and the notch frequency distance (NFD) are considered as objective measures to compare the modeled and measured HRTFs at different elevations in the median plane. The RSD was previously defined in Eq. 6.8 in Section 6.2.

The NFD expresses the distance between $HRTF_{modeled}$ and $HRTF_{measured}$ by comparing the notch frequency positions for a particular subject [260]:

$$NFD_1 = \log_2 \{ f_{N1}(HRTF_{modeled}) / f_{N1}(HRTF_{measured}) \} [\text{oct.}], \quad (7.10)$$

$$NFD_2 = \log_2 \{ f_{N2}(HRTF_{modeled}) / f_{N2}(HRTF_{measured}) \} [\text{oct.}], \text{ and} \quad (7.11)$$

$$NFD = |NFD_1| + |NFD_2| [\text{oct.}], \quad (7.12)$$

where f_{N1} and f_{N2} denote the frequencies of the notch positions $N1$, $N2$, respectively.

The modeled rear-directional HRTFs in the median sagittal plane for the three sub-

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.7: Comparison of measured and modeled HRTFs in the median plane for subject A for different elevations.

jects is shown in Fig. 7.10. It can be seen from Fig. 7.10 that the modeled rear-directional HRTFs match the measured rear-directional HRTFs well. The RSD values are plotted in Fig. 7.11. The RSD values were found to be below 4 dB for the frontal directions. However, the RSD values are higher for the modeled rear-directional HRTFs compared to the frontal directional modeled HRTF. This is because of the rear-directional HRTF,

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.8: Comparison of measured and modeled HRTFs in the median plane for subject R for different elevations.

which is based on an average front-back spectral difference of the subjects taken from the CIPIC database. It can also be noted that the RSD error values are higher ($RSD > 4dB$) for lower elevation angles below the head.

Table 7.1 compares the notch frequency distances between the modeled and the measured HRTFs. From Table 7.1, we can immediately note that the frontal projection pinna

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER
FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION
HEADPHONES

Figure 7.9: Comparison of measured and modeled HRTFs in the median plane for subjects K for different elevations.

spectral notches is able to model the elevation dependent notches well. The mean NFD residual error for the $N1$ notch is found to be 0.0311, 0.0069, and 0.0406, respectively for the subjects A, K, and R. For the $N2$ notch, the mean NFD residual error is found to be in the range of 0.02 for all the subjects. The total NFD residual error also varies from 0.02-0.06. For higher elevation angles, the notch $N2$ slowly disappears. For all the

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.10: Modeled and measured rear-directional HRTFs in the median sagittal plane. The rear-directional filter is applied to the frontal HRTFs to obtain the rear-directional HRTFs. HRTFs for 20°, 60°, and 80° are shown for all the three subjects.

three subjects, the notch $N2$ disappeared for angles more than 60° elevation and thus, it is not shown in the table for 80°. In the next section, the proposed model is validated with the help of subjective analysis.

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.11: Root mean square spectral differences for the modeled frontal and the rear HRTFs for all the elevation angles.

7.7 Subjective analysis

Subjective experiments were carried out for 15 subjects to validate the modeled HRTFs at various elevations using the frontal projection headphones. The following stimulus conditions were presented to the subjects for whom the elevation localization were tested.

1. Modeled individualized HRTFs using frontal projection headphones.
2. Modeled individualized HRTFs using conventional headphones.
3. Measured individualized HRTFs using conventional headphones.
4. Non-individualized HRTFs using conventional headphones.

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

N1 notch

(a)

Subjects	HRTFs	Extracted N1 notch frequencies (Hz) for different elevation angles						
		-40°	-20°	0°	20°	40°	60°	80°
A	Measured HRTFs	5327	7027	8527	11170	12720	16930	17960
	Modeled HRTFs	5026	6886	8226	11040	12650	16700	17700
	NFD (oct.)	-0.08	-0.03	-0.05	-0.01	-0.008	-0.02	-0.02
K	Measured HRTFs	5370	7070	8570	11300	12850	17060	17960
	Modeled HRTFs	5327	7077	8527	11260	12890	16970	17830
	NFD (oct.)	-0.01	0.001	-0.008	-0.005	0.005	-0.008	-0.011
R	Measured HRTFs	4900	5706	7106	10310	11900	15980	17230
	Modeled HRTFs	5100	6000	7400	9969	11990	15700	17100
	NFD (oct.)	0.058	0.073	0.059	-0.049	0.011	-0.026	-0.011

(b)

N2 notch

Subjects	HRTFs	Extracted N2 notch frequencies (Hz) for different elevation angles					
		-40°	-20°	0°	20°	40°	60°
A	Measured HRTFs	8110	8910	11110	17490	18950	21200
	Modeled HRTFs	8280	9080	11280	16930	18800	21500
	NFD (oct.)	0.03	0.027	0.021	-0.047	-0.012	0.02
K	Measured HRTFs	7170	7970	10770	13660	16110	19800
	Modeled HRTFs	6900	7750	10500	13360	16030	19759
	NFD (oct.)	-0.06	-0.04	-0.04	-0.03	-0.01	-0.003
R	Measured HRTFs	7620	8820	12020	13920	16410	20580
	Modeled HRTFs	7490	8890	11890	13620	16160	20200
	NFD (oct.)	-0.025	0.012	-0.02	-0.031	-0.022	-0.027

(c)

NFD Total

Subjects	Residual Error		
	N1 (oct.)	N2 (oct.)	Total (oct.)
A	0.0311	0.0263	0.0221
K	0.0069	0.0291	0.036
R	0.0406	0.0221	0.0627

Table 7.1: Comparison of the notch frequency positions in the modeled and the measured HRTFs in terms of the notch frequency distance (NFD)

Subjects had to undergo an audiometry test before the commencement of the experiment and no subject reported any hearing loss. The basic stimulus used was a sequence of four bursts of cosine square gated (30-ms) WGN of 300-ms each separated by an interval

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

Figure 7.12: Subjective responses for measured (a)Conventional headphones) and modeled individualized HRTFs (b)Conventional headphones c) Frontal projection headphones) in the median plane. Responses for the non-individualized HRTFs (d)) is also plotted. Note that the errors are greater for the measured non-individualized HRTFs.

of 30 ms silence, as in the experiments in Section 6.3.2. Subjects were asked to report the “apparent” direction of elevation of the virtual source presented. In this experiment, virtual sources were created for the angles -80° to 260° at a resolution of 20° . Virtual sources at different elevations were synthesized for six azimuth angles in the frontal and rear directions.

It should be noted that in the proposed modeling technique, the frontal projection

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

headphones are required only during the beginning of the modeling stage for obtaining the frontal projection responses. Once the frontal projection response is obtained, the HRTFs for different elevations can be directly calculated using the techniques described in the previous section. These HRTFs obtained modeled from the frontal projection headphones responses can now be played back even using the conventional headphones (Stimulus 2). However, Type-1 equalization has to be used for conventional headphones playback in order to compensate for the headphones response.

Figure. 7.12 shows the subjective responses for the modeled and measured individualized HRTFs at different elevations in the median plane. The front-back confusions were investigated for each of 4 different stimulus conditions (Fig. 7.13). Front-back confusions were found to be 12.38%, 13.33 %, and 13.67 % for the measured (Fig. 7.12a) and the modeled (Fig. 7.12b) HRTFs played back over conventional and the frontal projection headphones (Fig. 7.12c). However, the front-back confusions for the non-individualized case were much higher (Fig. 7.12d). ANOVA analysis proved that the front-back confusions for the measured and the modeled HRTFs were similar with no statistically significant differences ($p > 0.05$). The front-back confusions in the non-individualized HRTFs were statistically significant compared to the modeled and measured HRTFs ($p < 0.01$).

Localization error was obtained by computing the angular differences between the actual elevation angle and the responded elevation angle. The front-back confusions were removed prior to the calculation of the localization error. ANOVA test conducted showed that the localization differences between the measured and the modeled HRTFs at different elevations is statistically not significant [$F(52, 2) = 0.35, p = 0.8$]. However, the localization errors with the non-individualized HRTFs were statistically significant as compared to the modeled individualized elevation department HRTFs [$F(35, 1) = 7.5p < 0.01$]. Localization error for the elevation angles below and rear-directions were slighter larger than for frontal directions, however, the ANOVA analysis showed no significant differ-

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER
FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION
HEADPHONES

Figure 7.13: Front-back confusions for the 4 stimulus conditions.1) Measured individualized HRTFs 2) Modeled HRTFs over conventional headphones 3) Modeled HRTFs over frontal projection headphones 4) Measured non-individualized HRTFs.

ences.

7.8 Summary

In this chapter, the elevation dependent individualized HRTFs were modeled in the median plane using the frontal projection headphones playback. The frontal projection headphone playback along with a cascade of signal processing blocks models all the necessary spectral features required for elevation perception. It was found that the proposed HRTF model with both conventional, as well as frontal projection playback, displayed performance similar to the measured individualized HRTFs. Since the monaural spectral cues essentially do not vary much with azimuth, the model can be directly extended to

CHAPTER 7. INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS IN THE MEDIAN PLANE USING FRONTAL PROJECTION HEADPHONES

any sagittal plane (other than the mid-sagittal) by incorporating a frequency-independent time delay.

In the next chapter, signal processing techniques for the natural sound rendering over headphones for digital media content is presented. A special 4-channel 3D audio headphones is developed that improves the spatial quality and thus helps in natural sound rendering. Subjective experiments are carried out to validate the proposed natural sound rendering system.

Chapter 8

Natural sound rendering for digital media content using 3D audio headphones: An integration of signal processing techniques

In this chapter, a natural sound rendering technique for typical digital media content (stereo, multichannel) is described. A unique four-emitter headphones known as the 3D audio headphones is developed that holds the conventional side-emitter as well as the frontal projection emitter on each side. Signal processing techniques are introduced for natural sound rendering of digital media content to assist human listening. Key challenges associated with natural sound rendering over headphones are discussed. The work presented in this chapter has been published in the IEEE Signal Processing Magazine in the March 2015 issue [35].

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

8.1 The need for natural sound rendering

In most of these digital entertainment applications, listening is seldom from the physical sound sources but instead from playback devices, such as headphones or loudspeakers. Headphones, by virtue of their convenience and portability, are typically chosen as the preferred playback device, especially for personal listening. Therefore, to assist headphone listening, it is critical for the sound to be rendered in a way that listeners can perceive it as natural as possible. In this context, natural sound rendering essentially refers to rendering of the original sound scene using headphones to create an immersive listening experience and the sensation of “being there” at the venue of the acoustic event. To achieve natural sound rendering, the virtual sound rendered should exactly emulate all the spatial cues of the original sound scene, as well as the individual spectral characteristics of the listener’s ears. In this chapter, we mainly consider the most widely used channel-based audio as the input signals for the natural sound rendering system, though some of the signal processing techniques discussed could also be used in other audio formats, such as object-based format and ambisonics [269, 270].

In recent years, the design criteria for commercial headphones have undergone significant development. At Harman, Olive *et al.* investigated the best target responses for designing headphones based on the listener’s preference for the most natural sound [139]. Creating realistic surround sound in headphones has become a common pursuit of many headphone technologies from Dolby, DTS, etc. Furthermore, personalized listening experience and incorporating the information of listening environment are the recent trends in headphone industry. These trends in headphones share one common objective: to render natural sound in headphones.

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

8.2 Challenges

In natural listening, we listen to the physical sound sources in a particular acoustic space, with the sound waves undergoing diffraction and interference with different parts of our morphology (torso, head and pinna) before reaching the eardrum. Listeners also get valuable interaural cues for sound localization with head movements. However, headphone listening is inherently different from natural listening, as the sources we are listening to are no longer physical sound sources but are recorded and edited sound materials. These differences between natural and headphone listening lead to various challenges in rendering natural sound over headphones, which can be broadly classified into the following three categories:

1. From the perspective of source, the sound scenes rendered for headphone listening should comprise not only the individual sound sources but also the features of the sound environment. Listeners usually perceive these sound sources to be directional, i.e., coming from certain directions. Moreover, in most of the digital media content, the sound environment is usually perceived by the listener to be diffuse (partially). This perceptual difference between the sound sources and the sound environment requires them to be considered separately in natural sound rendering [270]. Though there are other formats that can represent the sound scenes (e.g., object-based, ambisonics), the convention for today's digital media is still primarily channel-based format. Hence, the focus of this chapter lies in the rendering of channel-based audio, where sound source and environment signals are mixed in each channel [270].

In channel-based signals, where only the sound mixtures are available (assuming one mixture in every channel), it is necessary to extract the source signals and environment signals, which can be quite challenging. Furthermore, most of the traditional recordings are processed, and mixed for optimal playback over loudspeakers, rather

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

than headphones. Direct playback of such recordings over headphones results in an unnatural listening experience, which is mainly due to the loss of crosstalk, and localization issues.

2. From the perspective of medium, headphone listening does not satisfy free-air listening conditions as in natural listening. Since HPTF is not flat, equalization of the headphone is necessary. However, this equalization is tedious and challenging, as the headphone response is highly dependent on the individual anthropometrical features and also varies with repositioning.
3. From the perspective of receiver, the omission of listener's individualized filtering with the outer ear in headphone listening often leads to coloration and localization inaccuracies. These individualized characteristics of the listener are lost when the sound content is recorded or synthesized non-individually. Furthermore, the sound in headphone listening is not adapted to the listener's head movements, which departs from a natural listening experience.

8.3 Signal processing techniques

To tackle the above challenges and enhance natural sound rendering over headphones for digital media content, digital signal processing techniques are commonly used. In Fig. 8.1, we summarize the differences between natural listening and headphone listening, and introduce the corresponding signal processing techniques to tackle these challenges, which are:

1. Virtualization: to match the desired playback for the digital media content.
2. Sound scene decomposition using blind source separation (BSS) and primary-ambient extraction (PAE): to optimally facilitate the separate rendering of sound sources

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Figure 8.1: A summary of the differences between natural listening and headphone listening and the corresponding signal processing techniques to solve these challenges for natural sound rendering. The main challenges and their corresponding signal processing techniques in each category (source, medium, and receiver) are highlighted and their interactions (not shown here) are further discussed in the chapter.

and sound environment.

3. Individualization of HRTF: to compensate for the lost or altered individual filtering of the sound in headphone listening.
4. Equalization: to preserve the original timbral quality of the source and alleviate the adverse effect of the inherent headphone response.
5. Head tracking: to adapt to the dynamic head movements of the listener.

The remainder of this chapter is structured as follows. Virtualization and head tracking, due to their high interactions, are explained together in Section 8.4, followed by the decomposition of sound scenes in Section 8.5. Sections 8.6 and 8.7 describe individualization and equalization, respectively. These signal processing techniques are integrated and evaluated using subjective tests in Sections 8.8 and 8.10, respectively. The development

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Figure 8.2: Virtualization of (a) multichannel loudspeaker signals $x_m(n)$ (adapted from [8]), and (b) multiple sources $s_k(n)$ and environment signals $a_L(n)$, $a_R(n)$, $y_L(n)$, $y_R(n)$ are the signals sent to the left and right ear, respectively. Note that head tracking can be used to update the selected directions of HRTFs/BRIRs.

of 3D audio headphone is described in Section 8.9. Finally, the conclusions and future trends are presented in Section 8.11.

In digital media, sound is typically mixed for loudspeaker rather than headphone playback. The spatial sound to be rendered naturally over headphones should emulate the natural propagation of the acoustic waves emanated from the loudspeaker to the eardrum of the listener. To emulate stereo or surround sound loudspeaker rendering over headphones, virtualization techniques based on HRTF corresponding to the loudspeaker positions are commonly used. Given these acoustic transfer functions (i.e., HRTFs), the virtualization technique is applicable to any multichannel loudspeaker setup, be it stereo, 5.1, 7.1, 22.2, or even loudspeaker arrays in wave-field synthesis. As shown in Fig.8.2(a), for every desired loudspeaker position, the signal in the m^{th} channel $x_m(n)$ is filtered

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

with the corresponding HRTF $h_{xmL}(n)$, $h_{xmR}(n)$, and summed before being routed to the left and right ears [12, 271], respectively, as:

$$y_L(n) = \sum_{m=1}^M h_{xmL}(n) * x_m(n), \text{ and} \quad (8.1)$$

$$y_R(n) = \sum_{m=1}^M h_{xmR}(n) * x_m(n), \quad (8.2)$$

where $*$ denotes convolution and M is the total number of channels. When the HRTFs are directly applied to multichannel loudspeaker signals, the rendered sound scenes in headphone playback suffer from inaccurate virtual source directions, lack of depth, and reduced image width [271, 272].

