3D AUDIO REPRODUCITON: NATURAL AUGMENTED REALITY HEADSET AND NEXT GENERATION ENTERTAINMENT SYSTEM USING WAVE FIELD SYNTHESIS

RISHABH RANJAN

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

2016

Acknowledgment

I take this opportunity to thank my supervisor, Assoc. Prof. Woon-Seng Gan, for giving me the opportunity to work on this exciting project. His enthusiasm and exuberance have helped me to carry out my research with composure and confidence. He has been appreciative of my ideas and encouraged critical thinking and technical writing, which have helped me improve my skills.

I would also like to use this opportunity to thank to my colleagues and peers at the Digital Signal Processing (DSP) Laboratory, NTU, especially Mu Hao, Jian Jun, KoKo, Santi, Bhan, Duy Hai, Tatsuya, Valentin, Kushal Anand, Abhishek Seth, Anushree and Anusha for their invaluable suggestions and support. I am extremely fortunate to got a great lab atmosphere to work and thrive in. Our laboratory executive, Mr. Yeo has been enthusiastic in providing me with help and support on issues related to networks and software at DSP.

I am, especially, grateful to my best lab buddies Kaushik Sunder and Apoorv Agha, without whom I could not have completed by PhD. There has been several days when they have helped me with my experiments tirelessly leaving their sleep aside. I will definitely miss the amazing moments we have had working together.

I am also grateful to my dear friends and fellow PhD candidates Ronak Bajaj, Divya Rao, Vipra Guneta, Achiranshu Garg, Abhinava Chaitanya, Ravi Kishore, Naidu and Sandeep for their companionship and support.

I also use this opportunity to thank my friends back in India, Sumit Raj, Satish Prasad, Jairaj Bhattacharya, Anshul Singh, Aman Gupta, Charvi Dhoot, Gaurav

Agarwaal, Sufal Roongta, Mohit Srivastava and Animesh Thakur for their constant support and motivation.

Finally, I am indebted to my family and my parents, for their prayers and encouragement. I thank them for their understanding and their efforts to support me in pursuing higher studies.

Table of Contents

Abstract **xxi**

1 Introduction **1**

1.1 Spatial Audio Overview 1

1.2 Enabling Natural Listening over Headphones and Loudspeakers . . . 7

1.3 Contribution of the Thesis 10

1.4 Structure of the Thesis 12

2 Binaural Technology : A Literature Review **14**

2.1 Head-Related Transfer Functions 14

2.2 Sound Localization 18

2.2.1 Inter-aural Cues (ITD and ILD) 18

2.2.2 Spectral Cues 20

2.2.3 Individualized HRTFs 21

2.2.4 Other Cues 21

2.3 Binaural Synthesis over Headphones 22

2.4 Headphone Equalization 25

2.5 Conclusions 26

3 Wave Field Synthesis using Loudspeaker Arrays **28**

3.1 Wave Field Synthesis: An Overview 28

3.2 Principle of Wave Field Synthesis 29

3.3 Practical Approximations for Wave Field Synthesis 33
3.4 Practical Constraints and Solutions 37
3.5 Evolution of Wave Field Synthesis 43
3.6 Future Trends and Conclusions 45

4 Natural Listening Over Headphones in Augmented Reality using Adaptive Filtering 48

4.1 Introduction 49
4.2 Headphones Effect on Direct Sound Spectrum 51
4.3 Natural Listening via Natural Augmented Reality Headset based on Adaptive Filtering 54
4.3.1 Proposed headset Structure 55
4.3.2 HMTF measurements and observations 57
4.3.3 Case I: Only real source present 58
4.3.4 Case II: Only virtual source present 61
4.3.4.1 Case II results: Conventional FxNLMS Vs Modified FxNLMS 66
4.3.4.2 Hybrid Adaptive Equalizer (Hybrid FxNLMS) 68
4.3.4.3 Case II Results: Hybrid FxNLMS Vs Others 71
4.3.5 Case III: Both virtual and real source present (Augmented reality): HAE with online adaptive estimation 74
4.4 Listening Test 79
4.4.1 Listening test results 83
4.5 Conclusion 87

5 Practical Limitations, Solutions and Extensions of Natural Augmented Reality Headset 89

5.1 Current Practical Limitations Overview 90
5.2 BRIRs acquisition using NAR headset 91

5.2.1 Continuous BRIR acquisition using NLMS 93
5.2.2 Continuous BRIRs acquisition results - *Without Head Rotation* 98
5.2.3 Continuous BRIRs acquisition results - *With Head Rotation* . 100
5.3 Adaptive Equalization for Non-stationary Virtual Signals 102
5.4 Detection and Fast Estimation of Headphone Transfer Function in Natural Augmented Reality Headset 105
5.5 Adaptive Equalization for Non-stationary Virtual signals with Adaptive Estimation of External signals 112
5.6 Conclusions and Further Improvements 118
5.6.1 Conclusions 118
5.6.2 Further Improvements 119

6 A Hybrid Speaker Array-Headphone System for Immersive Audio Reproduction 121

6.1 Introduction 122
6.2 Related Works 125
6.3 Proposed Hybrid System 127
6.4 Wave Field Synthesis Renderer 128
6.4.1 Spatial aliasing in the reproduced sound field 131
6.5 Virtual Wave Field Synthesis over Headphones 132
6.5.1 Virtual Wave Field Synthesis results 134
6.6 Frontal Auditory Scene Processing 138
6.6.1 Frontal array Playback with Headphone Isolation Compensation 138
6.6.2 Frontal array Playback with reproduction of high frequency components over Natural Augmented Reality headset 140
6.7 Subjective Study 144
6.7.1 Listening test 1 - *Localization test* 144
6.7.2 Listening test 2 - *Sound coloration test* 150
6.7.3 Listening test 3 - *Overall audio quality test* 154

6.8 Limitations of the Proposed Hybrid system 157
6.9 Conclusions 158

7 Fast and Efficient Real-Time GPU Based Implementation of Wave Field Synthesis 161

7.1 Introduction 162
7.2 Related Work 163
7.3 GPU and CUDA Overview 164
7.3.1 Thread Blocks, Warp and Coalesced Memory access 165
7.3.2 GPU Overheads 167
7.4 Real-Time WFS Framework 169
7.4.1 Computational complexity of WFS 171
7.5 GPU Implementation 171
7.5.1 Pre-filtering of multiple sources 173
7.5.2 WFS driving signals computation 173
7.5.3 WFS synthesized binaural signals computation 177
7.6 Simulation Results 177
7.7 Conclusions 180

8 Conclusions and Future Works 182

8.1 Conclusions 182
8.2 Future Works 190

Author's Publications 193

Bibliography 194

List of Figures

1.1 (a) Classification of Spatial audio and (b) Time-line of evolution of spatial audio [1] 2

1.2 Typical 5.1 Stereo system 4

1.3 Wave Field Synthesis using loudspeaker arrays shown in black and entire interior is the listening zone 5

2.1 Free-field sound transmission model - adapted from [2] 16

2.2 Transfer function along the ear canal from different points for three azimuths. Approximate measurement positions are shown in the ear diagram. Extracted from [2] to better illustrate the transfer function. 17

2.3 Headphone sound transmission model-adapted from [3] 23

2.4 Binaural recording/synthesis and reproduction over headphones . . . 24

3.1 Block Diagram of a WFS reproduction system 29

3.2 (a) Huygens Principle Realization (b) Rayleigh representation of Huygens Principle 30