To solve these problems in virtualization of multichannel loudspeaker signals and achieve a faithful reproduction of the sound scenes, the HRTFs should be applied to the individual source signals that are usually extracted (using BSS, PAE) from the loudspeaker signals (i.e. mixtures). In this virtualization, as shown in Fig. 8.2(b), the sources are rendered directly using the HRTFs of the corresponding source directions $h_{skL}(n)$, $h_{skR}(n)$:

$$y_L(n) = \sum_{k=1}^K h_{skL}(n) * s_k(n) + a_L(n), \text{ and} \quad (8.3)$$

$$y_R(n) = \sum_{k=1}^K h_{skR}(n) * s_k(n) + a_R(n), \quad (8.4)$$

where K is the total number of sources, $s_k(n)$ is the k^{th} source in the multichannel signal, and the environment signals $a_L(n)$, $a_R(n)$, are the rendered signals representing the sound environment perceived at two ears. To render the acoustics of the environment, the environment signals can be either synthesized according to the sound environment [273] or extracted from the mixtures. Techniques like decorrelation [271, 274] and artificial

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

reverberation [275] are commonly employed to render the environment signals in order to create a more diffuse and natural sound environment.

Furthermore, adding the reverberation of sources (or the loudspeaker signals in virtualization of multichannel loudspeaker signals) can also improve the realism of the reproduced sound scene [108]. Therefore, in virtualization, it is quite common to use binaural room impulse response (BRIR) [12,271], that encapsulates HRTFs and reverberation. On this note, selecting the correct amount of early reflections, as well as late reverberation, is critical to recreate a faithful sound environment [12]. In general, the BRIR that matches the sound environment of the scene or BRIR of a mixing studio are considered to be more suitable [139]. As discussed in Section 8.2, natural sound rendering requires the accurate reproduction of both the sound sources and the sound environment. Compared to the virtualization of multichannel loudspeaker signals (Fig. 8.2(a)), the latter technique of virtualizing the source and environment signals (Fig. 8.2(b)) is more desirable as it is closer to natural listening [272, 274, 276]. These virtualization techniques can also be incorporated into spatial audio coding systems, such as binaural cue coding [277], spatial audio scene coding [271], and directional audio coding [269].

In virtualization, the directions of the sources (or the loudspeakers in virtualization of multichannel loudspeaker signals as in Fig. 8.2(a)) have to be calibrated according to the head movements (as in natural listening). To fulfil this need, the HRTFs/BRIRs in the virtualization are updated on the fly based on these head movements that are often tracked by a sensor (e.g., accelerometer, gyroscope, camera, etc.). The latency between the head tracking and sound rendering should be such that the localization accuracy is not affected (latency < 60 ms) [275]. Such a head tracking system when incorporated in the virtualization process can provide useful dynamic cues to resolve the localization conflicts [12] and enhance natural sound rendering [108, 275]. It is noted that head tracking is more critical for the directional sources, but less important for the

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

diffuse signals like environment signals and late reverberation [275]. This is because the perception of diffuse signals is less affected by head movements.

Recreating the perception of distance of the sources close to natural listening is another critical aspect in virtualization for natural sound rendering. However, the challenges in simulating accurate distance perception are aplenty. The ability of human beings to accurately estimate the distance has long been known to be poorer compared to their direction localization ability even in the physical listening space [12]. Virtual listening over headphones further hinders the distance perception as it leads to inside-the-head localization (IHL) of sound [12]. IHL of sound is caused by several factors, such as the use of non-individualized HRTFs, absence of equalization, lack of reverberation, impedance mismatch due to the presence of headphones [42]. Presence of individualized HRTFs, equalization and reverberation can improve the externalization of sound but does not ensure accurate distance perception [12]. Direct-to-reverberation energy ratio is found to be the most critical cue for absolute distance perception, even though the intensity, loudness, and binaural cues can provide relative cues for distance perception [12]. Since reverberation is an essential cue for both distance perception and perception of a real environment context, a veridical simulation of the reverberation is highly imperative for natural sound rendering [12]. However, accurate simulation of distance perception is challenging, since reverberation entirely depends on the room characteristics. The correct amount of reverberation to be added to simulate distance perception in a particular room can be obtained only by carrying out acoustical measurements.

8.4 Sound source decomposition using BSS and PAE

To achieve natural sound rendering in headphones, two important constituents of the sound scenes are required in the virtualization, namely, the individual sound sources and characteristics of the sound environment. However, this information is usually not

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

directly available to the end user. One has to work with the existing digital media content that is available, i.e., the mastered mix distributed in channel-based formats (e.g., stereo, 5.1). Therefore, to facilitate natural sound rendering, it is necessary to extract the sound sources and/or sound environment from their mixtures. In this section, we discuss two types of techniques applied in sound scene decomposition, namely, BSS and PAE.

8.4.1 Decomposition using BSS

Extracting the sound sources from the mixtures, often referred to as BSS, has been extensively studied in the last few decades. The basic mixing model in BSS can be considered as anechoic mixing, where the sources $s_k(n)$ in each mixture $x_m(n)$ have different gains g_{mk} and delays τ_{mk} . Hence, the anechoic mixing is formulated as follows:

$$x_m(n) = \sum_{k=1}^K g_{mk} s_k(n - \tau_{mk}) + e_m(n), \forall m \in \{1, 2, \dots, M\}, \quad (8.5)$$

where $e_m(n)$ is the noise in each mixture, which is usually neglected for most cases. Note that estimating the number of sources is quite challenging and it is usually assumed to be known in advance [278]. This formulation can be simplified to represent instantaneous mixing by ignoring the delays, or can be extended to reverberant mixing by including multiple paths between each source and mixture. An overview of the typical techniques applied in BSS is listed in Table 8.1.

Based on the statistical independence and non-Gaussianity of the sources, ICA algorithms have been the most widely used techniques in BSS to separate the sources from mixtures in the determined case, where the numbers of mixtures and sources are equal [278]. In the over-determined case, where there are more mixtures than sources, ICA is combined with PCA to reduce the dimension of the mixtures, or combined with least-squares (LS) to minimize the overall mean-square error (MSE) [279]. In practice, the under-determined case is the most common, where there are fewer mixtures than

**CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES**

Table 8.1: Overview of typical techniques in BSS

Objective: To extract K ($K > 2$) sources from M mixtures		
Case		Typical techniques
Determined: $K = M$		ICA [278]
Over-determined: $K < M$		ICA with PCA or LS [278, 279]
Under-determined: $K > M$	$M > 2$	ICA with sparse solutions [278, 280]
	$M = 2$	Time-frequency masking [273, 281–283]
	$M = 1$	NMF [284–286]; CASA [287]

sources. For the under-determined BSS, sparse representations of the sources are usually employed to increase the likelihood of sources to be disjoint [280]. The most challenging under-determined BSS is when the number of mixtures is two or lesser, i.e., in stereo and mono signals.

Stereo signals (i.e., $M = 2$), being one of the most widely used audio format, have been the focus in BSS. Many of these BSS techniques can be considered as time-frequency masking and usually assume one dominant source in one time-frequency bin of the stereo signal [282]. In these time-frequency masking based approaches, a histogram for all possible directions of the sources is constructed, based on the range of the bin-wise amplitude and phase differences between the two channels. The directions, which appear as peaks in the histogram, are selected as source directions. These selected source directions are then used to classify the time-frequency bins, and to construct the mask. For every time-frequency bin (n, l) , the k^{th} source at m^{th} channel $\hat{S}_{mk}(n, l)$ is estimated as:

$$\hat{S}_{mk}(n, l) = \psi_{mk}(n, l) X_m(n, l), \tag{8.6}$$

where the mask and the m^{th} mixture are represented by $\psi_{mk}(n, l)$ and $X_m(n, l)$, respectively. In the case of single-channel (or mono) signals, the separation is even more challenging, since there is no inter-channel information. Hence, there is a need to look into the inherent physical or perceptual properties of the sound sources. Non-negative matrix factorization (NMF) based approaches are extensively studied and applied in single-channel

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

BSS in recent years. The key idea of NMF is to formulate an atom-based representation of the sound scene [284], where the atoms have repetitive and non-destructive spectral structures. NMF usually expresses the magnitude (or power) spectrogram of the mixture as a product of the atoms and time varying non-negative weights in an unsupervised manner. These atoms, after being multiplied with their corresponding weights, can be considered as potential components of sources [286]. Another technique applied in single-channel BSS is the computational auditory scene analysis (CASA) that simulates the segregation and grouping mechanism of human auditory system [287] on the model-based representation (monaural case) of the auditory scenes. An important aspect worth considering is the directions of the extracted sources, which can usually come as a by-product in multichannel BSS. In single-channel BSS, this information of source directions has to be provided separately.

8.4.2 Decomposition using PAE

In most sound scenes, the mixture comprises not only the dry sources but also the reverberation and ambient sound, which are contributed by the acoustics of the surrounding space. Therefore, the mixing model of the sources in BSS usually does not match with the actual sound scenes. In this chapter, we refer to the dominant sources as primary (or direct) components, while the signals contributed by the sound environment as ambient (or diffuse) components. The primary and ambient components are perceived to be directional and diffuse, respectively. Different rendering methods should be applied to the primary and ambient components [272, 273] due to their perceptual differences. Therefore, rendering of natural sound scenes requires the decomposition of the mixtures into primary and ambient components [272, 273, 276]. Since stereo is still the most widely used format for digital media content, our discussion on the decomposition using primary-ambient extraction is focused on stereo signals ($M = 2$).

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

In PAE, we often follow some intuitive signal models as discussed in [269, 271, 273, 274, 288]. In the m^{th} channel, the mixture $x_m(n)$ is assumed to be the sum of the primary component $p_m(n)$ and ambient component $a_m(n)$, i.e., $x_m(n) = p_m(n) + a_m(n)$. The discrimination of directional primary components and diffuse ambient components is mainly based on their inter-channel correlations, where the primary and ambient components in the two channels are assumed to be correlated and uncorrelated, respectively. In the basic mixing model for PAE, the primary components are assumed to be amplitude panned, while the ambient components are of approximately equal levels in all channels.

Based on these assumptions, various approaches are proposed in PAE for stereo signals. Similar to BSS, time-frequency masking approaches are introduced to extract ambient components $\hat{A}_m(n, l)$ [273, 288, 288] and these approaches can be generalized as :

$$\hat{A}_m(n, l) = X_m(n, l) \psi_A(n, l), \quad (8.7)$$

where $0 \leq \psi_A(n, l) \leq 1$ is the real-valued ambient mask at time-frequency bin (n, l) . Time-frequency bins, which have high inter-channel correlation, are considered to be primary components (or mostly primary components in the soft masking case), whereas low correlation bins are more likely to be ambient components.

Several linear estimation based PAE approaches were also introduced in [36], which exploits the differences between the two channels of the stereo signal to perform the primary-ambient extraction, including PCA based approaches [288] and LS based approaches. In these approaches, the extracted primary components $\hat{p}_0(n)$, $\hat{p}_1(n)$ and

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Table 8.2: Comparison between BSS and PAE in sound scene decomposition.

	BSS	PAE
Objective	To obtain useful information about the original sound scene from given mixtures, and facilitate natural sound rendering.	
Common characteristics	<ul style="list-style-type: none">• Usually no prior information, only mixtures.• Based on certain signal models• Require objective as well as subjective evaluation	
Basic mixing model	Sums of multiple sources (independent, non-Gaussian, etc.)	Primary components, (highly correlated)+ Ambient components (uncorrelated)
Techniques	ICA [278], Speech, music sparse solutions [280], time-frequency masking [282], NMF [284, 285], CASA [287], etc.	PCA [271, 288], LS [274, 289], time-frequency masking [273, 288], signal model classification [290], time/ phase-shifting [36, 291, 292], using pairwise correlations [293]
Typical applications	Speech, music	Movie, gaming
Related applications	Speech enhancement, noise reduction, speech recognition, music classification	Sound reproduction, sound localization, coding
Limitations	<ul style="list-style-type: none">• Small number of sources• Sparseness/disjoint• No/simple environment	<ul style="list-style-type: none">• Small number of sources• Sparseness/disjoint• Low ambient power• Primary ambient components uncorrelated

ambient components $\hat{a}_0(n)$, $\hat{a}_1(n)$ are expressed as weighted sums of the mixtures:

$$\begin{bmatrix} \hat{p}_0(n) \\ \hat{p}_1(n) \\ \hat{a}_0(n) \\ \hat{a}_1(n) \end{bmatrix} = \begin{bmatrix} w_{P0,0} & w_{P0,1} \\ w_{P1,0} & w_{P1,1} \\ w_{A0,0} & w_{A0,1} \\ w_{A1,0} & w_{A1,1} \end{bmatrix} \begin{bmatrix} x_0(n) \\ x_1(n) \end{bmatrix}. \tag{8.8}$$

The solutions for the weights in Eq.8.8 are derived based on different performance-related criteria [36]. More specifically, PCA extracts the primary components having maximum variance, and extracts the ambient components having minimum variance with the constraint that the primary and ambient components are uncorrelated, while

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

LS extracts these components having minimum MSE. Based on the study by He *et al.* [36], it is recommended that PCA based approaches should be used for signals that contains dominant primary components (e.g. gaming), while LS based approaches are preferred for signals that contain a balanced mix of primary and ambient components (e.g., movies). In addition, to deal with more complex types of input signals that do not fit into the basic mixing model, other techniques have also been introduced, such as, signal model classification [290], compensation of time/phase differences in primary components [36, 291, 292], and using pairwise correlations [293], and adaptive frequency bin partitioning for multiple sources in primary components [289]. Furthermore, though it is possible to extend the framework of PAE from stereo signals to multichannel signals, e.g., [293], more comprehensive studies on PAE for multichannel signals are required.

8.4.3 A comparison between BSS and PAE

Both BSS and PAE are extensively applied in sound scene decomposition, and a comparison between these approaches is summarized in Table 8.2. The common objective of BSS and PAE is to extract useful information (mainly the sound sources and their directions) about the original sound scene from the mixtures, and to use this information to facilitate natural sound rendering. On this note, there are three common characteristics in BSS and PAE. First, only the mixtures are available and usually no other prior information is given. Second, the extraction of the specific components from the mixtures is based on certain signal models. Third, both techniques require objective and subjective evaluation.

As discussed earlier, the applications of different signal models in BSS and PAE lead to different techniques. In BSS, the mixtures are considered as the sums of multiple sources, and the independence among the sources is one of the most important characteristics. In contrast, the mixing model in PAE is based on human perception of directional sources

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

(primary components) and diffuse sound environment (ambient components). The perceptual difference between primary and ambient components is due to the directivity of these components, which can be characterized by their correlations. The applications that adopted BSS and PAE also have distinct differences. BSS is commonly used in speech and music applications, where the clarity of the sources is usually more important than the effect of the environment. On the other hand, PAE is more suited for the reproduction of movie and gaming sound content, where the ambient components also contribute significantly to the naturalness and immersiveness of the sound scenes. Subjective experiments revealed that BSS and PAE based headphone rendering can improve the externalization and enlarge the sound stage with minimal coloration [272].

Despite the recent advances in BSS and PAE, the challenges due to the complexity and uncertainty of the sound scenes still remain to be resolved. One common challenge in both BSS and PAE is the increasing number of audio sources in the sound scenes, while only a limited number of mixtures (i.e., channels) are available. In certain time-frequency representations, the sparse solutions in BSS and PAE would require the sources to be sparse and disjoint [280]. Considering the diversity of audio signals, finding a robust representation for different types of audio signals is extremely difficult. The recorded or post-processed source signals might even be filtered due to physical or equivalently simulated propagation and reflections.

Moreover, the audio signals coming from adverse environmental conditions (including reverberation, and strong ambient sound) usually degrade the performance of the decomposition. These difficulties can be addressed by studying the features of the resulting signals and by obtaining more prior information on the sources, the sound environment, the mixing process [286], and combining with visual information of the scene.

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

Figure 8.3: (a) Human ears act as a natural filter in physical listening. (b) The natural HRTF filter is modeled by a digital filter using various individualization techniques.

8.5 Individualization of HRTFs

Binaural technology is the most promising solution for delivering spatial audio in headphones, as it is the closest to natural listening (Fig. 8.3). Unlike conventional microphone recordings, which are meant for loudspeaker playback, the binaural signals are recorded or synthesized at the ears of the listener. The different components of a binaural audio system have already been explained in Chapter 2. Need of individualized HRTFs arises from the fact that HRTFs are highly idiosyncratic and use of non-individualized HRTFs degrade the spatial audio perception. There are various individualization techniques to obtain the individualized HRTFs from acoustical measurements, anthropometric features of the listener, customizing generic HRTFs with perceptual feedback or frontal projec-

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

tion of sound, as summarized in Table 4.1. In this chapter, frontal projection headphone playback is used as the preferred individualization technique.