3.3 General Sound scene in an enclosed volume (Figure adapted from [4]) 31

3.4 General Sound scene in an enclosed volume with degenerated surface to a plane (Figure adapted from [4]) 32

3.5 WFS in practical scenario 33

3.6 Geometry used in WFS formulations 34

3.7 Example of a multiple line array in WFS (Figure extracted from [5]) . 37

3.8 Truncation Effects for WFS reproduction of finite array with Δx = 0.1 m (a) Monochromatic plane wave signal with frequency 800 Hz (b) Low pass filtered pulse with cut-off frequency = 1,500 Hz 38
3.9 Same as Figure 3.8 with tapering applied at the extremes of the loudspeaker array 39
3.10 Tapering applied to N-shaped array and linear array 39
3.11 Enlarged listening area using additional side array with linear array for Monochromatic plane wave signal with frequency 800 Hz, Δx = 0.1 m (a) Linear array only in the front (b) Additional side array with linear array in the front 40
3.12 Spatial sampling of the loudspeaker array for a source far away 41
3.13 Plane wave reproduction with spatial aliasing, Δx = 0.2m (a) Monochromatic plane wave f = 1,500 Hz (b) Low pass filtered pulse with cut-off frequency = 2,200 HZ 42
3.14 A look at various WFS developments 44

4.1 Bruel & Kjaer dummy head with different type of headphones used for the HRTF measurements. 52
4.2 Effects of four types of headphones on the direct sound source spectrum at ear drum of the dummy head 53
4.3 (a) Proposed NAR headset structure (top) and prototype using open CA headphones with two microphones (below) (b) Headphone modified transfer functions (HMTF) (c) Transfer functions measurement set up 56
4.4 Measured modified transfer functions ($H_{int}(z)$and $H_{ext}(z)$) for two azimuths (Top: Ipsilateral ear; Bottom: Contralateral ear) 57
4.5 Case I: Only real source present scenario and corresponding signal flow block diagram 59

4.6 Case II: Only Virtual source present scenario and corresponding signal flow block diagram 61
4.7 Conventional FxNLMS Block Diagram for virtual source reproduction 63
4.8 Block Diagram of modified FxNLMS for virtual source reproduction . 64
4.9 Comparison between Conventional FxNLMS and Modified FxNLMS for 40^o azimuth (Top: Ipsilateral ear; Bottom: Contralateral ear) . . . 66
4.10 Spectral distortion comparisons for both the approaches over 0 to 180^o azimuths (Top: Ipsilateral ear; Bottom: Contralateral ear) . . . 68
4.11 Block diagram of hybrid FxNLMS using conventional FxNLMS and modified FxNLMS algorithm 69
4.12 Hybrid adaptive equalizer performance for 40^o azimuth (Top: Ipsilateral ear; Bottom: Contralateral ear) 72
4.13 Spectral distortion score comparisons: Hybrid FxNLMS versus others (Top: Ipsilateral ear; Bottom: Contralateral ear) 73
4.14 Spectral distortion score for Hybrid FxNLMS for elevated sources . . 74
4.15 Case III: Both virtual and real source present scenario and corresponding signal flow block diagram 75
4.16 Residual error plots for hybrid FxNLMS with and without real source. Virtual source is positioned at 0^o azimuth, while real sound is coming from 40^o azimuth and added to virtually reproduced signal at m_{int}. (Top: Ipsilateral ear; Bottom: Contralateral ear) 76
4.17 Block diagram of hybrid adaptive equalizer with online adaptive estimation of $r_{int}(n)$ 77
4.18 Results for hybrid FxNLMS with and without adaptive estimation. Simulation set up is kept same as in (Figure 4.16). (Top: Ipsilateral ear; Bottom: Contralateral ear) 78
4.19 Listening test setup (: Elevated speaker; : Azimuth speaker) . 80
4.20 Source confusion % for the three listening sets 82

4.21 Box plot showing subjective scores for sound similarity and source position similarity 84

5.1 Measurement set up for single channel continuous BRIR acquisition with head-tracking 93

5.2 System identification using NLMS [6] 94

5.3 Perfect sweep signal showing continuous transition from 1 period to another. $N = 2048$ samples 97

5.4 Residual errors (in green) for BRIR acquisition using NLMS for (a) Noisy captured signal (b) Noiseless signal with 10 repetitions of actual recorded signal 98

5.5 Environment noise of the measurement room recorded at subject's ear 98

5.6 Magnitude frequency responses of estimated BRIRs compared with exponential sine sweep method (a) Left ear (b) Right ear 99

5.7 Continuous BRIR estimation results with head rotation in clockwise direction (Top: Head orientation in clockwise direction; Middle: Residual error for left ear; Bottom: Residual error for right ear) . . . 100

5.8 Estimated BRTFs with continuous head rotation for 3 head azimuths (Left: 0^0; Middle: 14^0 ; Right: 27^0) 101

5.9 Proposed adaptive equalizer for NAR headset extended for non-stationary virtual signals 103

5.10 Results for the proposed adaptive equalizer with training period of 1 second using white noise signal (Top: $h_{hp}(n) = \widehat{h}_{hp}(n)$; Bottom: $h_{hp}(n) \neq \widehat{h}_{hp}(n)$) 104

5.11 Modified block diagram of adaptive equalizer with online detection of change in HPTF (highlighted in grey box) 106

5.12 Results of HPTF detection for two measured physical model of HPTFs using (Equation 5.11). (Window size of $W = 4096$ samples i.e., around 100 msec is used) 107

5.13 Modified block diagram of adaptive equalization of NAR headset with HPTF estimation. Training and playback phase is shown in grey colors indicating that virtual signal playback is stopped, while HPTF estimation is going on. 108

5.14 Results for the adaptive equalization of NAR headset with HPTF estimation and compared with the reference case as well as adaptive equalization without HPTF estimation (Top: Residual error plots for training phase white noise ; Bottom: Residual error plots for a virtual speech signal) Simulation is performed for azimuth: 40^o and ipsilateral ear. 110

5.15 Proposed adaptive equalizer with adaptive estimation of real signals for NAR headset extended for non-stationary virtual signals 113

5.16 Results for the proposed adaptive equalizer without and with presence of external signals: Virtual source is positioned at 0^o azimuth, while external sound is coming from 40^o azimuth and added to virtually reproduced signal at m_{int}
(Top: Ipsilateral ear ; Bottom: Contralateral ear) 114

5.17 Results for the proposed adaptive equalizer with adaptive estimation of external signals. Simulation set up is kept same as of Figure 5.16 and step-size (μ_r) for adaptive estimation is taken as 0.4.
(Top: Ipsilateral ear ; Bottom: Contralateral ear) 115

5.18 Modified adaptive equalizer with adaptive estimation for non-stationary virtual signals resolving causality issue between $r_{ext}(n)$ and $r_{int}(n)$ . . 116

5.19 Trade-off between MSE and step size for different delay values ($\triangle_r$). Simulation is performed for azimuth: 40^o and contralateral ear. 117

6.1 Proposed hybrid system in a home scenario 124

6.2 Proposed hybrid system structure (: Physical speakers; : Virtual speakers) 127

6.3 Overall hybrid system block diagram 128
6.4 WFS Sound field plots for monochromatic source with frequency 2,000 Hz. Green circle indicates the listener position and red circle indicates the virtual source position. (Top: Point source; Bottom: Focussed source) 130
6.5 WFS Sound field plots for monochromatic source with frequency 6,000 Hz. Listener area is indicated by green square. f_{al} =3,000 Hz (Top: Far-field non-focused source; Bottom: Focussed source) 131
6.6 Virtual WFS using binaural synthesis over NAR headset 132
6.7 Multichannel adaptive equalization for virtual WFS using NAR headset 133
6.8 Measurement setup for WFS rendering of three virtual WFS non-focused sources 135
6.9 Compensated virtual WFS frequency response Vs Measured physical WFS frequency response for frontal WFS array. Three virtual WFS sources (left, center and right) were rendered as shown in Figure 6.8. 136
6.10 Impulse response plots: Real Vs Virtual WFS (Black: Measured IR Grey: Estimated IR) 137
6.11 Headphone isolation compensation of NAR headset for frontal processing 139
6.12 Headphone isolation compensation filters estimation 139
6.13 Headphone isolation compensation results for a WFS virtual source 1 m behind the array 141
6.14 Hybrid WFS frontal playback: High frequency reproduction using virtual densely spaced speakers. Black circle represents the physical speakers with inter-spacing of 9 cm, while grey circle denotes the virtual speakers with inter-spacing of 1 cm. 142
6.15 Frontal playback processing block diagram using hybrid WFS method 142

6.16 Frequency responses illustration for frontal auditory scene processing using hybrid WFS method. WFS_9 in top-left figure represents the speaker array response with inter-spacing of 9 cm emulating phsyical frontal array. WFS_1 in top-right represents speaker array response with spacing of 1 cm emulating virtual WFS. Bottom left figure shows the low-pass spectra of WFS_9 and high-pass spectra of WFS_1. Finally, bottom right figure represents the combined hybrid WFS response. Frequency responses were simulated for a virtual source 0.5 m behind the frontal array with listener position at 0.5 m in front of the array. 143
6.17 Target azimuth directions of virtual sources used in listening test . . 145
6.18 Azimuth accuracy for frontal virtual sources: Mean and their 95% confidence intervals 146
6.19 Azimuth accuracy for rear and side virtual sources: Center line in the box represents the median value, while edges of the box are 25 and 75 percentiles responses. Top and bottom lines represent the extreme subject responses, while outliers are shown in red. 148
6.20 Mean externalization grades with 95% confidence interval for both frontal as well as rear and side sources 149
6.21 Mean elevation grades with 95% confidence interval for both frontal as well as rear and side sources 149
6.22 Mean locatedness grades with 95% confidence interval for both frontal as well as rear and side sources 150
6.23 Experiment setup for sound coloration test. Source positions are indicated by green circle, while red circle indicates the two listener positions. 151
6.24 Mean coloration grades with 95% confidence interval for 5 reproduction methods 152

6.25 Frequency spectra of drum beats stimulus versus pink noise stimulus 153

6.26 Overall audio quality grades: Mean scores with their 95% confidence interval 156

7.1 NVIDIA GPU architecture overview. Adapted and modified from [7] 164

7.2 Predicted data transfer time using (Equation 7.2) for PCIe bandiwdth of 8 GB/s corresponding to Nvidia Tesla C2075 GPU and $\alpha = 10\mu s$. 168

7.3 Real-time WFS processing framework with processing blocks (PB1, PB2 and PB3) 170

7.4 Block-wise processing of incoming source data and pre-filtering of one source ($M = 512$) : Stage (a) 172

7.5 Individual driving signals computation and its kernel matrix ($M = 512$) : Stage (b) 173

7.6 CUDA kernel execution times for different TB configurations ($Ns = 100$, $L = 161$, $M = 512$, $Ls = 1$;$l \times m = 256$) : Stage (b) 174

7.7 Reduction sum for computation of driving signals : Stage (c) 176

7.8 Percentage improvement in execution times due to different GPU optimization techniques over GPU non-optimized implementation for PB1+PB2 ($Ns = 100, L = 161, M = 512, Ls = 1$) 179

7.9 Average Execution times and Peak throughput of overall system (PB1 + PB2) ($M = 512$, $Ls = 1$) 180

8.1 Natural listening using NAR headset 190

List of Tables

4.1 Mean subjective scores along with their 99 percentile intervals for the three listening sets 87

6.1 Spectral distortion scores (dB) for the virtual WFS over headphones . 136

7.1 Computational Complexity of different computation stages in PB1 (MAD: Multiply/Addition; ADD: Addition) 171

7.2 Average execution times (msec) for WFS processing blocks ($Ns = 1$, $L = 161$, $M = 512$, $Ls = 1$, $dimx = dimz = 256$) 178

List of Abbreviations

ANC	Active noise control
AR	Augmented reality
ARA	Augmented reality audio
ARE	Augmented reality environment
BRIR	Binaural room impulse response
BRTF	Binaural room transfer function
CA	Circumaural
CARROUSO	Creating assessing and rendering in real-time of high quality audio visual environments
CUDA	Compute unified device arcitecture
DLMS	Delayed least mean square
FxDLMS	Filtered-x delayed least mean square or delayed FxLMS
FxLMS	Filtered-x least mean square
FxNLMS	Filtered-x normalized least mean square
GPU	Graphics processing unit
HAE	Hybrid adaptive equalizer

HMTF	Headphone modified transfer function
HOA	Higher order ambisonics
HPTF	Headphone transfer function
HRIR	Head-related impulse response
HRTF	Head-related transfer function
ILD	Interaural level difference
ITD	Interaural time difference
KIH	Kirchhoff helmholtz integral
LFE	Low frequency effects
LMS	Least mean square
MSE	Mean square error
MSPS	Meaga samples per second
MUSHRA	Multiple Stimuli with hidden reference and anchor
NAR	Natural augmented reality
NLMS	Normalized least mean square
PCA	Principal component analysis
PE	Power estimate
PFS	Personal field speaker
PSEQ	Perfect sequence
RIR	Room impulse response

SA	Supraaural
SC	Spectral cues
SD	Spectral distortion
SIMD	Single instruction multiple data
SS-MSE	Steady state mean square error
TB	Thread block
VBAP	Vector base amplitude panning
VRE	Virtual reality environment
WFS	Wave field synthesis

List of Symbols

θ	Azimuth angle
ϕ	Elevation angle
d	Distance from center of head
w	Angular frequency
$\overrightarrow{r}$	Position vector
$\triangle$	delay
P	Sound pressure
P_{ed}	Sound pressure at ear drum
P_{bec}	Sound pressure at blocked ear canal
P_{oec}	Sound pressure at open ear canal
P_c	Sound pressure at center of head
$HRTF(\theta, \phi, d, w, ear)$	HRTF as function of θ, ϕ, d, w and ear
$HPTF(w, ear)$	HPTF as function of w and ear
$P(\overrightarrow{r}, w)$	Sound pressure as function of $\overrightarrow{r}$ and w
$D_{WFS}(n, w)$	n^{th} WFS driving signal in frequency domain
$d_{WFS}(n, t)$	n^{th} WFS driving signal in time domain
Δx	Loudspeaker inter-spacing
f_{al}	Aliasing frequency
m_{int}	internal microphone
m_{ext}	external microphone
$H_{int}(z)$	HMTF at m_{int}

$H_{ext}(z)$	HMTF at m_{ext}
$H_{int}^{v}(z)$	HMTF at m_{int} for virtual sound reproduction
$H_{ext}^{v}(z)$	HMTF at m_{ext} for virtual sound reproduction
$H_{he}(z)$	Headphone-effect transfer function
$h_{hp}(n)$	Headphone impulse response
$\hat{h}_{hp}(n)$	Headphone impulse response estimate

Abstract

Sound plays an important role in our day-to-day activities. We inherently use it for interacting, listening to music, watching movies in home or cinemas, playing video games, having video-conferencing, etc. The main purpose of 3D audio reproduction is to emulate a natural listening experience to the user via playback devices. Sound can be reproduced at listeners' ears over either headphones or loudspeakers/loudspeaker array. However, the rendering of sound to be played back over headphones and loudspeakers are very different. It is important for these two playback methods to faithfully reproduce the sounds to provide listener a natural listening experience. Headphones are mainly used for private listening, while loudspeakers (or loudspeaker arrays) are meant for shared listening among a group of listeners. This thesis focuses on both the headphones and loudspeakers based reproduction mechanisms with emphasis on augmented and virtual reality applications, respectively.