8.6 Equalization

Equalization of headphones is critical to ensure natural sound perception. It is observed that, in the absence of headphone equalization, the front-back reversals are increased and the elevation localization is distorted [12, 42, 188]. The different equalization techniques for both binaural and stereophony content have been listed out in Table 2.4 in Chapter 2. Type-2 equalization is used in this study, since the frontal projection headphones are used for playback.

8.7 Integration of natural sound rendering techniques

An integration of these signal processing techniques for natural sound rendering reviewed in this chapter is depicted in Fig. 8.4. The original sound sources along with their environmental information are represented as a sound mixture after the mixing process. The sound scenes from the mix are then decomposed into primary components (sources) and/or ambient components (environment) using BSS and/or PAE. The extracted primary components, which are basically directional sound sources as perceived by the listener, can be rendered using (individualized) HRTFs [12]. Ambient components are rendered in a manner to recreate a natural sound environment. Modeling the acoustics of the natural sound environment by adding the correct amount of early reflections and reverberation also helps in enhancing the perception of the sound environment as well as veridical distance, which is critical for natural listening. Moreover, a suitable individualization technique has to be applied to the directional sources, such that the rendered sound scenes played over headphones are maximally tailored for the individual

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Figure 8.4: Natural sound rendering system for headphones: an integration of all the signal processing techniques reviewed in this paper.

listener. The use of a robust equalization technique can significantly reduce the adverse coloration of the source. Finally, the influence of the head movements on the rendered sound can be taken into account by incorporating head tracking in virtualization.

In general, natural sound rendering requires both the spatial and timbral quality of the reproduced sound to be realistic. For digital media content that contains plenty of spatial cues (e.g., movies, games), all the five techniques reviewed are important in creating a sense of immersiveness. For other content, where the timbral quality is of utmost importance (e.g., music recordings), a subset of the techniques (e.g., individualization, equalization) are sufficient in natural sound rendering.

8.8 3D Audio headphones

To render natural sound, a four emitter headphone [34] is developed that holds both conventional side emitter, as well as the frontal projection emitter on each side [144] (Fig. 8.5). From the study carried out in earlier chapters, we find that the localization is en-

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL PROCESSING TECHNIQUES

Figure 8.5: (a) Skeleton of the 3D audio headphones holding both the frontal projection as well as the conventional side emitter (b) A finished 3D audio headphone prototype.

hanced with the frontal projection headphone devoid of any confusion. The conventional side emitter performs poorer compared to the frontal projection headphone in localization experiments, and thus, is more suited to playback the ambience or the environment signal. To create a prototype of the 3D audio headphones for further testing, the earcups holding the frontal and side emitters were 3D printed [34]. Subjective experiments are reported in the next section to validate the natural sound rendering system using the 3D audio headphones.

8.9 Subjective experiments

Subjective experiments were carried out to validate the reviewed natural sound rendering system by comparing it with the conventional stereo playback system. A total of

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Table 8.3: Stimuli used in the experiments.

Type	Track	Duration
Binaural	Motorcycle in the storm	1:07
	Bee at the waterfall	0:20
Movie	Brave	2:59
	Prometheus	2:24
Gaming	Batterfield 3	1:49

18 subjects (15 males and 3 females), who were all between 20-30 years old, participated in this listening experiment. None of the subjects reported any hearing loss. The test was conducted in a semi-anechoic listening chamber at Nanyang Technological University, Singapore. The two systems of headphone listening tested in this experiment were:

1. **Conventional stereo system:** The materials are directly played back over headphones without any processing.
2. **Natural sound rendering system:** The signal processing techniques introduced in the chapter were applied to the audio content. In this study, we chose PAE as the sound scene decomposition method, since our primary interest lies in movie and gaming audio content that contains the individual sound sources and the sound environment [292].

Individualization is carried out by frontal projection headphone pinna cues during playback and does not require any individual acoustical experiments, anthropometric data or training [144]. In the virtualization process, the frontal emitters are used to render the directional sources, while all the emitters (both frontal and side) are used to render the sound environment. Type-2 EQ is applied to the frontal emitters for source rendering [144], and diffuse-field Type-1 EQ is applied to the side projection headphones. Head tracking has not been incorporated in this system.

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Figure 8.6: A screenshot of the GUI designed for the test

The stimuli used in this experiment were binaural, movie, and gaming tracks, which contain plenty of spatial cues (Table 8.3). Each track was played back using the two headphone playback systems tested here. The tracks corresponding to the two systems were named “A” and “B” and played back in a random order. The listening tests were conducted in a double-blind manner, where both the experimenter and the subjects were unaware of the order of the stimuli. In this experiment, four audio quality measures were considered to evaluate the performance of the two systems. Their descriptions are given below:

1. **Sense of direction:** how clear or distinct are the perceived directions of the sound objects?
2. **Externalization:** how clear is the stimulus perceived outside the head?
3. **Ambience:** how clear and natural is the ambience of the sound environment perceived?

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Figure 8.7: MOS of the 4 measures for each track: (a) sense of direction; (b) externalization; (c) ambience; and (d) timbral quality.

4. **Timbral quality:** how realistic is the timbral quality of the sound?

Subjects were asked to give the scores for the four measures pertaining to each of the two tracks “A” and “B”. The scores were based on a 0-100 scale (Recommendation, 2003) where subjects rated 0-20 (Bad), 21-40 (Poor), 41-60 (Fair), 61-80 (Good), and 81-100 (Excellent). Finally, the subjects were also required to indicate their overall preference for the two tracks by selecting one of the following three choices: “Prefer A”, “Not sure”, or “Prefer B”. To carry out this experiment, a GUI (Fig.8.6) was created, which randomized the order of the stimuli, and automatically stored the responses of the subjects in a file computed across all the 18 subjects and 5 stimuli. The responses of the subjects were analyzed for both the sound rendering systems. The mean opinion scores (MOS) of each of the four measures for all the five tracks are shown in Fig. 8.7. It was observed that the subjects had rated higher for the natural sound rendering system than the conventional stereo system. This observation was consistent for all the measures across all the five tracks. The scatter plot in Fig. 8.8 implies that most of the subjects gave a higher score for the natural sound rendering system for all the four measures.

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Figure 8.8: Scatter plot for 4 measures: (a) sense of direction; (b) externalization; (c) ambience; and (d) timbral quality.

Figure 8.9 shows the overall comparison between the two systems in terms of the mean opinion score (MOS), scatter plot and the overall preference of the subjects. While the MOS for the conventional stereo system for all the measures were around 60, the natural sound rendering system performed much better with MOS of over 70. An analysis of variance (ANOVA) was conducted to generalize these results to the whole population of listeners. The p-values were found to be very small ($\ll 0.01$) for all the measures, indicating that the improved performance of the natural sound rendering system over the conventional stereo system is statistically significant. The overall preference of the subjects across all the five tracks is shown in Fig. 8.9(c). The pie chart suggests that 61% of the subjects preferred the natural sound rendering, while only 33% preferred the conventional stereo rendering.

To sum up the subjective test results, we found that the natural sound rendering system using the various signal processing techniques explained in this chapter enhances the listening experience compared to a conventional stereo system. Additionally, the presence of head tracking in the system will only improve the natural sound rendering, as observed in several studies [108].

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Figure 8.9: Comparison of the overall performance: (a) MOS for 4 measures; (b) scatter plot for all scores; and (c) preference of the tracks.

8.10 Conclusions

With the advent of low cost, low power, small form factor, and high speed multi-core embedded processors, we can now implement the above signal processing techniques in real-time and embed processors into the headphone design. However, various implementation issues regarding the computation cost of sound scene decomposition, HRTF/BRIR filtering in virtualization, and equalization as well as the latency in head tracking, should be carefully considered. A unique four-emitter 3D audio headphone is developed at the DSP Lab in NTU [34, 35]. The natural sound rendering system has been psychophysically validated and found to perform much better than the conventional stereo headphone playback system.

Besides the five types of techniques discussed in this chapter, there have been other efforts to enhance the natural experience of headphone listening. To enable the natural pass-through of the sound from outside world without coloration, headphones can be designed with suitable acoustically transparent materials. When this is not effective,

CHAPTER 8. NATURAL SOUND RENDERING FOR DIGITAL MEDIA
CONTENT USING 3D AUDIO HEADPHONES: AN INTEGRATION OF SIGNAL
PROCESSING TECHNIQUES

Table 8.4: ANOVA results for MOS of the four measures.

Measures	<i>F</i> -Statistic Values	<i>p</i> values
Sense of direction	$F(1,178) = 48.54$	6×10^{-11}
Externalization	$F(1,178) = 23.11$	3×10^{-6}
Ambience	$F(1,178) = 45.34$	2×10^{-10}
Timbral quality	$F(1,178) = 120.25$	3×10^{-4}

microphones integrated into headphones and associated signal processing techniques, such as equalization [294], and active noise control [295] are employed. The headphones with built-in microphones open a new dimension to augment the listening experience with the physical world [296].

The future of headphones especially for assistive listening applications would be the one where listeners cannot differentiate between the virtual acoustic space created from headphone playback and the real acoustic space. This would require the combined effort from the whole audio community comprising the headphone manufacturers, sound engineers and the audio scientists. More information about the content production has to be distributed from the content developers to the end user to enhance the extraction process. Moreover, obtaining and exploiting every individual’s anthropometrical features or hearing profiles is crucial for a natural listening experience. Finally, with more sensors, such as GPS, gyroscopes, and microphones that can be integrated into headphones, future headphones can be more location-aware, listener-aware, and hence become more intelligent and assistive.

Chapter 9

Conclusions and Future Works

9.1 Conclusions

In this thesis, we have investigated into some of the core challenges of binaural audio reproduction, i.e. individualization and equalization. Solutions for the challenges posed in binaural audio to obtain a highly customized or individualized 3D audio are presented in this thesis. We have looked into all the three dimensions of spatial audio, namely, azimuth, elevation, and distance.

Individualized HRTFs were modeled in order to obtain a veridical perception of 3D audio. Binaural audio rendering over conventional side projection headphones were found to be not accurate and highly subject-dependent. This leads to several issues like in-head localization, timbral coloration, and directional reversals. Binaural audio reproduction using frontal projection headphones were studied as part of this thesis. The frontal projection headphones were used due to its unique structure where the transducer is placed in the front of the ear unlike the conventional side emitter headphones.

Most of the individualization techniques in literature require either a) highly tedious individualized acoustical measurements or b) measuring individual anthropometric measurements with the help of 3D scans or c) perceptual experiments. All these techniques

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

are highly expensive in terms of the resources required and the measurements have to be carried out for every individual that is highly impractical.

An extensive distance-dependent HRTF database was created to compare and validate the modeled HRTFs using the frontal projection headphones. The HRTF database known as the DSP Lab HRTF database is available for download for free at <http://eeeweb.nyu.edu.sg/DSPLab/DspLabHRTF/>.

The main challenge of individualization that lingers today is to identify an individualization technique, which does not require any tedious subject-dependent acoustical measurements. What is ideally required is a plug and play type system that can allow the listeners to customize their HRTFs and enjoy immersive sound on their personal devices. The development of the frontal projection headphones is a positive step towards this goal. The frontal projection headphones project the sound from the front and automatically embed the listener's pinna cues at the eardrum spectrum.

The first application of the frontal projection headphones playback shown in this thesis is its use in individualizing the non-individualized binaural synthesis (Chapter 4). It was observed that the frontal projection, generated pinna spectral cues, that are close to the measured individualized HRTFs. Due to the frontal projection of the non-individualized binaural sound, the highly idiosyncratic pinna cues are superposed with the non-individual spectral cues at the eardrum. Subjective experiments indicated that the presence of individualized pinna cues improved the frontal perception of the 3D sound. However, there was no such improvement in the playback using a conventional side projection headphones. In fact, the side projection headphones displayed a rear-perceptual bias as most of the virtual sounds were localized behind the head.

Equalization of headphones is again a challenging aspect of binaural playback process. Headphones, due to its non-flat transfer function, color the input sound spectrum and degrades the spatial perception. Conventional Type-1 equalization is a subject-dependent

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

equalization technique and thus, a generic equalization filter that can work well for everyone cannot be achieved. A subject-independent robust equalization technique is desired. In this thesis, a subject-independent robust equalization technique, known as the Type-2 equalization technique is developed for the frontal projection headphones playback technique. The Type-1 equalization technique is calculated using the free-field response of the headphone. Thus, it is enough that the Type-2 equalization filter is calculated only once for frontal projection playback on any listener. The front-back confusions with the Type-2 equalization technique reduced almost by 50 %, when compared with the non-individualized binaural playback using side projection headphones.

The frontal projection headphones playback along with the Type-2 equalization has been found to be an useful individualization technique for non-individualized binaural synthesis. However, the presence of non-individualized spectral cues along with the individualized pinna cues the eardrum, leads to timbral coloration of the sound scene. To investigate this, a set of spectral flattening experiments were carried out, where the high-frequency spectral cues in the non-individualized HRTF were removed in different octave BWs (2-oct, 1-oct high, 1-oct mid, 1-oct low, 1/2-oct high, 1/2-oct mid high, 1/2-oct mid low, 1/2-oct low, Baseline) by spectral flattening. The spectral flattened HRTFs using frontal headphone projection were compared with baseline conditions using both frontal as well as side projection headphones. Subjective experiments proved that subjects displayed similar localization performance in the 2-octave BW spectral flattened condition when compared to the baseline condition. The rate of front-back confusions with the frontal projection headphones were found to be 20%, while the front-back confusions were much higher (70%) with the side projection headphones. These set of results proved that the non-individualized high frequency spectral cues can be removed in the frontal projection playback individualization technique. The individualized spectral cues created by the frontal projection playback are self-sufficient in delivering a realistic immersive

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

experience.

The frontal projection individualization technique, which was validated in Chapter 4, is then extended to model the individualized distance-dependent HRTFs (Chapter 6). HRTFs in the far-field in general do not vary much with distance. However, HRTFs in the near-field show high variation with distance. Moreover, the need of an acoustic point source arises to avoid scattering between the source and the head during measurements. Measuring individualized HRTFs in the near-field due to the above reasons becomes even more tedious. Thus, the next challenge is to model these highly individualized HRTFs in the proximal region. In Chapter 6, the frontal projection headphone playback technique along with a cascade of signal processing blocks is used to individualize these distance-dependent HRTFs. Before the modeling process, two important questions were needed to be addressed:

- Does headphone equalization affect distance perception?
- What are the important perceptual cues that affect distance perception?

To answer the first question, the effect of equalization on distance perception with different headphones (closed, open-back, and open) were investigated with the help of detailed perceptual experiments (Chapter 5). It was found that the equalization of the headphones is critical for accurate distance perception. Imperfect equalization often led to either IHL or poor estimation of the distance. Moreover, the high-frequency spectral cues did not have any perceptual effect on the the distance localization performance. The low-frequency equalization of the headphone is thus, extremely critical, when accurate distance-perception is desired. It was also found that both non-individualized and individualized equalization had a similar effect on distance perception. This is precisely due to the lack of dependence of distance perception on the idiosyncratic high-frequency cues. The open, and open-back headphones had lesser number of IHL compared to the closed-back headphones. The frontal projection headphones is an open headphone and thus,

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

the low-frequency response of the frontal projection headphone is extremely poor. This may severely distort the distance perception. Therefore, it is necessary to compensate the low-frequency response of the frontal projection headphones to avoid degradation of distance perception.

In the first stage of the modeling process, the important cues that affect distance perception in the near-field were identified. The auditory parallax and the ILD cues were mainly investigated for their role in distance perception. Detailed subjective experiments were carried out in which, the stimuli were generated using synthesized auditory parallax and ILD cues using both individualized and non-individualized HRTFs. Both the objective and subjective results proved that the role of auditory parallax in the presence of ILD cues was insignificant. Thus, the ILD cues were found to be the most prominent cues for distance perception in anechoic conditions. Thus, the auditory parallax cues were not considered in the modeling process. The distance-dependent ILD effects were modeled by the spherical head model. In addition, the Type-2 equalization filter, and a low-compensation filter to compensate for the low-frequency response of the frontal projection headphone were needed. The modeled distance-dependent HRTFs modeled using the frontal projection headphones were compared with the measured distance-dependent individualized HRTFs and validated with both objective and subjective experiments. To model the rear directional HRTFs, rear-directional filters were developed based on the spectral differences between the frontal and rear directional HRTFs of all the subjects in the CIPIC database. The RSD scores for the ipsilateral HRTF modeling in frontal directions was always found to be less than 4 dB, while the RSD score increased for rear directions. Furthermore, for the contralateral ear, the RSD score increased even more due to the presence of ipsilateral pinna cues in the contralateral ear spectrum. However, subjective experiments suggested that subjects could perceive both frontal and rear directions well with the modeled HRTFs.