The first part of the thesis investigates natural listening over headphones in augmented reality using adaptive filtering techniques. We developed a natural augmented reality (NAR) headset with two pairs of binaural microphones attached to open headphones (one internal and one external microphone on each side). This work focuses on enabling natural listening via adaptive equalization of headset to ensure that the virtual sounds are reproduced perceptually as close to real sounds as possible in any listener environment, while also being aware of the external sound sources. The key objective is to minimize the large localization errors (front-back

confusions), in-head localization as well as the timbre differences between virtual and real sounds. Modified adaptive filtering based on filtered-x normalized least mean square (FxNLMS) algorithms is proposed in this work to adapt the headphone synthesized signals to sound exactly like physical sounds, while equalizing for the individual headphone response. The adaptive equalization is further extended for the case when external sounds also present. Results show that the proposed adaptive algorithm approaches the desired response with minimum mean square error and converges faster than the conventional FxNLMS algorithm. The proposed method is found to be equally effective in the presence of external sounds. Subjective test using individualized binaural room-related impulse responses shows that listeners could not distinguish between the real and virtual sounds most of the times.

Next, we emphasize on the spatial sound reproduction in a home entertainment scenario using multi-channel loudspeaker setups. Spatial sound systems aim at creating realistic sound experience to the listeners with uniform sound fields in the entire listening area. Conventional surround sound systems, which are most widely used as home theater systems, are based on multi-channel stereophony, like 5.1, 10.2 and higher surround channel system. These systems require multiple loudspeakers to be placed in fixed configuration but often constrained by the room size and the best impression only achieved at the sweet spot. Sound field reproduction systems like wave field synthesis (WFS) is based on the principle of natural propagation of sound waves, and hence can create replica of true sound field uniformly over an extended listening area. WFS virtual sources are localized much accurately as compared to the stereophonic phantom sources. However, WFS based systems require hundreds of densely spaced loudspeakers enclosing the listener area and thus, difficult to realize in homes. In addition, practical approximations of WFS, such as finite, discrete and line array of loudspeakers limit, the performance of WFS with reduced listening area, sound coloration and horizontal plane only reproduction. Therefore, a combination of WFS and binaural synthesis over the NAR headset

is proposed to overcome the practical and physical limitations of the WFS. The proposed hybrid system enables frontal loudspeaker array playback using WFS, which provides strong frontal localization cues, while rear and side auditory scene is played back via NAR headset using virtual WFS to complete an entire 360^o auditory scene presentation. Furthermore, the use of virtual WFS over headphones helps in minimizing sound coloration above spatial aliasing frequency of physical array with the help of virtual densely spaced speaker array. Both objective and subjective experiments are carried out to evaluate the performance of the proposed setup. In particular, a detailed subjective study is carried out to investigate the performance of the proposed hybrid system with regard to sound localization and sound coloration.

Finally, a fast and efficient real-time GPU based implementation of WFS is presented to enhance the system throughput by exploiting the inherent massive parallelism in WFS based system comprising hundreds of densely spaced loudspeakers. The main goal of this work is to develop a real-time high throughput scalable platform for the hybrid WFS setup, which would need multiple driving signals, as well as WFS synthesized binaural signals using virtual WFS at the same time.

To summarize, in this thesis we aimed to reproduce natural listening over headphones for personal listening in augmented reality environment, as well as creating an immersive listening experience for user using WFS in home scenarios.

Chapter 1

Introduction

We are used to perceive sound in a three-dimensional (3D) world. In order to reproduce real world sound in an enclosed room or theater, extensive study on how spatial sound can be created, continued to be an active research topic for the past decades. Spatial audio is an illusion of creating sound objects that can be spatially positioned in a 3D space by passing original sound tracks through a sound rendering system and reproduced through multiple transducers, which are either distributed over the listening space or positioned very close to the listeners' ears. The reproduced sound field aims to achieve a perception of spaciousness and sense of directivity of the sound objects. Ideally, such a sound reproduction system should give listeners a sense of immersive 3D sound experience.

1.1 Spatial Audio Overview

Spatial audio can primarily be divided into three types of sound reproduction techniques, namely, loudspeaker stereophony, binaural technology, and reconstruction using synthesis of natural wave field (which includes Ambisonics and Wave Field Synthesis), as shown in Figure 1.1(a).

The history of spatial audio dates back to late 1800, with the very first inventions being the gramophone and phonograph [8] used in monophonic (only one channel)

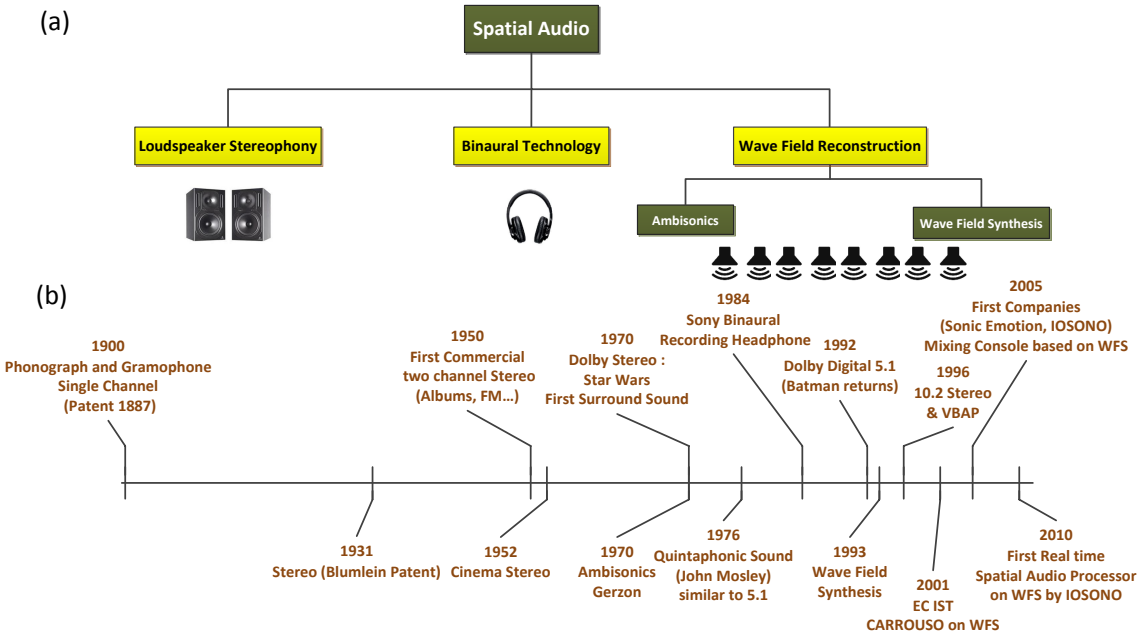


Figure 1.1: (a) Classification of Spatial audio and (b) Time-line of evolution of spatial audio [1]

sound recording. As shown in the time-line in Figure 1.1(b), there have been major advancements in terms of both technical and perceptual aspects in the last century. Spatial sound systems have evolved over the years from a two-channel stereo system to multichannel surround sound system. These surround systems are not only limited to cinemas and auditoriums, but are also being adapted in home entertainment systems. Conventional headphones, which employ a pair of small emitters, aims to produce high quality sound close to ears and they do not need to account for inaccuracies due to surroundings in contrast to loudspeakers. Nowadays, multiple emitters are embedded inside the ear cup to create a virtual surround sensation in 3D surround headphones. Modern electroacoustic systems [9, 10] have improved significantly with new functionalities to adapt or correct the sound field in a given room acoustic. Towards the end of the 19th century, new reproduction techniques like Ambisonics [11, 12], and Wave Field Synthesis (WFS) [13] (Figure 1.1(b)), which uses the principle behind physical sound wave propagation in air and provide true sound experience in any environment, were introduced to overcome the limitations

of stereo systems.

Two-channel stereophony is the oldest and simplest audio technology, first patented by Blumlein [14], which has been progressively extended to multichannel stereophony systems, through 5.1, 7.1, 10.2, and 22.1 surround sound systems (Note that in the x.y surround sound format representation, x indicates the number of full bandwidth channel and y indicates the number of low frequency channels, known as low frequency effects (LFE) sub channel). These multichannel systems have been widely used in cinema, home entertainment, and gaming to create an immersive surround sound experience. Figure 1.2 shows a typical setup of a 5.1 stereo system with 3 front and 2 rear loudspeakers. It uses rear speakers to enhance the ambience sound quality and center speaker to enhance the frontal perception. The disadvantages of multichannel stereophony system are the localization of phantom sources and sweet spot. In other words, a phantom source can only be located along the lines connecting two loudspeakers and listeners will only be able to experience the best surround sound effect at the sweet spot or focal point of all multichannel speakers. Further advancement of the two channel stereophony was developed as Vector Base Amplitude Panning (VBAP) by Pulkki [15] to spatial audio with multiple loudspeakers in an arbitrary two or three dimensional placement. This enables virtual sound source positioning anywhere in 3D space on the active triangle formed by three loudspeakers.

Binaural technology is another approach to reproduce sound signals naturally, also known as the Binaural Stereo [2, 16]. Binaural technology consists of recording, as well as reproduction of natural sound scenes at the two ears. Sound signals are recorded using a pair of microphones positioned inside the ears of dummy head or inside the ear canal of the actual human listener. Several patents can be found in the literature describing the binaural synthesis and methods to create 3D auditory display [17–19]. Recorded sounds can be reproduced accurately at the ears by filtering the source signals using acoustic transfer functions between source location

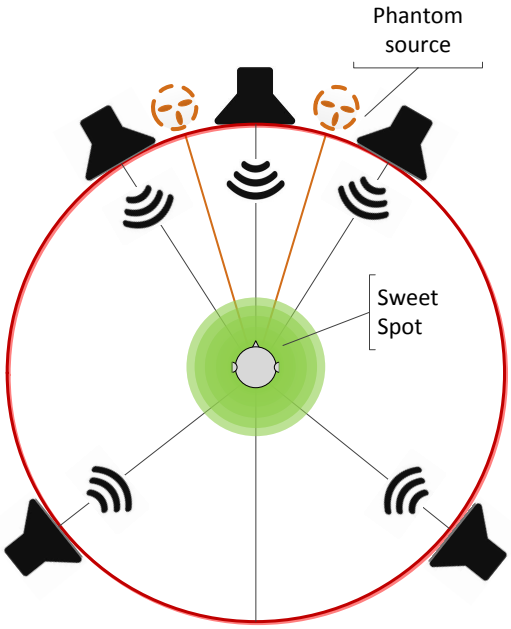


Figure 1.2: Typical 5.1 Stereo system

and both ears, which are popularly known as the head related transfer function (HRTF) [20]. HRTF contains three important cues of interaural time differences (ITD), interaural level differences (ILD), and spectral cues (SC). These cues are essential for us to correctly localize and perceptually visualize the sound scenes. ITD and ILD cues are respectively, the time and level differences at the listeners' ears in accordance with Rayleigh's duplex theory [21], which states that low frequency sounds are localized using time cues, while high frequency sounds are localized using interaural level cues [22]. Spectral cues account for spectral modification due to sound interaction of external ear parts like pinna, concha etc. As a result, binaural reconstruction can produce excellent spatial awareness and sound color under given circumstances. Binaural signals can be played back via loudspeaker or headphones. Direct reproduction of binaural signals through loudspeakers suffers from the problem of cross-talk between the left and right ear signals. A cross-talk cancellation system must be inserted between the loudspeakers and binaural processing in order to achieve accurate 3D audio display. Binaural reproduction using headphones are the most efficient way, as the signals are correctly reproduced at each ear and do not

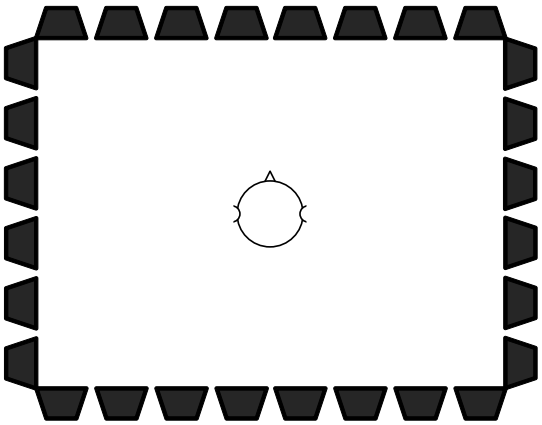


Figure 1.3: Wave Field Synthesis using loudspeaker arrays shown in black and entire interior is the listening zone

suffer from any distortion due to environments. But, there have been several inherent limitations of binaural sound reproduction through headphones, which includes front-back confusions, in head localization, and incorrect perception of the elevation of virtual sound sources [16]. These limitations of binaural technology are due to highly idiosyncratic ear features that results in confusion and human beings are unable to disambiguate between ITD, ILD, and SC cues when listening to binaurally synthesized signal derived from generic or non-individualized HRTF representations [23–26] based on manikin recordings or anthropometric data.

Both multichannel stereophony and binaural technology are widely used in cinemas, auditoriums, home entertainments and headphones playback. However, their inherent limitations, which rely mainly on psychoacoustic principles in creating a fully immersive environment, have inspired researchers to look into a more natural ways of reproducing 3D sounds. The two technologies, which use the concept of natural propagation of wave fields, are Ambisonics and WFS. In contrast to stereophony and headphones playback, the wave-field based approach use holographic principles to synthesize true sound field rather than relying on psychoacoustic principles for recreating sound scenes. With the help of loudspeaker arrays, both Ambisonics and WFS are able to synthesize natural sound environment in an enlarged listening area with perfect sound source localization. Ambisonics was first proposed by Gerzon

[12] in 1970, while Berkhout [13] invented WFS in 1988. Although both approaches follow the same basic principles but the difference lies in the detailed mathematical derivations. Traditional Ambisonics approach, which is based on spherical harmonics function and limited to four channels, has been extended to higher orders enabling use of unlimited loudspeakers in an arbitrary placement providing high spatial resolution as well as increasing the size of sweep spot. There are different variants of the higher order Ambisonics (HOA) [27–31] exist in the literature. There also have been recent advancements in spherical harmonics based sound field for 2D and 3D sound reproduction [32–39]. To practically realize the 3D sound field reproduction in real rooms, theater and auditoriums, multiple circular loudspeaker arrays were employed over limited region of interest providing a flexible loudspeaker array placement [40, 41]. The main advantage of Ambisonics is that it can synthesize exact sound field for any number of loudspeakers arranged in circular/spherical geometry. However, the computation of driving signals for loudspeakers in Ambisonics is highly complex due to the spherical harmonics functions and also depending on particular geometry, while driving functions for WFS can be realized very efficiently as weight and delay[42, 43]. Detailed comparison between HOA and WFS can be found in [42, 44]. According to Francis Rumsey [8], "*Ambisonics being mainly a collection of elegant principles and signal representation forms rather than a particular implementation.*" Few works have been reported in the recent past to standardize the HOA as sound field coding format for flexible speaker playback layouts and to facilitate the exchange of HOA data [45–47]. Extensive research is being carried out recently in developing a universal 3D audio standard MPEG-H, which supports many types of speaker setups like stereo, 5.1, 22.1, binaural reproduction, reproduction setups using object based coding as well as HOA content. Broadcasters and corporates like BBC and Dolby atmos have expressed interest in Ambisonics and are actively doing research to bring this technology into market.