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

Localization in the median plane is severely distorted when non-individualized HRTFs are used. The individualization of HRTFs using the frontal projection playback technique can also be extended to the median plane. The most critical cues for elevation perception are the monaural pinna spectral cues in the form of notches ($N1$, $N2$) and peaks ($P1$) in the HRTF spectrum. Study of elevation cues suggested that these individualized pinna notches $N1$ and $N2$ shifted in their center frequencies monotonically with increasing elevation angle. In Chapter 7, individualized HRTFs were modeled in the median plane using the frontal projection headphones. In this modeling process, it is first required to measure the frontal projection headphones response for the listener whose individualized HRTFs are to be modeled. Using the Raykar's algorithm [8], the position of the notches $N1$, $N2$, and peak $P1$ are extracted. Similarly, the positions of the notches $N1$, $N2$, and peak $P1$ and the variation of notch positions with elevation is studied for the CIPIC database. The average of the notch-shift variation with elevation across all the subjects is then utilized to shift the notch position in the measured frontal projection response. These shifts in notch positions with elevation simulates the elevation perception in the median plane.

Objective measures such as RSD and NFD were calculated for the measured and the modeled HRTFs for all the elevation angles. RSD scores for the modeled HRTFs were found to be around 2 dB for frontal angles above the head. However, the RSD scores increased to 4 dB, when the source position moved below the head. The RSD scores were higher (3-3.5 dB) even for rear directions due to the approximation by the use of an average rear-directional filter. NFD is a measure of the notch frequency distance (in octaves) for the two important notches $N1$ and $N2$ for the modeled and the measured HRTFs. It was found that the total residual error across all the directions was between 0.02-octave to 0.06-octave BW. This variation can be considered to be a very small value as most of the notch features in the HRTF are 1-octave BW wide. Moreover, subjective

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

experiments were conducted for the modeled median plane HRTFs using both frontal and side projection headphones. Perceptual experiments indicated that the modeled HRTFs in the median plane is a good approximation to the true individualized HRTFs. These set of results could be easily extended to other sagittal planes using a frequency-independent ITD.

We have seen that the frontal projection headphone playback technique can be used to model the individualized HRTFs across the azimuth, distance, and elevation. Though this ensures accurate spatial reproduction, it does not deliver a completely natural sound as perceived in real life. A special 3D audio headphone is developed, which is a four emitter headphone containing both the frontal projection as well as the conventional side emitter headphone on each side. Direct playback of multimedia content (music, games, and movies) on conventional headphones results in a very unnatural perception. The multimedia stereo content is first separated into primary sources and the ambient signal using a primary-ambient extraction (PAE) algorithm. The extracted primary cues are played back through the frontal emitters, while the ambience signal is routed to both the conventional side emitters as well as the frontal projection emitters. Type-2 equalization is applied to the frontal projection emitter and a Type-1 diffused-field equalization is applied for the side projection emitter. Listening experiments conducted showed that subjects could indeed perceive the sounds to truly natural and immersive. The MOS scores for the natural sound rendering system was always higher than 70 for all the measures, while the MOS scores were around 55-60 for the conventional stereo system. Overall preference studies indicated that 61% of the subjects preferred the natural sound rendering system, while 33% of the subjects preferred conventional stereo system. A small percentage of the listeners were unsure about their preference between both the systems.

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

9.2 Future works

The work carried out in this thesis has enormous scope, especially with the increase in devices supporting virtual reality. Headphones offering private listening space are ideally suited for such applications. Several refinements can be done to the proposed individualization scheme using frontal projection headphones. The Type-2 equalization filter is currently obtained by first measuring the free-field response of the headphone in an anechoic test box and then taking an inverse of the free-field response. This might not be entirely accurate and may differ from the boundary conditions, when a listener wears the frontal projection headphone. In order to ensure accurate measurements, experiments could be conducted using different types of materials covering the ear. The position of the microphones also play an important role during the measurement of the free-field response. All these factors are critical and can be investigated in future.

In this thesis, the timbral coloration due to the frontal projection headphones were not reported. However, it can be valuable to know the amount of timbral coloration introduced by the frontal projection headphones. Extensive timbral coloration experiments (objective and subjective) can be carried out to further validate the modeled HRTFs using frontal projection responses. An important extension of the individualization scheme would be to integrate the frontal projection system with head-tracking. Head-movements give important interaural cues in real life to resolve the ambiguity in the cone of confusion region. It is hypothesized that additional head-tracking will only improve the immersiveness with frontal projection headphones. The 3D audio headphone that is currently used for virtual reality can be extended to augmented reality with the help of additional hear-through microphones. In addition, a more comprehensive theoretical modeling of the frontal projection headphones and the multi-emitter 3D audio headphone is needed in future. Furthermore, to validate the 3D audio reproduction completely, the

CHAPTER 9. CONCLUSIONS AND FUTURE WORKS

spatial sampling of the sound field in the earcup and compensation of near field effects have to be well investigated. It would also be interesting to compare the natural sound rendering system with natural sound perception instead of conventional stereo system. The individualization techniques can also be integrated with several commercial systems like the DTS Headphone:X [297], Microsoft Hololens [298],Oculus rift [299], Dolby 7.1 headphones [300], etc.

Author's Publications

Journals

- [1] Sunder, K., Tan, E.-L., and Gan, W.-S. "Individualization of binaural synthesis using frontal projection headphones," *J.Audio Eng.Soc.*, 61, 989-1000, 2013.
- [2] Sunder, K., Jianjun, H., Tan, E.-L., and Gan, W.-S. "Natural sound rendering for headphones: Integration of signal processing techniques," *IEEE Signal Processing Magazine*, 32, 100-113, 2015.
- [3] Sunder K., Gan, W.-S, and Tan, E.-L., "Modeling distance-dependent individual head-related transfer functions in the horizontal plane using frontal projection headphones" *J. Acoust. Soc. Am.*, 138, 150-171, 2015.
- [4] Sunder K., Gan, W.-S, "Individualization of head-related transfer functions in the median plane using frontal projection headphones" Submitted to the *J.Audio Eng.Soc.*,

Conferences

- [5] Sunder, K., Tan, E.-L., and Gan, W.-S. "On the study of frontal-emitter headphone to improve 3-D Audio playback," *Proc. 133rd AES Convention*, San Francisco, 2012.
- [6] Sunder, K., Tan, E.-L., and Gan, W.-S. "Effect of headphone equalization on auditory distance perception," *Proc.137th AES Convention*, Los Angeles, 2014.
- [7] Ko Ko, P., Sunder, K., and Gan, W.-S. "Android framework implementation of 3D Audio with range control," *Proc.137th AES Convention*, Los Angeles, 2014.

Databases

- [8] Sunder, K. (2014). "DSPLab HRTF Database."
<http://eeeweba.ntu.edu.sg/DSPLab/DspLabHRTF/>

Bibliography

- [1] B and K, “HATS - Type 4128-C Dummy Head Documentation,” <http://www.bksv.com/products/transducers/ear-simulators/head-and-torso/hats-type-4128c.aspx>.
- [2] B. Gardner and K. Martin, “HRTF measurements of a KEMAR dummy-head microphone,” *MIT Media Laboratory Perceptual Computing Technical Report*, vol. 280, no. 280, pp. 1–7, 1994.
- [3] Neumann, “Neumann KU100 Dummy Head Documentation,” https://www.neumann.com/?lang=en&id=current_microphones&cid=ku100_description.
- [4] Head-Acoustics, “HMSIV Aachen Head Documentation,” https://www.head-acoustics.de/eng/nvh_hms_IV.htm.
- [5] 01dB METRAVIB, “MK2B, Acoustic Artificial Head Documentation.”
- [6] F. Christensen, C. B. Jensen, and H. Mller, “The design of valdemar-an artificial head for binaural recording purposes,” in *Proc. 109th Audio Engineering Society Convention, Los Angeles*, 2000.
- [7] V. Larcher, J.-M. Jot, and G. Vandernoot, “Equalization methods in binaural technology,” in *Proc. 105th Audio Engineering Society Convention, San Francisco*, 1998.

BIBLIOGRAPHY

- [8] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 364–374, 2005.
- [9] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1829–1834, 1974.
- [10] F. Rumsey, *Spatial audio*. Focal Press, 2001.
- [11] J. Blauert, "Spatial hearing revised edition," *Cambridge, MA: Massachusetts Institute of Technology*, 1997.
- [12] D. R. Begault, *3-D sound for virtual reality and multimedia*. Cambridge, MA: AP Professional, 2000.
- [13] ITU-R, "Multi-channel stereophonic sound system with or without accompanying picture," 1993.
- [14] T. Holman, *Surround sound: up and running*. Taylor & Francis, 2008.
- [15] K. Hamasaki, K. Hiyama, and R. Okumura, "The 22.2 multichannel sound system and its application," in *Proc. 118th Audio Engineering Society Convention, Barcelona*, 2005.
- [16] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.
- [17] J. Braasch, *Modelling of Binaural Hearing*. Springer Berlin Heidelberg, 2005, ch. 4, pp. 75–108.
- [18] H. F. Olson, *Modern Sound Reproduction*. Van Nostrand Reinhold, 1972.
- [19] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc.*, vol. 33, no. 11, pp. 859–871, 1985.

BIBLIOGRAPHY

- [20] A. J. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, vol. 36, no. 12, pp. 977–995, 1988.
- [21] R. Ranjan and W. Gan, "Wave field synthesis: The future of spatial audio," *Potentials, IEEE*, vol. 32, no. 2, pp. 17–23, March 2013.
- [22] W. Gardner, "Transaural 3-d audio," Tech. Rep., 1995.
- [23] W. Gardner, "3-d audio using loudspeakers," Ph.D. dissertation, 1997.
- [24] B. S. Atal and M. Schroeder, "Apparent sound source translator," Feb 22, 1966 1966.
- [25] D. H. Cooper and J. L. Bauck, "Prospects for transaural recording," *J. Audio Eng. Soc.*, vol. 37, no. 1/2, pp. 3–19, 1989.
- [26] O. Kirkeby, P. A. Nelson, and H. Hamada, "The'stereo dipole': A virtual source imaging system using two closely spaced loudspeakers," *J. Audio Eng. Soc.*, vol. 46, no. 5, pp. 387–395, 1998.
- [27] A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 1071–1074, 2000.
- [28] S.-M. Kim and W. Choi, "On the externalization of virtual sound images in headphone reproduction: A wiener filter approach," *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3657–3665, 2005.
- [29] K. Sunder, E.-L. Tan, and W.-S. Gan, "Effect of headphone equalization on auditory distance perception," in *Proc. 137th Audio Engineering Society Convention, Los Angeles*, 2014.

BIBLIOGRAPHY

- [30] P. Ko Ko, K. Sunder, and W.-S. Gan, "Android framework implementation of 3d audio with range control," in *Proc. 137th Audio Engineering Society Convention, Los Angeles*, 2014.
- [31] K. Sunder, "DSPLab HRTF Database," <http://eeeweba.ntu.edu.sg/DSPLab/DspLabHRTF/>, 2014.
- [32] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, and H. Ziegelwanger, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Proc. 134th Audio Engineering Society Convention, Rome*, 2013.
- [33] K. Sunder and W. S. Gan, "Modeling individualized hrtfs in the median plane using frontal projection headphones," *J.Audio Eng.Soc.*, vol. Submitted, 2015.
- [34] W. Gan and E. Tan, "Listening device and accompanying signal processing method," Patent PCT/SG2012/000 116, 2012.
- [35] K. Sunder, H. Jianjun, T. Ee Leng, and G. Woon-Seng, "Natural sound rendering for headphones: Integration of signal processing techniques," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 100–113, 2015.
- [36] J. He, E.-L. Tan, and W.-S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 505–517, 2014.
- [37] C. Frauenberger and M. Noistering, "3d audio interfaces for the blind," in *Proc. 9th International Conference on Auditory Display (ICAD2003)*.
- [38] S. Goose, C. Moller, and Iler, "A 3d audio only interactive web browser: using spatialization to convey hypermedia document structure," in *Proc. 7th ACM international conference on Multimedia (Part 1), Orlando*, 1999, pp. 363–371.

BIBLIOGRAPHY

- [39] C. Papadopoulos, C. Kyriakakis, A. Sawchuk, and X. He, "Cyberseer: 3d audio-visual immersion for network security and management," in *Proc. of the 2004 ACM workshop on Visualization and data mining for computer security, Washington DC*, 2004, pp. 90–98.
- [40] R. S. Bolia, W. R. D'Angelo, and R. L. McKinley, "Aurally aided visual search in three-dimensional space," *J. Hum. Fac. Erg. Soc. P*, vol. 41, no. 4, pp. 664–669, 1999.
- [41] O. Carlander, L. Eriksson, and P.-A. Oskarsson, *Handling Uni- and Multimodal Threat Cueing with Simultaneous Radio Calls in a Combat Vehicle Setting*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4555, ch. 31, pp. 293–302.
- [42] R. Nicol, *Binaural Technology*, ser. Audio Engineering Society Monograph. AES, 2010.
- [43] D. Brungart and W. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, p. 1465, 1999.
- [44] F. L. Wightman and D. J. Kistler, "Headphone simulation of freefield listening. i: Stimulus synthesis," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 858–867, 1989.
- [45] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *J. Acoust. Soc. Am.*, vol. 61, no. 6, pp. 1567–1576, 1977.
- [46] S. Carlile and D. Pralong, "The locationdependent nature of perceptually salient features of the human headrelated transfer functions," *J. Acoust. Soc. Am.*, vol. 95, no. 6, pp. 3445–3459, 1994.

BIBLIOGRAPHY

- [47] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217, 1995.
- [48] W. G. Gardner and K. D. Martin, "Hrtf measurements of a kemar," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1995.
- [49] E. Shaw, "Transformation of sound pressure level from the free field to the eardrum in the horizontal plane," *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1848–1861, 1974.
- [50] F. Firestone, "The phase difference and amplitude ratio at the ears due to a source of pure tone," *J. Acoust. Soc. Am.*, vol. 2, no. 2, pp. 260–270, 1930.
- [51] J. Middlebrooks, J. Makous, and D. Green, "Directional sensitivity of sound pressure levels in the human ear canal," *J. Acoust. Soc. Am.*, vol. 86, no. 1, pp. 89–108, 1989.
- [52] P. M. Morse and K. Ingard, "Uno: Theoretical acoustics," 1968.
- [53] M. Vorländer, "Past, present and future of dummy heads," *Proc. Acústica, Guimarães, Portugal*, 2004.
- [54] M. Morimoto and Y. Ando, "On the simulation of sound localization," *J. Acoust. Soc. Jp.*, vol. 1, no. 3, pp. 167–174, 1980.
- [55] F. Wightman and D. Kistler, "Headphone simulation of freefield listening. ii: Psychophysical validation," *J. Acoust. Soc. Am.*, vol. 85, no. 3, pp. 868–878, 1989.
- [56] A. W. Bronkhorst, "Localization of real and virtual sound sources," *J. Acoust. Soc. Am.*, vol. 98, no. 5, pp. 2542–2553, 1995.

BIBLIOGRAPHY

- [57] F. Wightman and D. Kistler, "Measurement and validation of human hrtfs for use in hearing research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 429–439, 2005.
- [58] F. M. Wiener and D. A. Ross, "The pressure distribution in the auditory canal in a progressive sound field," *J. Acoust. Soc. Am.*, vol. 18, no. 2, pp. 401–408, 1946.
- [59] F. M. Wiener, "On the diffraction of a progressive sound wave by the human head," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 143–146, 1947.
- [60] E. Shaw, "Ear canal pressure generated by a free sound field," *J. Acoust. Soc. Am.*, vol. 39, no. 3, pp. 465–470, 1966.
- [61] M. Burkhard and R. Sachs, "Anthropometric manikin for acoustic research," *J. Acoust. Soc. Am.*, vol. 58, no. 1, pp. 214–222, 1975.
- [62] C. Searle, L. Braida, D. Cuddy, and M. Davis, "Binaural pinna disparity: another auditory localization cue," *J. Acoust. Soc. Am.*, vol. 57, no. 2, pp. 448–455, 1975.
- [63] E. Shaw and M. Vaillancourt, "Transformation of sound pressure level from the free field to the eardrum presented in numerical form," *J. Acoust. Soc. Am.*, vol. 78, no. 3, pp. 1120–1123, 1985.
- [64] D. Zotkin, R. Duraiswami, E. Grassi, and N. Gumerov, "Fast head-related transfer function measurement via reciprocity," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2202–2215, 2006.
- [65] K. Fukudome, T. Suetsugu, T. Ueshin, R. Idegami, and K. Takeya, "The fast measurement of head related impulse responses for all azimuthal directions using the continuous measurement method with a servo-swiveled chair," *Applied Acoustics*, vol. 68, no. 8, pp. 864–884, 2007.