In WFS, secondary sound sources (responsible for reproduction of sound field

produced by real sources) create a replica of true sound field in an extended listening area while retaining spatial as well as temporal aspects of physical sound field using the acoustic holography approach [48]. WFS based reproduction system aims to accurately synthesize the sound field within the entire listening space, i.e., sweet spot is everywhere in the listening room (Figure 1.3). WFS duplicate the sound field generated by primary sources (real sources which produce the sound) with the help of loudspeaker arrays acting as secondary sources in an enlarged listening area. Source localization is possible anywhere in the physical space, which is only limited by the extent of visible area covered by configuration of loudspeaker arrays and listener. Also unlike binaural listening, virtual source does not move with listener's movement. Listeners feel as if they are in a real environment and sounds appear to come from where they are meant to be.

1.2 Enabling Natural Listening over Headphones and Loudspeakers

As we discussed in Section 1.1, sound can be reconstructed at the listener's ears either via headphones or loudspeakers (or loudspeaker arrays). Headphones, with the help of two emitters, directly reproduce 3D sound close to ears and are widely used in personal audio applications. Loudspeakers (or loudspeaker arrays), on the other hand, reproduce immersive sound experience for several listeners and are widely used in cinemas, theater or home entertainment systems. Both the playback devices renders sound differently and their perceptual properties are also fundamentally different.

The main objective of this research is: *To enable natural listening experience to the users over both headphones and loudspeakers via different reproduction mechanisms.* Natural listening can be realized in either an augmented reality environment (ARE) or virtual reality environment (VRE). Both augmented and virtual reality

aim to immerse listeners in the presented auditory environment, though they achieve this via contrasting approaches. Augmented reality blend virtual sound objects into the real world and allows interaction with them seamlessly, while being continuously aware of the physical world. Virtual reality immerses the user in virtual auditory environment, which is different from real world. Therefore, listeners feel as if they were teleported to the virtual environment such that virtual objects are perceived similar to how we hear sound in real world.

Augmented reality (AR), which composes of virtual and real world environments, is becoming one of the major topics of research interest due to the advent of wearable devices. Today, AR is commonly used as assistive display to enhance the perception of reality in education, gaming, navigation, sports, entertainment, simulators, etc. However, most of the past works have mainly concentrated on the visual aspects of AR. Auditory events are one of the essential components in human perception in daily life but the augmented reality solutions have been lacking in this regard till now compared to visual aspects. Therefore, there is a need for natural listening in AR systems to give a holistic experience to the user. For this reason, am important objective of this thesis is on the natural listening in an ARE for headphones playback as against virtual reality, where listeners are isolated from the real world. The main challenges in headphones playback via binaural synthesis is due to the use of non-individualized HRTFs resulting in:

- Front-back/Up-down confusion
- In-head localization
- Sonic difference between real and virtual sounds

In addition, headphone transfer function (HPTF), which is the electro-acoustical transfer function from headphone emitter to listeners' ears, modifies the intended spectrum and may result in unnatural listening experience. Since HPTFs are also unique to every individual, headphone equalization must be accustomed to every

listener for accurate reproduction of virtual sounds. Furthermore, room reverberations and dynamic head movement cues are also very important for externalization of virtual sources. In this work, a natural augmented reality (NAR) headset is developed using open headphones employing adaptive filtering techniques to achieve natural listening. Open headphones are used to allow the external sounds to pass through the headphones without much attenuation. Adaptive filtering with the help of sensing microphones attached to the ear cup ensures that the virtual sound is reproduced as close as possible to real sound, while equalizing for any change in the individual HPTF due to the re-positioning of headphones. Adaptive equalization is extended for the augmented reality mode, i.e. when virtual sounds are reproduced in the presence of external sounds.

Loudspeakers based reproduction system aims at engrossing the listener into virtual acoustic scene, especially for multimedia applications such as gaming, movies. These systems are becoming more popular in domestic use with the advancement in multichannel stereo setup, where a sense of virtual auditory environment is created with the help of many loudspeakers encompassing the listening space. However, these multichannel surround sound setup require loudspeakers to be arranged in fixed configuration and is often constrained by the room size. Additionally, listener movements are restricted to only the sweet spot, i.e., center of the listening area. The second aim is to explore the use of WFS based setup in home entertainment scenarios to realize natural listening experience for virtual reality applications. WFS based on holographic sound reproduction recreates sound similar to natural sound propagation and thus, can provide perfect sound field reproduction with the help of densely spaced loudspeaker arrays. However, WFS based systems too have several limitations in practical implementations:

- Reduction of listening area: Due to the finite length of the loudspeaker array, listener area is reduced.

- Horizontal plane reproduction: Due to the impracticality of placing loudspeak-

ers everywhere in the 3D plane, virtual sources can only be reproduced correctly in the horizontal plane.

- Sound coloration of sound field: Reconstructed sound field is only correct up to aliasing frequency due to the spatial sampling of the loudspeaker array. Thus, widely spaced loudspeaker arrays suffer from severe spatial aliasing of the sound field.

- Too many loudspeakers: WFS requires many loudspeakers to entirely enclose the listening area, which is difficult to realize in practice, especially in a big living room or homes.

We also investigate the perceptual properties of headphones reproduction of virtual WFS and compare with the actual WFS based loudspeaker system. A virtual WFS system with densely spaced loudspeaker array can achieve very high fidelity sound reproduction with no spatial aliasing and is only limited by the computational capacity of the system. NAR headset is employed to emulate the virtual WFS over headphones. Finally, a hybrid speaker array-headphones system is proposed by combining WFS using loudspeaker array and binaural synthesis using NAR headset to overcome physical and practical limitations of the WFS.

1.3 Contribution of the Thesis

The main contributions of this thesis are listed below:

1. A new headphones configuration (NAR headset) is presented with two pairs of binaural microphones (one internal and one external) for natural listening. Subjective test based on individualized binaural room transfer functions (BRTFs) reveals that the subjects cannot distinguish between real and virtual sounds.

2. Adaptive headphone equalization is proposed for the NAR headset using the two microphones. Modified versions of the filtered-x normalized least mean square algorithm (FxNLMS) are proposed to achieve a faster convergence with optimum mean square error (MSE) as compared to the conventional FxNLMS. The main advantage of adaptive equalization is that it compensates for any variations in individual headphone response as compared to generic and imperfect headphone equalization in conventional headphones.

3. The main goal of the NAR headset is to interact with the virtual sound objects in the presence of physical sound sources. Therefore, online adaptive estimation of the external sounds is introduced to ensure optimum convergence of the adaptive equalization and virtual sources are coherently blended with the real world environments.

4. Fast and continuous measurement of individualized BRTFs for human subjects is developed based on normalized least mean square (NLMS) technique using perfect sweep sequences [49, 50]. Head tracking is further employed to incorporate dynamic head movement cues for NAR headset.

5. We address some of the practical limitations of the NAR headset:

   (a) Adaptive equalization for non-stationary virtual signals.

   (b) Online detection and fast estimation of secondary path headphone response.

   (c) Resolving causality issues in adaptive estimation of external signals.