BIBLIOGRAPHY

- [66] G. Enzner, "Analysis and optimal control of lms-type adaptive filtering for continuous-azimuth acquisition of head related impulse responses," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'08)*, pp. 393–396.
- [67] G. Enzner, "3d-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, pp. 325–328.
- [68] T. Ajdler, L. Sbaiz, and M. Vetterli, "Dynamic measurement of room impulse responses using a moving microphone," *J. Acoust. Soc. Amer*, vol. 122, no. 3, pp. 1636–1645, 2007.
- [69] IRCAM, "Listen hrtf database," <http://recherche.ircam.fr/equipes/salles/listen/index.html>, 2003.
- [70] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The cipic hrtf database," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01)*. IEEE, pp. 99–102.
- [71] Tohoku, "Hrtf database," <http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf/>, 2001.
- [72] T. Nishino, Y. Nakai, K. Takeda, and F. Itakura, "Estimating head related transfer function using multiple regression analysis," *IEICE Trans. A*, vol. 84, pp. 260–268, 2001.
- [73] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of hrtfs on the horizontal plane using physical features," *Applied Acoustics*, vol. 68, no. 8, pp. 897–908, 2007.
- [74] A. A. of Sciences, "Hrtf database," <http://www.kfs.oeaw.ac.at/index.php?lang=en>.

BIBLIOGRAPHY

- [75] P. Majdak, M. J. Goupell, and B. Laback, “3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training,” *Atten Percept Psychophys*, vol. 72, no. 2, pp. 454–469, 2010, 20139459[pmid] *Atten Percept Psychophys*.
- [76] TU-Berlin, “HRTF database,” <http://audio.qu.tu-berlin.de/?p=641>.
- [77] Oldenburg-University, “HRTF database,” <http://medi.uni-oldenburg.de/hrir/html/documentation.html>.
- [78] K. Genuit and N. Xiang, “Measurements of artificial head transfer functions for auralization and virtual auditory environment,” in *Proc. 15th International Congress on Acoustics*, vol. 2, pp. 469–472.
- [79] K. Riederer, “Head-related transfer function measurements,” Ph.D. dissertation, 1998.
- [80] J. Blauert, M. Brueggen, K. Hartung, A. W. Bronkhorst, R. Drullmann, G. Reynaud, L. Pellieux, W. Krebber, and R. Sottek, “The audis catalog of human hrtfs,” in *Proc. 16th International Congress on Acoustics*, 1998, pp. 2901–2902.
- [81] B. P. Bovbjerg, F. Christensen, P. Minnaar, and X. Chen, “Measuring the head-related transfer functions of an artificial head with a high-directional resolution,” in *Proc. 109th Audio Engineering Society Convention, New York*, 2000.
- [82] S. Takane, D. Arai, T. Miyajima, K. Watanabe, Y. Suzuki, and T. Sone, “A database of head-related transfer functions in whole directions on upper hemisphere,” *Acoustical science and technology*, vol. 23, no. 3, pp. 160–162, 2002.
- [83] E. Grassi, J. Tulsi, and S. Shamma, “Measurement of head-related transfer functions based on the empirical transfer function estimate,” in *Proc. 9th International Conference on Auditory Display (ICAD2003)*, pp. 119–121.

BIBLIOGRAPHY

- [84] B. Xie, X. Zhong, D. Rao, and Z. Liang, "Head-related transfer function database and its analyses," *Science in China Series G: Physics, Mechanics and Astronomy*, vol. 50, no. 3, pp. 267–280, 2007.
- [85] G. Yu, Y. Liu, B. Xie, and Q. Zhong, "Fast measurement system and super high directional resolution head-related transfer function database," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 3304–3304, 2012.
- [86] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300–321, 1995.
- [87] L. Rayleigh and A. Lodge, "On the acoustic shadow of a sphere. with an appendix, giving the values of legendre's functions from p_0 to p_{20} at intervals of 5 degrees," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 203, no. 359–371, pp. 87–110, 1904.
- [88] D. Leakey, B. M. Sayers, and C. Cherry, "Binaural fusion of low-and high-frequency sounds," *J. Acoust. Soc. Am.*, vol. 30, pp. 222–223, 1958.
- [89] G. B. Henning, "Detectability of interaural delay in highfrequency complex waveforms," *J. Acoust. Soc. Am.*, vol. 55, no. 1, pp. 84–90, 1974.
- [90] C. Trahiotis and L. R. Bernstein, "Lateralization of bands of noise and sinusoidally amplitude-modulated tones: Effects of spectral locus and bandwidth," *J. Acoust. Soc. Am.*, vol. 79, no. 6, pp. 1950–1957, 1986.
- [91] B. M. Sayers, "Acoustic image lateralization judgments with binaural tones," *J. Acoust. Soc. Am.*, vol. 36, no. 5, pp. 923–926, 1964.
- [92] F. Toole and B. M. Sayers, "Lateralization judgments and the nature of binaural acoustic images," *J. Acoust. Soc. Am.*, vol. 37, no. 2, pp. 319–324, 1965.

BIBLIOGRAPHY

- [93] G. Von Békésy and E. G. Wever, *Experiments in hearing*. McGraw-Hill New York, 1960, vol. 8.
- [94] F. Wightman and D. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661, 1992.
- [95] S. Weinrich, "Improved externalization and frontal perception of headphone signals," in *Proc. 92nd Audio Engineering Society Convention, Vienna*, 1992.
- [96] B. Rayleigh, *The theory of sound*. Macmillan, 1896, vol. 2.
- [97] G. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Am.*, vol. 62, no. 1, pp. 157–167, 1977.
- [98] H. Wallach, "On sound localization," *J. Acoust. Soc. Am.*, vol. 10, no. 4, pp. 270–274, 1939.
- [99] A. Musicant and R. Butler, "The influence of pinnae-based spectral cues on sound localization," *J. Acoust. Soc. Am.*, vol. 75, no. 4, pp. 1195–1200, 1984.
- [100] S. Oldfield and S. Parker, "Acuity of sound localisation: A topography of auditory space: II. pinna cues absent," *Perception*, vol. 13, no. 5, pp. 601–617, 1984.
- [101] J. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Proc. 98th Audio Engineering Society Convention, New York*, 1995.
- [102] V. Larcher, "Techniques de spatialisation des sons pour la réalité virtuelle," Ph.D. dissertation, 2001.
- [103] R. Woodworth and H. Schlosberg, *Experimental psychology*. Holt, 1954.

BIBLIOGRAPHY

- [104] D. Brungart, "Near-field virtual audio displays," *Presence.Teleop.Virt*, vol. 11, no. 1, pp. 93–106, 2002.
- [105] R. Butler and K. Belendiuk, "Spectral cues utilized in the localization of sound in the median sagittal plane," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1264–1269, 1977.
- [106] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, "Localization using nonindividualized headrelated transfer functions," *J. Acoust. Soc. Am.*, vol. 94, p. 111, 1993.
- [107] F. Wightman and D. Kistler, "Multidimensional scaling analysis of head-related transfer functions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'93)*. IEEE, pp. 98–101.
- [108] D. Begault, E. Wenzel, A. Lee, and M. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, 2001.
- [109] J. Blauert, "Sound localization in the median plane," *Acta Acustica united with Acustica*, vol. 22, no. 4, pp. 205–213, 1969. [Online]. Available: <http://www.ingentaconnect.com/content/dav/aaua/1969/00000022/00000004/art00004>
- [110] P. Bloom, "Creating source elevation illusions by spectral manipulation," *J. Audio Eng. Soc.*, vol. 25, no. 9, pp. 560–565, 1977.
- [111] H. L. Han, "Measuring a dummy head in search of pinna cues," *J. Audio Eng. Soc.*, vol. 42, no. 1-2, pp. 15–37, 1994.
- [112] E. Macpherson and W. Center, "On the role of head-related transfer function spectral notches in the judgement of sound source elevation," *Auditory Display:*

BIBLIOGRAPHY

- Sonification, Audification, and Auditory Interface, SFI Studies in the Sciences of Complexity*, edited by G. Kramer and S. Smith (Addison-Wesley, Reading, MA), pp. 187–194, 1994.
- [113] E. Langendijk and A. Bronkhorst, “Contribution of spectral cues to human sound localization,” *J. Acoust. Soc. Am.*, vol. 112, no. 4, pp. 1583–1596, 2002.
- [114] D. Toledo, “Technical and perceptual issues on head-related transfer functions sets for use in binaural synthesis,” PhD Thesis, 2011.
- [115] M. Gardner and R. Gardner, “Problem of localization in the median plane: effect of pinnae cavity occlusion,” *J. Acoust. Soc. Am.*, vol. 53, no. 2, pp. 400–408, 1973.
- [116] E. Shaw and R. Teranishi, “Sound pressure generated in an external ear replica and real human ears by a nearby point source,” *J. Acoust. Soc. Am.*, vol. 44, no. 1, pp. 240–249, 1968.
- [117] A. J. Watkins, “Psychoacoustical aspects of synthesized vertical locale cues,” *J. Acoust. Soc. Am.*, vol. 63, no. 4, pp. 1152–1165, 1978.
- [118] C. Rodgers, “Pinna transformations and sound reproduction,” *J. Audio Eng. Soc.*, vol. 29, no. 4, pp. 226–234, 1981.
- [119] R. Humanski and R. Butler, “The contribution of the near and far ear toward localization of sound in the sagittal plane,” *J. Acoust. Soc. Am.*, vol. 83, no. 6, pp. 2300–2310, 1988.
- [120] J. Middlebrooks, “Narrowband sound localization related to external ear acoustics,” *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 2607–2624, 1992.
- [121] F. Asano, Y. Suzuki, and T. Sone, “Role of spectral cues in median plane localization,” *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 159–168, 1990.

BIBLIOGRAPHY

- [122] E. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3248–3259, 1996.
- [123] M. Itoh, K. Iida, and M. Morimoto, "Individual differences in directional bands in median plane localization," *Applied Acoustics*, vol. 68, no. 8, pp. 909–915, 2007.
- [124] S. Roffler and R. Butler, "Factors that influence the localization of sound in the vertical plane," *J. Acoust. Soc. Am.*, vol. 43, no. 6, pp. 1255–1259, 1968.
- [125] E. A. G. Shaw, *Acoustical Characteristics of the Outer Ear*. John Wiley & Sons, Inc., 2007, pp. 1325–1335.
- [126] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479, 2001.
- [127] S. K. Roffler and R. A. Butler, "Localization of tonal stimuli in the vertical plane," *J. Acoust. Soc. Am.*, vol. 43, no. 6, pp. 1260–1266, 1968.
- [128] R. Bcklein, "The audibility of frequency response irregularities," *J. Audio Eng. Soc.*, vol. 29, no. 3, pp. 126–131, 1962.
- [129] B. Moore, S. Oldfield, and G. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 khz," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 820–836, 1989.
- [130] M. B. Gardner, "Some monaural and binaural facets of median plane localization," *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1489–1495, 1973.
- [131] R. Butler and N. Planert, "The influence of stimulus bandwidth on localization of sound in space," *Atten Percept Psychophys*, vol. 19, no. 1, pp. 103–108, 1976.

BIBLIOGRAPHY

- [132] P. Hofman, J. Van Riswick, and A. Van Opstal, "Relearning sound localization with new ears," *Nature neuroscience*, vol. 1, no. 5, pp. 417–421, 1998.
- [133] H. Møller, M. Sørensen, C. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.
- [134] H. Møller, "Fundamentals of binaural technology," *Applied acoustics*, vol. 36, no. 3, pp. 171–218, 1992.
- [135] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *J. Acoust. Soc. Am.*, vol. 100, no. 6, pp. 3785–3793, 1996.
- [136] K. I. McAnally and R. L. Martin, "Variability in the headphone-to-ear-canal transfer function," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 263–266, 2002.
- [137] G. Theile, "On the standardization of the frequency response of high-quality studio headphones," *J. Audio Eng. Soc.*, vol. 34, no. 12, pp. 956–969, 1986.
- [138] D. Hammershøi and H. Møller, "Sound transmission to and within the human ear canal," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 408–427, 1996.
- [139] S. Olive, T. Welti, and E. McMullin, "Listener preferences for different headphone target response curves," in *Proc. 134th Audio Engineering Society Convention, Rome*, 2013.
- [140] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 218–232, 1995.
- [141] G. Lorho, "Subjective evaluation of headphone target frequency responses," in *Proc. 126th Audio Engineering Society Convention, Munich*, 2009.

BIBLIOGRAPHY

- [142] F. Fleischmann, A. Silzle, and J. Plogsties, "Identification and evaluation of target curves for headphones," in *Proc. 133rd Audio Engineering Society Convention, San Francisco*, 2012.
- [143] S. Olive and T. Welti, "The relationship between perception and measurement of headphone sound quality," in *Proc. 133rd Audio Engineering Society Convention, San Francisco*, 2012.
- [144] K. Sunder, E.-L. Tan, and W.-S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 989–1000, 2013.
- [145] R. L. Martin, K. I. McAnally, and M. A. Senova, "Free-field equivalent localization of virtual audio," *J. Audio Eng. Soc.*, vol. 49, no. 1/2, pp. 14–22, 2001.
- [146] M. Park, S. Choi, S. Kim, and K. Bae, "Improvement of front-back sound localization characteristics in headphone-based 3d sound generation," in *Proc. 7th International Conference on Advanced Communication Technology (ICACT 2005)*, vol. 1. IEEE, pp. 273–276.
- [147] F. Volk, F. Heinemann, and H. Fastl, "Externalization in binaural synthesis: effects of recording environment and measurement procedure," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 6419–6424, 2008.
- [148] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, p. 3678, 1996.
- [149] G. Plenge, "On the differences between localization and lateralization," *J. Acoust. Soc. Am.*, vol. 56, no. 3, pp. 944–951, 1974.

BIBLIOGRAPHY

- [150] N. Durlach, A. Rigopulos, X. Pang, W. Woods, A. Kulkarni, H. Colburn, and E. Wenzel, "On the externalization of auditory images," *Presence. Teleop. Virt.*, vol. 1, no. 2, pp. 251–257, 1992.
- [151] J. M. Loomis, C. Hebert, and J. G. Cicinelli, "Active localization of virtual sounds," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1757–1764, 1990.
- [152] E. M. Wenzel, "What perception implies about implementation of interactive virtual acoustic environments," in *Proc. 111th Audio Engineering Society Convention, New York*, 2001.
- [153] G. Wersnyi, "Effect of emulated head-tracking for reducing localization errors in virtual audio simulation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 247–252, 2009.
- [154] D. R. Begault, "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *J. Audio Eng. Soc.*, vol. 40, no. 11, pp. 895–904, 1992.
- [155] T. Brookes and C. Treble, "The effect of non-symmetrical left/right recording pinnae on the perceived externalisation of binaural recordings," in *Proc. 118th Audio Engineering Society Convention, Barcelona*, 2005.
- [156] C. I. Cheng, "Visualization, measurement, and interpolation of head-related transfer functions (hrtfs) with applications in electro-acoustic music," Ph.D. dissertation, 2001.
- [157] G. Wersnyi and A. Illnyi, "Test signal generation and accuracy of turntable control in a dummy-head measurement system," *J. Audio Eng. Soc.*, vol. 51, no. 3, pp. 150–155, 2003.

BIBLIOGRAPHY

- [158] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 55, no. 7/8, pp. 623–637, 2007.
- [159] A. Dobrucki, P. Plaskota, P. Pruchnicki, M. Pec, M. Bujacz, and P. Strumillo, "Measurement system for personalized head-related transfer functions and its verification by virtual source localization trials with visually impaired and sighted individuals," *J. Audio Eng. Soc.*, vol. 58, no. 9, pp. 724–738, 2010.
- [160] M. R. Schroeder, "Integrated impulse method measuring sound decay without using impulses," *J. Acoust. Soc. Am.*, vol. 66, no. 2, pp. 497–500, 1979.
- [161] P. A. N. Briggs and K. R. Godfrey, "Pseudorandom signals for the dynamic analysis of multivariable systems," *Electrical Engineers, Proceedings of the Institution of*, vol. 113, no. 7, pp. 1259–1267, 1966.
- [162] N. Ream, "Nonlinear identification using inverse-repeatm sequences," *Proc. of the Institution of Electrical Engineers*, vol. 117, no. 1, pp. 213–218, 1970.
- [163] C. Dunn and M. J. Hawksford, "Distortion immunity of mls-derived impulse response measurements," *J. Audio Eng. Soc.*, vol. 41, no. 5, pp. 314–335, 1993.
- [164] N. Aoshima, "Computergenerated pulse signal applied for sound measurement," *J. Acoust. Soc. Amer*, vol. 69, no. 5, pp. 1484–1488, 1981.
- [165] A. Farina and E. Ugolotti, "Subjective comparison between stereo dipole and 3d ambisonic surround systems for automotive applications," in *Proc. 16th Audio Engineering Society Conference on Spatial Sound Reproduction, Finland*, 1999.
- [166] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. 108th Audio Engineering Society Convention, Paris*, 2000.

BIBLIOGRAPHY

-
- [167] M. Hiipakka, T. Kinnari, and V. Pulkki, "Estimating head-related transfer functions of human subjects from pressure-velocity measurements," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4051–4061, 2012.
 - [168] J. Chan and C. Geisler, "Estimation of eardrum acoustic pressure and of ear canal length from remote points in the canal," *J. Acoust. Soc. Am.*, vol. 87, no. 3, pp. 1237–1247, 1990.
 - [169] S. Hosoe, T. Nishino, K. Itou, and K. Takeda, "Development of micro-dodecahedral loudspeaker for measuring head-related transfer functions in the proximal region," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, vol. 5. IEEE, pp. V–V.
 - [170] Y. Hayakawa, T. Nishino, and K. Takeda, "Development of small sound equipment with micro-dynamic-type loudspeaker for hrtf measurement," in *Proc. 19th. International Congress on Acoustics, Madrid*, 2007.
 - [171] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1124–1132, 2009.
 - [172] R. Duda and W. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058, 1998.
 - [173] P. T. Calamia and E. L. Hixson, "Measurement of the head-related transfer function at close range," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 3117–3117, 1997.
 - [174] Y. Guang-Zheng, X. Bo-Sun, and R. Dan, "Effect of sound source scattering on measurement of near-field head-related transfer functions," *Chinese Physics Letters*, vol. 25, no. 8, p. 2926, 2008.

BIBLIOGRAPHY

- [175] S. Mller and P. Massarani, "Transfer-function measurement with sweeps," *J.Audio Eng.Soc.*, vol. 49, no. 6, pp. 443–471, 2001.
- [176] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*. Prentice-hall Englewood Cliffs, 1989, vol. 2.
- [177] N. Damera-Venkata, B. L. Evans, and S. R. McCaslin, "Design of optimal minimum-phase digital fir filters using discrete hilbert transforms," *IEEE Trans. on Signal processing*, vol. 48, no. 5, pp. 1491–1495, 2000.
- [178] J. Sandvad and D. Hammershoi, "Binaural auralization, comparison of fir and iir filter representation of hirs," in *Proc. 96th Audio Engineering Society Convention, Amsterdam*, 1994.
- [179] P. Minnaar, J. Plogsties, and F. Christensen, "Directional resolution of head-related transfer functions required in binaural synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 919–929, 2005.
- [180] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and subjective evaluation of head-related transfer function filter design," *J. Audio Eng. Soc.*, vol. 47, no. 4, pp. 218–239, 1999.
- [181] A. Kulkarni, S. Isabelle, and H. Colburn, "On the minimum-phase approximation of head-related transfer functions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'95)*. IEEE, pp. 84–87.
- [182] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 6, no. 5, pp. 476–488, 1998.
- [183] L. Savioja, J. Huopaniemi, T. Lokki, and R. Vninen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, 1999.

BIBLIOGRAPHY

- [184] R. O. Duda, C. Avendano, and V. R. Algazi, *An adaptable ellipsoidal head model for the interaural time difference*, ser. International Conference on Acoustics Speech and Signal Processing (ICASSP). IEEE, 1999, pp. 965–968.
- [185] J. Pernaux, “Spatialisation du son par les techniques binaurales: application aux services de telecommunications,” Ph.D. dissertation, 2003.
- [186] Z. Xiao-Li and X. Bo-Sun, “A novel model of interaural time difference based on spatial fourier analysis,” *Chinese Physics Letters*, vol. 24, no. 5, p. 1313, 2007.
- [187] K. Watanabe, K. Ozawa, Y. Iwaya, Y. Suzuki, and K. Aso, “Estimation of interaural level difference based on anthropometry and its effect on sound localization,” *J. Acoust. Soc. Am.*, vol. 122, no. 5, pp. 2832–2841, 2007.
- [188] S. Xu, Z. Li, and G. Salvendy, *Individualization of head-related transfer function for three-dimensional virtual auditory display: a review*. Springer, 2007, pp. 397–407.
- [189] D. S. Brungart, G. Romigh, and B. D. Simpson, “Rapid collection of hrtfs and comparison to free-field listening,” in *Proc. International Workshop on the Principles and Applications of Spatial Hearing (IWPASH’09)*.
- [190] Nagoya-University, “HRTF database,” <http://www.ciair.coe.nagoya-u.ac.jp/db/>.
- [191] B. U. Seeber and H. Fastl, “Subjective selection of non-individual head-related transfer functions,” in *Proc. 9th International Conference on Auditory Display (ICAD 2003)*, vol. 3, pp. 259–262.
- [192] W. L. Martens, “Rapid psychophysical calibration using bisection scaling for individualized control of source elevation in auditory display,” in *Proc. 8th International Conference on Auditory Display (ICAD 2002)*.

BIBLIOGRAPHY

-
- [193] J. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1480–1492, 1999.
 - [194] C.-J. Tan and W.-S. Gan, "User-defined spectral manipulation of hrtf for improved localisation in 3 d sound systems," *Electronics letters*, vol. 34, no. 25, pp. 2387–2389, 1998.
 - [195] A. Silzle, "Selection and tuning of hrtfs," in *Proc. 112th Audio Engineering Society Convention, Munich*, 2002.
 - [196] K. H. Shin and P. Youngjin, "Enhanced vertical perception through head-related impulse response customization based on pinna response tuning in the median plane," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 91, no. 1, pp. 345–356, 2008.
 - [197] P. Runkle, A. Yendiki, and G. Wakefield, "Active sensory tuning for immersive spatialized audio," in *Proc. International Conference on Auditory Display (ICAD'00)*.
 - [198] K. J. Fink and L. Ray, "Tuning principal component weights to individualize hrtfs," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'12)*. IEEE, pp. 389–392.
 - [199] Y. Hur, S.-P. Lee, Y.-c. Park, and D.-h. Youn, "Efficient individualization of hrtf using critical-band based spectral cues control," in *Proc. 124th Audio Engineering Society Convention, Amsterdam*, 2008.
 - [200] S. Shimada, N. Hayashi, and S. Hayashi, "A clustering method for sound localization transfer functions," *J. Audio Eng. Soc.*, vol. 42, no. 7/8, pp. 577–584, 1994.
 - [201] A. Bondu, S. Busson, V. Lemaire, and R. Nicol, "Looking for a relevant similarity criterion for hrtf clustering: a comparative study," in *Proc. 120th Audio Engineering Society Convention, Paris*, 2006.

BIBLIOGRAPHY

- [202] M. Dellepiane, N. Pietroni, N. Tsingos, M. Asselot, and R. Scopigno, "Reconstructing head models from photographs for individualized 3daudio processing," in *Computer Graphics Forum*, vol. 27. Wiley Online Library, pp. 1719–1727.
- [203] W. W. Hugeng and D. Gunawan, "Improved method for individualization of head-related transfer functions on horizontal plane using reduced number of anthropometric measurements," *J. Telecomm*, vol. 2, no. 2, pp. 31–41, 2010.
- [204] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in *Proc. International Conference on Computational Intelligence and Security*, vol. 2. IEEE, pp. 1829–1832.
- [205] M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, "Hrtf customization using multiway array analysis," in *Proc. 18th European Signal Processing Conference (EUSIPCO'10)*, pp. 229–233.
- [206] Q.-h. Huang and Y. Fang, "Modeling personalized head-related impulse response using support vector regression," *Journal of Shanghai University (English Edition)*, vol. 13, no. 6, pp. 428–432, 2009.
- [207] H. Hu, L. Zhou, H. Ma, and Z. Wu, "Hrtf personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, 2008.
- [208] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "Hrtf personalization using anthropometric measurements," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*. IEEE, pp. 157–160.
- [209] K. Sunder, E.-L. Tan, and W.-S. Gan, "On the study of frontal-emitter headphone to improve 3-d audio playback," in *Proc. 133rd Audio Engineering Society Convention, San Francisco*, 2012.

BIBLIOGRAPHY

- [210] D. Kistler and F. Wightman, "A model of headrelated transfer functions based on principal components analysis and minimumphase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, 1992.
- [211] S. Rodriguez and M. Ramirez, "Linear relationships between spectral characteristics and anthropometry of the external ear," in *Proc. 11th Meeting of the International Conference on Auditory Display (ICAD05)*, vol. 11.
- [212] S. Spagnol, D. Rocchesso, M. Geronazzo, and F. Avanzini, "Automatic extraction of pinna edges for binaural audio customization," in *IEEE 15th International Workshop on Multimedia Signal Processing (MMSP2013)*. IEEE, pp. 301–306.
- [213] R. O. Duda, V. R. Algazi, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *Proc. 113th Audio Engineering Society Convention, Los Angeles*, 2002.
- [214] O. Kirkeby, P. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 6, no. 2, pp. 189–194, 1998.
- [215] D. C. Montgomery, *Design and analysis of experiments*. John Wiley & Sons, 2008.
- [216] M. J. Evans, "Obtaining accurate responses in directional listening tests," in *Proc. 104th Audio Engineering Society Convention, Amsterdam*, 1998.
- [217] G. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, vol. 19, no. 4, pp. 71–87, 1995.
- [218] C. J. Tan and W. S. Gan, "Direct concha excitation for the introduction of individualized hearing cues," *J. Audio Eng. Soc.*, vol. 48, no. 7-8, pp. 642–653, 2000.

BIBLIOGRAPHY

- [219] D. S. Brungart, "Auditory localization of nearby sources. iii. stimulus effects," *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. 3589–3602, 1999.
- [220] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.
- [221] K. Sunder, W. S. Gan, and E.-L. Tan, "Modeling range-dependent individual head-related transfer functions in the horizontal plane using frontal projection headphones," *J. Acoust. Soc. Am.*, vol. To be published in July/August issue, 2015.
- [222] P. Zahorik, "Auditory display of sound source distance," in *Proceedings of 8th International Conference on Auditory Display*, pp. 1–7.
- [223] G. von Bksy, "The moon illusion and similar auditory phenomena," *American journal of psychology*, vol. 62, no. 4, pp. 540–552, 1949.
- [224] W. Simpson and L. D. Stanton, "Head movement does not facilitate perception of the distance of a source of sound," *American journal of psychology*, pp. 151–159, 1973.
- [225] H.-Y. Kim, Y. Suzuki, S. Takane, and T. Sone, "Control of auditory distance perception based on the auditory parallax model," *Applied Acoustics*, vol. 62, no. 3, pp. 245–270, 2001.
- [226] S. Stevens and M. Guirao, "Loudness, reciprocity, and partition scales," *J. Acoust. Soc. Am.*, vol. 34, no. 9B, pp. 1466–1471, 1962.
- [227] P. Zahorik and F. L. Wightman, "Loudness constancy with varying sound source distance," *Nature neuroscience*, vol. 4, no. 1, pp. 78–83, 2001.

BIBLIOGRAPHY

- [228] G. Stewart, "The acoustic shadow of a rigid sphere, with certain applications in architectural acoustics and audition," *Physical Review (Series I)*, vol. 33, no. 6, p. 467, 1911.
- [229] R. Hartley and T. Fry, "The binaural location of pure tones," *Physical Review*, vol. 18, no. 6, pp. 431–442, 1921.
- [230] J. Huopaniemi and K. A. Riederer, "Measuring and modeling the effect of source distance in head-related transfer functions," in *Proc. of 19th International Congress on Acoustics, Seattle*, 1998.
- [231] B. G. Shinn-Cunningham, J. Schickler, N. Kopo, and R. Litovsky, "Spatial unmasking of nearby speech sources in a simulated anechoic environment," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1118–1129, 2001.
- [232] K.-h. Shin and Y. Park, "Modeling of non-individualized head-related transfer functions for nearby sources," in *Proc. of 9th Western Pacific Acoustics Conference (WESPAC'06)*, 2006.
- [233] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *J. Acoust. Soc. Am.*, vol. 107, no. 3, pp. 1627–1636, 2000.
- [234] H. R. Hirsch, "Perception of the range of a sound source of unknown strength," *J. Acoust. Soc. Am.*, vol. 43, no. 2, pp. 373–374, 1968.
- [235] J. Molino, "Perceiving the range of a sound source when the direction is known," *J. Acoust. Soc. Am.*, vol. 53, no. 5, pp. 1301–1304, 1973.
- [236] Y. Tahara and H. Sakurai, "A tentative model for the localization of sound based on the simultaneity of time difference and level difference of the sound between

BIBLIOGRAPHY

- ears,” in *Proceedings of the Congress of the Acoustical Society of Japan, Japan*, 1974.
- [237] D. S. Brungart, “Preliminary model of auditory distance perception for nearby sources,” *Computational models of auditory function*, pp. 83–96, 2001.
- [238] R. Hershkowitz and N. Durlach, “Interaural time and amplitude jnds for a 500hz tone,” *J. Acoust. Soc. Am.*, vol. 46, no. 6B, pp. 1464–1467, 1969.
- [239] W. Rabinowitz, J. Maxwell, Y. Shao, and M. Wei, “Sound localization cues for a magnified head- implications from sound diffraction about a rigid sphere,” *Presence. Teleop. Virt.*, vol. 2, no. 2, pp. 125–129, 1993.
- [240] C. Ze-Wei, Y. Guang-Zheng, X. Bo-Sun, and G. Shan-Qun, “Calculation and analysis of near-field head-related transfer functions from a simplified head-neck-torso model,” *Chinese Physics Letters*, vol. 29, no. 3, p. 034302, 2012.
- [241] D. Ramani, D. Zotkin, and N. Gumerov, “Interpolation and range extrapolation of hrtfs,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP’04), Montreal*, 2004.
- [242] D. Menzies, “Calculation of near-field head related transfer functions using point source representations,” in *Ambisonics Symposium, IEM Graz*, 2009, pp. 23–28.
- [243] D. Romblo and B. Cook, “Near-field compensation for hrtf processing,” in *Proc. 125th Audio Engineering Society Convention, San Francisco*, 2008.
- [244] A. Kan, C. Jin, and A. van Schaik, “A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function,” *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2233–2242, 2009.

BIBLIOGRAPHY

- [245] P. Zahorik, "Distance localization using nonindividualized headrelated transfer functions," *J. Acoust. Soc. Am.*, vol. 108, no. 5, pp. 2597–2597, 2000.
- [246] C. Avendano, R. O. Duda, and V. R. Algazi, "Modeling the contralateral hrtf," in *Proc. 16th Audio Engineering Society Conference on Spatial Sound Reproduction, Rovaniemi*, 1999.
- [247] D. Mershon, W. Ballenger, A. Little, P. McMurtry, and J. Buchanan, "Effects of room reflectance and background noise on perceived auditory distance," *Perception*, vol. 18, no. 3, pp. 403–416, 1989.
- [248] D. H. Mershon and J. N. Bowers, "Absolute and relative cues for the auditory perception of egocentric distance," *Perception*, vol. 8, no. 3, pp. 311–322, 1979.
- [249] D. Mershon and L. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975. [Online]. Available: <http://dx.doi.org/10.3758/BF03204113>
- [250] B. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Distance perception of nearby sources in reverberant and anechoic listening conditions: Binaural vs. monaural cues," in *Assoc Res Otolaryn. Meeting*, vol. 23.
- [251] D. Brungart and B. Simpson, "Auditory localization of nearby sources in a virtual audio display," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, 2001, pp. 107–110.
- [252] C. J. Chun, H. K. Kim, S. H. Choi, S.-J. Jang, and S.-P. Lee, "Sound source elevation using spectral notch filtering and directional band boosting in stereo loudspeaker reproduction," *IEEE Trans. on Consumer Electronics*, vol. 57, no. 4, pp. 1915–1920, 2011.