6. Next generation entertainment system is presented by combining WFS loudspeaker array mounted on a television set for the frontal auditory scene, while NAR headset using virtual WFS array reproduce the rear and side auditory scenes. Measurement results show that temporal and spatial characteristics of the pure WFS based system is retained in the reconstructed virtual sound

field using NAR headset. Subjective study investigates the performance of the hybrid system in comparison to pure loudspeaker based WFS system or pure headphones based system with regards to sound localization, sound coloration, and overall audio quality.

7. Frontal auditory scene reproduction using WFS is enhanced by reproducing the high frequency content of original sound tracks above aliasing frequency over headphones, while low frequency contents are rendered using physical frontal array.

8. Multichannel version of the earlier proposed adaptive equalization is developed for virtual WFS reproduction over the NAR headset. The main purpose of the multichannel adaptive equalization is adaptive equalization of all the virtual WFS speakers in an enclosed WFS setup simultaneously.

9. Finally, a fast and efficient real-time GPU based implementation of WFS is presented for rendering of multiple virtual sources in a multiple-speaker multiple-listener set up. The main purpose of GPU implementation is to exploit the massive inherent parallelism in WFS to enhance system throughput.

1.4 Structure of the Thesis

The thesis consists of eight chapters, and its organization is as follows:

Chapter 2: This chapter presents the comprehensive review of binaural technology. In particular, spatial cues for sound source localization are discussed. Furthermore, we present existing techniques to synthesize binaural signals at listener's eardrum over headphones.

Chapter 3: In this chapter, we give a detailed survey of WFS theory and the mathematical formulations derived in the literature. Practical constraints of the

WFS and their existing solutions are also discussed. We conclude the chapter with future trends of WFS in commercial as well as home entertainment systems.

Chapter 4: In this chapter, a natural augmented reality (NAR) headset is introduced for AR applications using adaptive filtering techniques. Adaptive equalization techniques based on FxNLMS are introduced in this chapter. Three working modes of the NAR headset is discussed: 1) Only real sources present 2) Only virtual sources present, and 3) Both virtual and real sources present (Augmented reality mode). Both objective and subjective experiments to evaluate the performance of NAR headset are discussed.

Chapter 5: Practical limitations, solutions and extensions of the NAR headset are presented in this chapter. Fast continuous BRIRs acquisition for human subjects in dynamic scenarios based on NLMS algorithm is presented. Furthermore, an extension of the adaptive equalization techniques for NAR headset is presented with simulation results.

Chapter 6: A hybrid speaker array-headphone system for immersive audio reproduction in home scenarios is proposed in this chapter. Synthesis of virtual WFS over headphones is carried out using a multichannel version of the adaptive equalization of NAR headset. Hybrid WFS method is further introduced to enhance frontal auditory scene perception. A detailed subjective investigation for evaluation of the performance of the proposed methods is also presented.

Chapter 7: This chapter presents a fast and efficient real-time GPU implementation of a three-fold WFS framework. Different GPU-CPU optimization technique are presented to enhance the overall system throughput.

Chapter 8: Finally, the conclusions and future works of this thesis are described in this chapter.

Chapter 2

Binaural Technology : A Literature Review

With the help of just two ears, we are able to acquire all the auditory information about incoming sounds, such as distance, direction based on the time and level difference between the sounds received at the two ears. Thus, if the two ears' recorded signals, referred as binaural signals, can be reproduced exactly the same as in the direct listening case, a perfect replica of true auditory scene can be synthesized and subsequently, natural listening is attained. Binaural Technology, is more generally defined as *"a body of methods that involve the acoustic input signals to both ears of the listener for achieving practical purposes, e.g., by recording, analyzing, synthesizing, processing, presenting and evaluating such signals"*[51]. In this chapter, we present literature review of binaural technology and is organized as follows. Section 2.1 give a brief overview on head-related transfer functions followed by the sound localization cues in Section 2.2. Existing techniques on binaural synthesis over headphones and headphone equalization is presented in Section 2.3 and Section 2.4. Section 2.5 concludes the chapter.

2.1 Head-Related Transfer Functions

HRTF plays a significant role in localizing the sound image accurately. HRTF filters the source signal to account for the propagation of sound from the source to the

listeners' ears in a free-field listening environment. HRTF or head-related impulse response (HRIR), which is the time domain form of HRTF, is represented as pair for left and right ear. They consist of all the spatial characteristics related to source localization in 3D space and more importantly, reflections and refractions due to head, torso and pinnae of the listener [2, 16]. Typically, HRTFs are measured at the microphones positioned at ear drums of the acoustic simulator of dummy head or using probe microphones placed in the ear canal for human subjects. However, it is practically inconvenient to insert a miniature microphone near to the ear drum of the human subject. Fortunately, it has been found that the acoustic transfer function between a point at the entrance of the ear canal or inside the ear canal is independent of the source position [3, 52]. Therefore, acoustic transfer function measured from a point source to a point at entrance of blocked or open ear canal are also considered as HRTFs. Therefore, three widely used definitions of the HRTF can be written as:

$$HRTF_{ear\,drum}(\theta, \phi, d, w, ear) = \frac{P_{ed}}{P_c}, \tag{2.1}$$

$$HRTF_{blocked\,ear\,canal}(\theta, \phi, d, w, ear) = \frac{P_{bec}}{P_c}, \tag{2.2}$$

$$HRTF_{open\,ear\,canal}(\theta, \phi, d, w, ear) = \frac{P_{oec}}{P_c}, \tag{2.3}$$

with

$P_{ed} = Sound\,pressure\,at\,ear\,drum;$

$P_{bec} = Sound\,pressure\,at\,blocked\,ear\,canal;$

$P_{oec} = Sound\,pressure\,at\,open\,ear\,canal;$

$P_c = Sound\,pressure\,at\,center\,of\,head\,without\,head;$

where, θ, ϕ, d and w denotes the azimuth, elevation, distance and angular frequency. Sound transmission from point source to ear drum (2.1) can be split into direction

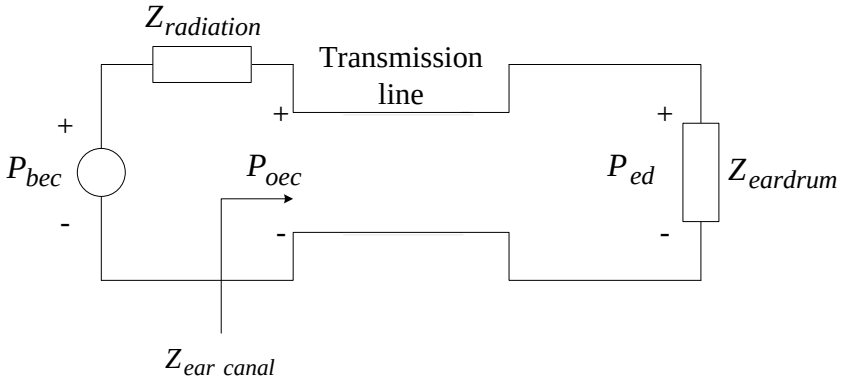


Figure 2.1: Free-field sound transmission model - adapted from [2]

dependent parts (2.2 or 2.3) and direction independent parts as depicted in the free-field model in Figure 2.1 as:

$$\frac{P_{ed}}{P_c} = \frac{P_{bec}}{P_c}\,\frac{P_{oec}}{P_{bec}}\,\frac{P_{ed}}{P_{oec}}. \tag{2.4}$$

$$\frac{P_{ed}}{P_c} = \frac{P_{oec}}{P_c}\,\frac{P_{ed}}{P_{oec}}. \tag{2.5}$$

The ratios P_{oec}/P_{bec} and P_{ed}/P_{oec} are direction independent and represent the pressure division ratio of sound pressures at ear canal entrance and acoustic transfer function along the ear canal, respectively. Although, neither of these ratios depend on the source location but are highly idiosyncratic [3]. Hence, HRTFs can be also be recorded either at blocked ear canal or open ear canal entrance. Blocked ear canal measurements (2.2) are most widely used on subjects as it provides good repeatability over open ear canal entrance measurements. However, in this thesis, we use open ear canal HRTF measurements (2.3) for human subjects as P_{oec} exists physically in case of natural listening situation, while P_{bec} exists only at the time of measurement when the ear canal is blocked. Moller [2] studied the transfer function from different points in the ear canal to ear drum using probe microphones on four subjects. The transfer function results for one typical subject is shown in