BIBLIOGRAPHY

- [253] G. Kendall, C. Rodgers, G. Kendall, and C. Rodgers, "The simulation of three-dimensional localization cues for headphone listening," in *International Computer Music Conference, Denton*, 1981, pp. 225–243.
- [254] C. Avendano, V. R. Algazi, and R. O. Duda, "A head-and-torso model for low-frequency binaural elevation effects," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99)*. IEEE, pp. 179–182.
- [255] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 2053–2064, 2002.
- [256] G. Woon-Seng, S.-E. Tan, E. Meng-Hwa, and C. Yong-Kim, "Elevated speaker projection for digital home entertainment system," *IEEE Trans. on Consumer Electronics*, vol. 47, no. 3, pp. 631–637, 2001.
- [257] S. Hwang and Y. Park, "Hrir customization in the median plane via principal components analysis," 2007.
- [258] M. Guldenschuh, A. Sontacchi, F. Zotter, and R. Holdrich, "Hrtf modeling in due consideration variable torso reflections," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 3080–3080, 2008.
- [259] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Applied Acoustics*, vol. 68, no. 8, pp. 835–850, 2007.
- [260] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae)," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 317–333, 2014.

BIBLIOGRAPHY

-
- [261] S. Nishioka, Y. Ishii, and K. Iida, "Just noticeable difference of frequency of the first peak of head-related transfer functions with regard to vertical angle perception," in *Proc. Autumn Annual Meeting of the Acoustical Society of Japan*, 2013.
 - [262] M. A. Blommer and G. H. Wakefield, "Pole-zero approximations for head-related transfer functions using a logarithmic error criterion," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 5, no. 3, pp. 278–287, 1997.
 - [263] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and zero modeling of head-related transfer functions," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 7, no. 2, pp. 188–196, 1999.
 - [264] E. A. Durant and G. H. Wakefield, "Efficient model fitting using a genetic algorithm: pole-zero approximations of hrtfs," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 10, no. 1, pp. 18–27, 2002.
 - [265] A. Kulkarni and H. S. Colburn, "Infinite-impulse-response models of the head-related transfer function," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1714–1728, 2004.
 - [266] B. Yegnanarayana, "Formant extraction from linearprediction phase spectra," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1638–1640, 1978.
 - [267] B. Yegnanarayana, D. Saikia, and T. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 32, no. 3, pp. 610–623, 1984.
 - [268] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
 - [269] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

BIBLIOGRAPHY

- [270] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.
- [271] M. M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'07)*, vol. 1. IEEE, pp. I-9–I-12.
- [272] J. Breebaart and E. Schuijers, "Phantom materialization: A novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1503–1511, 2008.
- [273] C. Avendano and J.-M. Jot, "A frequency-domain approach to multichannel up-mix," *J.Audio Eng.Soc.*, vol. 52, no. 7/8, pp. 740–749, 2004.
- [274] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J.Audio Eng.Soc.*, vol. 54, no. 11, pp. 1051–1064, 2006.
- [275] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 33–42, 2011.
- [276] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in *Proc. 128th Audio Engineering Society Convention, London*, 2010.
- [277] C. Faller and F. Baumgarte, "Binaural cue coding-part ii: Schemes and applications," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 11, no. 6, pp. 520–531, 2003.
- [278] A. Hyvriinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

BIBLIOGRAPHY

- [279] M. Joho, H. Mathis, and R. H. Lambert, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," in *Proc. International Conference on Independent Component Analysis and Blind Signal Separation. Helsinki, Finland*, 2000.
- [280] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2010.
- [281] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *Proc. 7th International Conference on Digital Audio Effects (DAFX-04)*.
- [282] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [283] M. Cobos and J. J. Lopez, "Stereo audio source separation based on timefrequency masking and multilevel thresholding," *Digital Signal Processing*, vol. 18, no. 6, pp. 960–976, 2008.
- [284] T. Virtanen, "Sound source separation in monaural music signals," Ph.D. dissertation, 2006.
- [285] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley Publishing, 2009.
- [286] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

BIBLIOGRAPHY

- [287] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.
- [288] J. Merimaa, M. M. Goodwin, and J.-M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. 123rd Audio Engineering Society Convention*, New York, 2007.
- [289] J. He, E.-L. Tan, and W.-S. Gan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'14)*, 2014.
- [290] A. Hrm, "Classification of timefrequency regions in stereo audio," *J.Audio Eng.Soc.*, vol. 59, no. 10, pp. 707–720, 2011.
- [291] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2141–2150, 2007.
- [292] J. He, E.-L. Tan, and W.-S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'13)*. IEEE, pp. 266–270.
- [293] J. Thompson, B. Smith, A. Warner, and J.-M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pairwise correlations," in *Proc. 133rd Audio Engineering Society Convention*, San Francisco, 2012.
- [294] A. Hrm, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *J.Audio Eng.Soc.*, vol. 52, no. 6, pp. 618–639, 2004.

BIBLIOGRAPHY

- [295] D. W. Schobben and R. M. Aarts, "Personalized multi-channel headphone sound reproduction based on active noise cancellation," *Acta acustica united with acustica*, vol. 91, no. 3, pp. 440–450, 2005.
- [296] V. Valimaki, A. Franck, J. Ramo, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 92–99, 2015.
- [297] DTS, "Headphone:x, <http://www.headphonex.com/>."
- [298] Microsoft, "Hololens, <https://www.microsoft.com/microsoft-hololens/en-us>."
- [299] Oculus, "Oculus rift, <https://www.oculus.com/>."
- [300] Dolby, "Dolby 7.1 headphones, <http://www.dolby.com/us/en/technologies/dolby-headphone.html>."
- [301] J. Borwick, *Loudspeaker and headphone handbook*. Woburn, MA: CRC Press, 2012.
- [302] A. Mark R and L. J. Kelly, "Principles of headphone designa tutorial review," in *Proc. 21st Audio Engineering Society Conference on Audio at home, UK*, 2006.
- [303] C. Poldy, "Headphone fundamentals," in *Tutorial presented at the Proc. 120th Audio Engineering Society Convention, Paris*, 2006.

Appendices

Appendix A

Spatial oriented format for acoustics (SOFA)

The spatial oriented format for acoustics (SOFA) [32] is a format for storing HRTFs with a focus on interchangeability and extendability. SOFA aims at representing HRTFs in a standard format, which allows to store data such as a directional room impulse response (DRIR) measured with a microphone array excited by the loudspeaker array. SOFA aims at simplifying the development of programming interfaces for different platforms like Matlab, C ++, Octave. SOFA stores all the information in a single file by serializing the data in to a binary stream. SOFA files have an extension “.sofa”.

In the SOFA format, along with the HRTF data, several other metadata are stored in the form of objects that describe the complete experimental setup. Figure B.1 shows the typical HRTF/DRIR measurement setup.

A common property in all the measurement setups is that excitation signals are generated and microphones are used to record the incoming signals. The measurements are repeated while varying the position of the listener, the sound source, or both. In general, binaural measurement setups require only two microphones to record the left and right ear signals. However, some HRTF measurements may also require multiple microphones.

APPENDIX A. SPATIAL ORIENTED FORMAT FOR ACOUSTICS (SOFA)

Figure A.1: Typical HRTF/DRIR measurement setup

For example, a multichannel microphone array arranged around the listeners in a HRTF measurement system based on reciprocity, multichannel microphone arrays for measuring DRIRs, three microphone system in hearing aid applications, concert hall acoustics measurements. In the SOFA format, a microphone as a single receiver of the sound field is called the receiver, and the object comprising all the receivers is called the listener as shown in Fig. B.1. Similarly, the sources creating the excitation signal are called emitters and the object containing all the emitters is referred to as the source.

Receiver: Any acoustic sensor or microphone. The number of receivers is not limited in SOFA and defines the size of the data matrix.

Listener: Listener is the object containing all the receivers. For a binaural measurement, the listener consists of two receivers. For DRIRs, a listener might represent a microphone-array structure such as a sphere or a frame.

Emitter: An emitter is any acoustic excitation used for the measurement. The

APPENDIX A. SPATIAL ORIENTED FORMAT FOR ACOUSTICS (SOFA)

number of emitters is not limited in SOFA. The metadata describes the contribution of the particular emitter.

Source: Source is the object, which comprises all the emitters. In SOFA, a source might be a multi-driver loudspeaker, or a speaker array, etc. Only one source is considered but the source may incorporate an unlimited number of emitters.

Room: Room is the volume enclosing the measurement setup. In the case of a free-field measurement, the room is not considered. Other optional objects can be given as user-defined metadata of a measurement.

A.1 Objects

A.2 SOFA Conventions

SOFA conventions are definitions of data and metadata consistently describing particular HRTF/DRIR measurement setups. SOFA conventions are specified in order to meet the requirements from different application fields. Some of the SOFA conventions for which the SOFA files are publicly available and can be read/modified by a software package are:

- General FIR: General convention with FIR as DataType (no restrictions but DataType)
- General FIRE: General convention with FIRE as DataType (no restrictions but DataType)
- General TF: General convention with TF as DataType (no restrictions but DataType)
- SimpleHeadphoneIR: Conventions for IRs with a 1-to-1 correspondence between emitter and receiver. The main application for this convention is to store headphone IRs recorded for each emitter and each ear.

APPENDIX A. SPATIAL ORIENTED FORMAT FOR ACOUSTICS (SOFA)

Field ▾	Value	Min	Max
Obj <1x1 struct>			
Data	<1x1 struct>		
ListenerView_Units	'meter'		
ListenerView_Type	'cartesian'		
ListenerView	[1,0,0]	0	1
ListenerUp	[0,0,1]	0	1
EmitterPosition_Units	'meter'		
EmitterPosition_Type	'cartesian'		
EmitterPosition	[0,0,0]	0	0
SourcePosition_Units	'degree, degree, meter'		
SourcePosition_Type	'spherical'		
SourcePosition	<600x3 double>	0	359
ReceiverPosition_Units	'meter'		
ReceiverPosition_Type	'cartesian'		
ReceiverPosition	[0,-0.0900,0;0,0.0900,0]	-0.0900	0.0900
ListenerPosition_Units	'meter'		
ListenerPosition_Type	'cartesian'		
ListenerPosition	[0,0,0]	0	0
API	<1x1 struct>		
GLOBAL_ListenerShortName	'HATS dummy head'		
GLOBAL_DatabaseName	'DSPLAB'		
GLOBAL_Title	''		
GLOBAL_DateModified	'2014-09-26 15:27:27'		
GLOBAL_DateCreated	'25 September 2014'		
GLOBAL_Origin	''		
GLOBAL_RoomType	'free field'		
GLOBAL_References	''		
GLOBAL_Organization	''		
GLOBAL_License	'No license provided, as...		
GLOBAL_History	'Converted from the CIP...		
GLOBAL_DataType	'FIR'		
GLOBAL_Comment	'This is a range-depende...		
GLOBAL_AuthorContact	'Kaushik Sunder (kaushi...		
GLOBAL_ApplicationVersion	'0.4.3'		
GLOBAL_ApplicationName	'Demo of the SOFA API'		
GLOBAL_APIVersion	'0.4.3'		
GLOBAL_APIName	'ARI SOFA API for Matla...		
GLOBAL_SOFAConventions...	'0.4'		
GLOBAL_SOFAConventions	'SimpleFreeFieldHRIR'		
GLOBAL_Version	'0.6'		
GLOBAL_Conventions	'SOFA'		

Figure A.2: SOFA attributes contained in the HRTF database

- MultispeakerBRIR:
- SimpleFreeFieldHRIR: Free-field HRTFs stored as impulse responses, measured with an omnidirectional source for a single listener.
- SimpleFreeFieldTF: Similar to SimpleFreeFieldHRIR but uses TF as DataType covering special needs coming from HRTF simulations.
- SingleRoomDRIR: Room impulse responses measured with an arbitrary number of

APPENDIX A. SPATIAL ORIENTED FORMAT FOR ACOUSTICS (SOFA)

receivers (such as a microphone array) and an omnidirectional source in a single room.

In the DSPLab HRTF database developed, we use the SOFA convention “Simple-FreeFieldHRIR”. Matlab files are provided to extract the HRIR files from the SOFA format file.

For example, `[Obj,Dims] = NETCDFload('DSPLAB_HRTF_512samples.sofa','all');` returns an object ‘Obj’ which contains all the attributes of the HRTF database. The attribute “Data” is a structure, which contains the impulse responses in a 3D array “IR”.

`HRIR = extractHRIR(azim, r, nsamples);` extracts the HRIR for a given azimuth and distance. The matlab function finds the nearest azimuth and distance in the HRTF database and extracts the corresponding HRIR.

$n_{samples}$ is the number of samples in the HRIR; $n_{samples}$ can be 256 samples or 512 samples.

```
function HRIR = extractHRIR(azim, r, n-samples)
% This function extracts the HRIR from the database in SOFA format
% azim is the azimuth angle
% r is the distance of the HRTF in cm
% n-samples can be 256 or 512 samples
% HRIR contains the left and the right ear HRTF.
% Written by Kaushik Sunder, 23/10/2014

% Angles and Distances in the HRTF database
angles = [0 4:5:89 90 94:5:269 270 274:5:359];
dist = [35 45 50 60 75 80 90 140];

% Find the nearest azimuth and distance
[Y idx_azim]= min(abs(azim-angles));
```


APPENDIX A. SPATIAL ORIENTED FORMAT FOR ACOUSTICS (SOFA)

```

azim_nearest = angles(idx_azim);

[Y idx_dist]= min(abs(r-dist));
dist_nearest = dist(idx_dist);

% Obj contains all the attributes of the HRTF database
[Obj,Dims] = NETCDFload(['DSPLAB-HRIR-' num2str(n_samples) 'samples.sofa'],'all');

% Extracting all the HRIR information of size (600 X 2 X n_samples)
HRIR_all = Obj.Data.IR;
HRIR_all = permute(HRIR_all,[3,1,2]);

% Extracting the HRIR for azim and r
I = idx_azim + 75*(idx_dist-1);
HRIR_L = HRIR_all(:,I,1);
HRIR_R = HRIR_all(:,I,2);
HRIR = [HRIR_L HRIR_R];

end

```


Appendix B

Range dependent spherical head model

Duda and Martens [172] developed numerical equations which could model acoustical interactions on a spherical head for a point source. This study was carried out to investigate the behavior of the HRTF at close range. A spherical model of the head was considered in this study to analyze the near-field HRTFs. A mathematical analysis of the head response is first carried out. Subsequently, numerical techniques are introduced to evaluate the theoretical solution. Though the model is quite idealized, the extracted features are close to those observed in HRTFs in proximal region.

The complex solution for a source infinitely far from the sphere is first developed. This solution aids in computing the response at the surface of the sphere for a sound source located at any particular distance. The derivation of the range-dependent spherical head model is presented next [172].

APPENDIX B. RANGE DEPENDENT SPHERICAL HEAD MODEL

List of symbols used:

a	radius of the sphere (m)
h_m	m^{th} order spherical Hankel function
c	ambient speed of the sound (m/s)
f	frequency (Hz)
h	head-related impulse response
H	head-related transfer function relative to free field
H_ε	head-related transfer function relative to source
j_m	m^{th} order spherical Bessel function
k	acoustics wave number (/m)
p_{ff}	free-field pressure at the center of the sphere (kg/m^2)
p_s	pressure on the surface of the sphere (kg/m^2)
p_ε	pressure at a small surface surrounding the source (kg/m^2)
r	distance from the center of the sphere to the source (m)
r_ε	radius of a small sphere surrounding the source
S_ω	magnitude of flow from an ideal point source (m^3/s)
t	time(s)
θ	angle of incidence (rad)
λ	wavelength (m)
μ	normalized frequency
ρ	normalized distance to the source
ρ_0	density of air (kg/m^3)
ω	radian frequency (rad/s)

Lord Rayleigh first obtained the frequency domain solution for the acoustic wave diffraction by a rigid sphere. Consider the flow for a complex sinusoidal point source to be of the form $S_\omega e^{-i\omega t}$. The free-field pressure at a distance r from the source is given by:

$$p_{ff}(r, \omega, t) = -i\omega \frac{\rho_0 S_\omega}{4\pi r} e^{-i(kr - \omega t)}, \quad (\text{B.1})$$

where $k = \omega/c$. It is important to note the term $-i\omega$ in the free-field pressure equation. Multiplying by $-i\omega$ in the frequency domain corresponds to differentiation in the time domain. This implies that given a unit step function as representing the flow, the dirac impulse wave function would then represent the free-field pressure whose strength varies inversely with the distance to the source.

The sound waves get diffracted and thus, modified the pressure field in the presence

APPENDIX B. RANGE DEPENDENT SPHERICAL HEAD MODEL

Figure B.1: HRTFs at the left ear for various source directions and distances simulated by the spherical head model

APPENDIX B. RANGE DEPENDENT SPHERICAL HEAD MODEL

of the sphere. The solution can be written as:

$$p_s(r, a, \omega, \theta, t) = \frac{i\rho_0 c S_\omega}{4\pi a^2} \psi e^{-i\omega t}, \quad (\text{B.2})$$

where ψ is the infinite series expansion,

$$\psi = \sum_0^\infty (2m+1) P_m(\cos \theta) \frac{h_m(kr)}{h'_m(ka)}, r > a. \quad (\text{B.3})$$

Here θ is the incidence angle that corresponds to the angle between the vector from the source center to the sphere center and the vector to the measurement point on the surface of the sphere. Angle of incidence normal to the sphere is considered to be $\theta = 0^\circ$.

$$\mu = ka = f \frac{2\pi a}{c}. \quad (\text{B.4})$$

The Time $2\pi a/c$ corresponds to the time taken by a wave to travel once around the sphere. μ is the normalized frequency.

Defining the normalized distance to the source ρ by,

$$\rho = \frac{r}{a}. \quad (\text{B.5})$$

And the transfer function H by,

$$H = \frac{p_s}{p_{ff}}, \quad (\text{B.6})$$

Then,

$$H(\rho, \mu, \theta) = \frac{-\rho}{\mu} e^{-i\rho\mu} \psi, \quad (\text{B.7})$$

where

$$\psi(\rho, \mu, \theta) = \sum_{m=0}^\infty (2m+1) P_m(\cos \theta) \frac{h_m(\mu\rho)}{h'_m(\mu)}, \rho > 1. \quad (\text{B.8})$$

This HRTF (H) relates the pressure that would be present at the center of the sphere in

APPENDIX B. RANGE DEPENDENT SPHERICAL HEAD MODEL

free field to the pressure that is developed at the surface of the sphere. Equations B.7 and B.8 represent a range-dependent spherical head model and can be implemented in MATLAB. Using this equation, we can identify the important features due to the head shadowing effect and others in the near field.

Appendix C

Inverse filtering

If an acoustical impulse response contains reflections, there will be repeated and similar magnitude characteristics during sound propagation. This causes the impulse response to consist of a maximum-phase and a minimum phase component. In terms of Z transformation, the minimum phase components are within the unit circle, while the maximum phase components are outside the unit circle. Those components can be seen as numerator coefficients of a digital FIR filter. Inversion turns the numerator coefficients into denominator coefficients and those outside the unit circle will make the resulting filter (which is now an IIR filter) unstable. Now the inverse filter is stable but non-causal (left handed part of response towards negative times). To compensate this, the resulting response is shifted in time to make the non-causal part causal. But the “true” inverse is still an infinite one but is represented by an finite (Nfft-long) approximation. Due to this fact and due to the periodic nature of the DFT, the Nfft-long “snapshot” of the true inverse also contains overlapping components from adjacent periodic repetitions ($\Rightarrow$ “time aliasing”). Windowing the resulting response helps to suppress aliasing at the edges but does not guarantee that the complete response is aliasing-free. In fact, inverting non-minimum phase responses will always cause time aliasing - the question is not “if at all” but “to which amount” to Time-aliasing “limiters”:

APPENDIX C. INVERSE FILTERING

- Use of short impulse responses to be inverted ($\Rightarrow$ windowing prior to inverse filter design).
- Use of longer inverse filters ($\Rightarrow$ increasing FFT length).
- Avoids inversion of high-Q (narrow-band spectral drips/peaks with high amplitude) spectral components ($\Rightarrow$ regularization, smoothing)

In addition, the parameters should be chosen to minimize the left-sided part of the filter response to minimize perceptual disturbing pre-ringing.

Appendix D

Theoretical model for headphones

Loudspeakers are generally modeled by deriving their electrical or acoustical impedance equivalent circuits using a lumped-element technique [301]. The lumped-element technique can also be extended to model the headphones [302] [303].

An important assumption in such a modeling technique is that each element must not in itself exhibit wave behavior. In a mechanical analogy, if a vibrating diaphragm is considered as a lumped mass, all points in the diaphragm must oscillate with identical velocity, both in magnitude and phase. One of the limitations of the lumped element technique is that it is generally restricted to low-frequencies (upto 2 kHz), since most mechanical and acoustic systems behave modally (in a distributed than a lumped manner) within the audio bandwidth. However, since headphones are small devices, this technique can be used to illustrate some of the fundamental behaviors of the headphone system across a useful range of frequencies.

D.1 Pressure-chamber principle

Headphones differ from loudspeakers in the manner they produce sound at the ear. In the case of loudspeaker playback, the ear is immersed in a propagating sound field, while

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

in headphone playback, it registers the SPL in a leaky pressure chamber. The sound field of headphones is confined to a relatively small volume of approximately 30 cm^3 .

In a closed headphone, the force pumps on a cavity giving a sound pressure proportional to the excursion. Sound pressure level in the cavity can be assumed to be distributed uniformly in the volume even in presence of mild leaks upto around 2 kHz, where the wavelength is still larger compared to the dimensions of the cavity [301]. For intermediate frequencies where the influence of leaks is low, the cavity can be regarded as a pressure chamber. Here, the pressure is in phase with the volume displacement of the transducer membrane and its amplitude is proportional to it.

SPL in a pressure chamber for a simple leak free headphone at low-frequencies for a small piston displacement can be expressed as [301]:

$$p = \frac{S \times x}{C_A} \quad (\text{D.1})$$

where C_A is the acoustic compliance of the pressure chamber which is described by $\frac{V}{\gamma \times P_{atmos}}$. V is the volume of the chamber, γ is the Specific heat of the medium (1.4 for air), P_{atmos} is the atmospheric pressure.

D.2 Analogies between different domains :

In a typical headphone working principle, the voltage excites the voice coil (Electrical) which in turn generates the membrane vibrations (Mechanical). These vibrations displace the air giving a perception of audible sound (Acoustical). Let us first describe the analogies between the acoustical, electrical and mechanical domains (Table D.1). For acoustical quantities, pressure p ($\frac{N}{m^2}$) is equivalent to voltage U , and volume velocity q ($\frac{m^3}{s}$) is equivalent to current I . For mechanical quantities, force F (N) is equivalent to voltage U , and velocity v (m) is equivalent to current I . The acoustical elements can

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Table D.1: Analogies between Electrical, Mechanical, and Acoustical domains

Energy type	Electrical (V,I)	Mechanical (F,v)	Acoustical (P,U)
Dissipative	RI	$R_m v$	$R_A U$
Potential (Spring)	$\frac{1}{j\omega L} I$	$\frac{k}{j\omega} v$	$\frac{C_A}{j\omega} U$
Kinetic (Inertia)	$j\omega L$	$j\omega M v$	$j\omega m U$

be expressed in the mechanical units, if the mechanical resistances and inductances are divided by S^2 , and mechanical condensers are multiplied by S^2 , where S is the cross-sectional area which is perpendicular to the flow direction [301].

$$R_A = \frac{R_m}{S^2} \quad L_A = \frac{m}{S^2} \quad C_A = \frac{C_m}{S^2} \tag{D.2}$$

It should be noted that subscripts A and m indicate acoustical and mechanical quantities, respectively.

D.3 Components of headphones :

Headphones mainly comprises of the following categories of building blocks [301] :

- Cavity C (Acoustic compliance) connected to the ground
- Acoustic bottlenecks, porous paper, holes, slits; R (Resistance), L (Acoustic Mass or inertance) in series;
- Compliant membrane : L (mass) , C (compliance), R (damping resistance), all in series ;
- Radiation impedances : (R, L in parallel);
- Mechanical-acoustical interface : Ideal transformer
- Electrical-mechanical interface : Gyrator

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Table D.2: Acoustic elements and their electrical equivalent circuits, and R , L , C are expressed in terms of the mechanical dimensions l , b , d (m), S (m²), V (m³). Air density is 1.2 Kg/m³, Speed of sound $c = 340$ m/s, Specific heat of air = 1.4, $P_{atmos} = 105$ Pa.

SI units throughout	Equivalent Circuit	R	L	C
<p>Hole</p>		$\frac{8\eta l}{\pi\left(\frac{d}{2}\right)^4}$	$\frac{4\rho l}{3\pi\left(\frac{d}{2}\right)^2}$	—
<p>Slit (b<d)</p>		$\frac{12\eta l}{bd^3}$	$\frac{6\rho l}{5bd}$	—
<p>Cavity Compliance V = Volume</p>		—	—	$\frac{V}{\gamma\times P_{atmos}}$
<p>Membrane S = area m = mass f_R = resonance</p>		$\cong 0$	$\frac{m}{S^2}$	$\frac{S^2}{4\pi^2mf_R^2}$
<p>Radiating Piston S = Area</p>		$\frac{\rho c_0}{S}$	$\frac{0.85\rho}{\sqrt{\pi S}}$	—

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Table D.2 indicates the acoustic elements present in headphones and their electrical equivalent circuits. Each of the components are explained in further detail.

Cavity: From equation, the acoustic compliance of a cavity can be expressed as:

$$C = 7.14 \times 10^{-6} \times V \quad (\text{acoustic Farad SI}) \quad (\text{D.3})$$

where V is in m^3 . One of the condenser terminals is invariably grounded.

Acoustical bottlenecks or constrictions : Resonances can be damped using porous elements. A large R (resistance) and a small L (airmass) is required to prevent from competing with R . Paper and woven resistances are typically favored since the pores though irregular, are very small. They can be regarded as a collection of holes in parallel. R decreases with increasing diameter to the 4th power, L only to the 2nd power (Table D.2). Therefore, small diameters have good damping.

Membranes: Membranes are represented by R , L , and C in series (Table D.2). These acoustic parameters can also be related to the mechanical ones R_m , m , C_m as shown using the effective membrane area S . Thiele-small procedure [301] can be used to obtain these parameters for the driver. L and C can be readily determined from the total mass and resonance if the effective area S is known. Membrane movement is largely stiffness controlled C or resistance controlled (damping R) or usually a combination of both C and R . This is unlike loudspeakers, which are predominantly mass controlled L .

Radiation Impedances: As most of the headphones are not entirely closed, radiating components are also present due to openings or leaks. They are often represented by a parallel combination of R and L . The real component R predominates at high frequencies, and represents energy loss by radiation. This component can be utilized for damping resonances. At low frequencies, the radiation impedance becomes reactive and is equivalent to a mass loading. Radiation impedances in headphones are in general very low compared with the impedances behind the openings where they radiate. Thus, they

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Figure D.1: Acoustic impedance analogue of a dynamic driver

Figure D.2: A simplified headphone-earcanal model

can be replaced by a short circuit in most approximations.

D.4 Model of a headphone-earcanal system :

A dynamic transducer is selected to facilitate a simple electro-mechanical modeling of the system. An acoustical (p, U) impedance analogue for a dynamic driver is shown in Fig. D.1. In Figure D.1, e represents the applied voltage, Bl represents the force factor (NA^{-1}), S_D is the diaphragm area (m^2), R_E = DC coil resistance, L_E = Coil inductance,

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

M_{AD} = moving diaphragm mass (kgm^{-4}), C_{AD} = suspension compliance ($m^{-5}N^{-1}$), R_{AD} = Suspension resistance (Nsm^{-5}), Z_{AB} = Acoustic backload ($Pa.sm^{-3}$), Z_{AF} = Acoustic frontload ($Pa.sm^{-3}$).

All the parameters except Z_{AB} and Z_{AF} are determined by the selection of the drive unit, and these Thiele-small parameters are obtained by conventional measurement techniques [302]. Figure D.2 shows the simplest design case of a driver in a sealed enclosure. The acoustic load is dependent on the input impedance to the ear canal $Z_{A,0}$, that is a function of canal length, cross-section, and termination impedance.

The input impedance of the earcanal $Z_{A,0}$ can be expressed in terms of the eardrum impedance and the cross-sectional area of the earcanal :

$$Z_{A,0} = \frac{Z_{A,T} + j\rho_0 c/S_1 \tan(kl)}{1 + j \frac{Z_{A,T}}{\rho_0 c/S_1} \tan(kl)} \quad (D.4)$$

The lumped impedance of an enclosed volume is given as :

$$C_A = \frac{V}{\gamma P_0} = \frac{V}{\rho_0 c^2} \quad (D.5)$$

Therefore, the acoustic loads Z_{AB} and Z_{AF} shown in Figure D.1 shown can be replaced by the network shown in Figure D.3. The pressure represents the pressure at the entrance of the ear canal.

The pressure is a function of driver velocity that takes a second-order bandpass response around a resonant frequency, which is determined by diaphragm mass and the series combination of suspension and backload compliances $\frac{C_{AD}C_{AB}}{C_{AD}+C_{AB}}$ [301]. It should be noted that (Volume of front cavity) generates a shunt compliance that acts as a low-pass filter. This suggests that the volume V_F should be kept as small as possible at the same time comfortably accommodating the pinnae of users. A larger coupler volume reduces the sensitivity. However, increased sensitivity of small coupler comes at a cost of narrow

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Figure D.3: Acoustic circuit analogue for a simple headphone-earcanal model

bandwidth of reproduction.

It is known that the subjective sensation of listening is dependent on the pressure at the eardrum. The pressures at each end of a cylindrical duct of known termination impedance can be related to each other as:

$$p_l = \frac{p_0}{\{\cos(kl) + j \frac{\rho_0 c}{S_1 Z_{A,T}} \sin(kl)\}} \quad (\text{D.6})$$

Earcanal can be considered as a cylindrical duct (two-port tube) generating resonances of infinite amplitude at 3 kHz ($\frac{\lambda}{4}$), 6 kHz ($\frac{2\lambda}{4}$), and 9 kHz ($\frac{3\lambda}{4}$) if terminated hard. The eardrum has an effect of damping these peaks by 12 dB. The transfer function p_{out}/p_{in} of any two-port system can in principle only depend on the two ports itself and the output impedance.

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Figure D.4: Types of headphones and schematic characteristic responses at low-frequencies. a) closed headphones b) open foam cushion c) cushion of impermeable foam with fixed resistance d) Integrated open-back headphones. Figure adapted from reference [301].

D.5 Analysis of leakages in headphones :

The lumped-element analysis is a useful tool to analyze the radiation due to leakages in headphones. Headphones do not rigidly attach to the side of the head leading to leakage of sound. The leakage can be uncontrolled as sound can leak through hair/ or

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Figure D.5: Simple model of headphone-earcanal system with leaks.

Figure D.6: Acoustic circuit analogue for a headphone model with leaks.

through porous cushions or through uncontrolled porosity in squashed foam [301].

However, in some designs, leakage can be controlled using rear vents as in the open-back headphones. Figure. D.4 shows the characteristic responses at low-frequencies for a) closed b) open foam cushion c) cushion of impermeable foam with fixed resistance d) vented rear cavity. Usually, leaks are desirable and a deliberate part of the acoustical

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

design. As a result of controlled leaks, the SPL in the cavity becomes more stable with respect to additional chance leaks. The two are like resistances in parallel, whichever is lower will dominate.

Figure. D.5 shows a simple model of headphone/ear system including leaks. Such leaks can be modeled as a branch pipe that further diverts driver volume velocity from the driver. Small losses generally exhibit lossy as well as inductively reactive behavior. As a first approximation, the leak may be modeled as a lumped mass :

$$M_{AL} = \frac{\rho_0 l_2}{S_2} \quad (D.7)$$

The impedance (Figure. D.6) acts as a high-pass filter, as one could expect considering the spectral effects of pushing the cup more firmly to the side of the head.

D.6 Closed Vs Open-back designs :

Leak impedances can significantly affect the pressure responses at the eardrum [302] [303]. The filtering effect due to these leaks cannot be compensated electronically as the spectral effects vary highly between users. The leak variability can be lessened by instituting a large, predetermined leak between and and / or free space. One such simple configuration is shown in Figure D.7. Figure D.8 shows the acoustic circuit analogue for an open-back headphone.

As long as the predetermined front/back leakage path has small impedance compared to the variable leakages, the spectrum at the eardrum can be maintained more constant between users that might otherwise be the case. However, the sensitivity of the device is also reduced considerably. In most of the cases, such a loss in sensitivity is tolerable as the drop in sound pressure level at the ear may be compensated by increasing the amplitude of the electrical input accordingly, as the driver unit remains linear in operation to a

APPENDIX D. THEORETICAL MODEL FOR HEADPHONES

Figure D.7: Simple model for an open-back headphone.

Figure D.8: Acoustic circuit analogue for an open-back headphone.

sufficiently large excursion. The open-back design lacks low-frequency emphasis unlike a closed-back design but the key feature about these open-back designs is the lesser significant effect on the pressure at the eardrum.

In this thesis, the circumaural headphones used were mainly open-back headphones due to this reason. The lumped-element method can also be used to model the frontal projection headphones (open headphones).