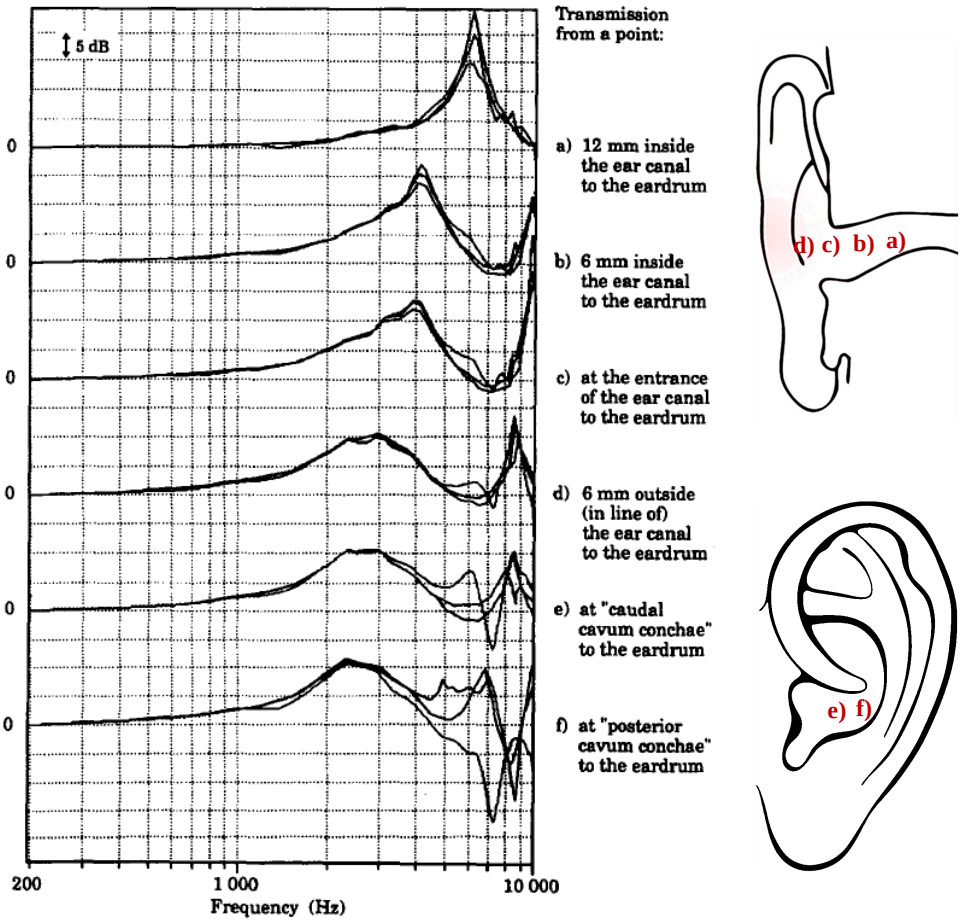


Figure 2.2: Transfer function along the ear canal from different points for three azimuths. Approximate measurement positions are shown in the ear diagram. Extracted from [2] to better illustrate the transfer function.

Figure 2.2 for three azimuths ($\theta = 0^o$, $90^o$and 180^o). It is found that any point from few millimeters outside of ear entrance to ear drum can be considered for HRTF measurements as acoustical transfer function of these points along ear canal were direction independent. Similar observations were reported by Middlebrooks *et al.* [52] for a total 356 source directions corresponding to loudspeaker positions in the median and azimuthal plane. Due to the practical in-feasibility of measuring HRTFs on human subjects, there have been several works in the literature to model the HRTFs from discrete finite set of measurements based on spherical harmonics [23, 53, 54], primary component analysis [25], and Fourier Spherical Bessel series [24, 55] and represent them using reduced number of parameters.

2.2 Sound Localization

HRTF comprises of three main binaural cues, namely, interaural time differences (ITD), interaural level differences (ILD) and spectral cues (SC) [2, 16]. Localization cues can mainly be divided into inter-aural cues, which depend on the difference between the two ear signals, and mono-aural spectral cues, which are attributed to the modification of source spectrum by the human body, torso, head and pinnae. Besides these three primary cues, head movement and room reflections also aid in localizing the sound sources in real life.

2.2.1 Inter-aural Cues (ITD and ILD)

According to Rayleigh duplex theory of directional sound localization, low frequencies are localized by time differences, while high frequencies above 1.5 kHz are localized using level differences [21]. Sound from the source propagates through the medium (air) and reaches the two ears at different time due to different acoustical path. Sound coming from one side of the head also get attenuated after reaching opposite ear as head acts as an acoustic obstacle. Thus, cues due to the difference between arrival of sound at two ears is known as ITD cues, while level differences between the sounds is called ILD cues.

ITD is the dominant cue in low frequency below 1.5 kHz, where head shadowing is weak as sound wavelength is more than the distance between the two ears, while from 1.5 kHz to 6 kHz, ILD prevails over ITD and helps in localizing the sound more accurately [56, 57]. Study by Wightman and Kistler [22] had shown that any stimulus with low frequency contents, auditory position of source is determined by ITD irrespective of the ILD. ITD is usually estimated by taking cross-correlation of low-pass filtered HRIRs (cut-off frequency of 2 kHz) and taking maximum value of the lags. ITD for a spherical head model is defined as [58]:

$$ITD = \frac{D}{2c}\left(\arcsin\left(\cos\phi\sin\theta\right) + \cos\phi\sin\theta\right), \tag{2.6}$$

where D is the diameter of head (roughly taken as 17.5 cm), and c is the speed of sound equal to 344 m/s. ITD can also be computed from the phase differences between the two ears as a function of frequency. It has been found that sensitivity of ITD is highest in low frequencies below 1.5 kHz and is 3/2 times more sensititve compared to that of ILD above 1.5 kHz. ILD is defined as the spectral magnitude difference between the left and right HRTFs. ILD is also defined as energy ratio of HRTFs for a limited frequency range (generally chosen between 1 kHz and 5 kHz) [59]:

$$ILD = 10\log_{10}\left(\frac{\int_{f1}^{f2}\left|HRTF_L(f)\right|^2 df}{\int_{f1}^{f2}\left|HRTF_R(f)\right|^2 df}\right) \tag{2.7}$$

ILD is also estimated as a function of frequency using energy ratios of the left and right auditory filters outputs $h_{l,w}$ and $h_{r,w}$ in each frequency band with center frequency w and for each azimuth θ as [60]:

$$ILD\left(w,\theta\right) = 10\log_{10}\left(\frac{\int_{-\infty}^{\infty}\left|h_{l,w}(\theta,t)\right|^2 dt}{\int_{-\infty}^{\infty}\left|h_{r,w}(\theta,t)\right|^2 dt}\right) \tag{2.8}$$

Auditory filter outputs, $h_{l,w}$ and $h_{r,w}$ in (2.8) are computed using the Auditory Modelling Toolbox [61] with filter spacing in terms of equivalent rectangular bandwidth. For wavelengths much less than the interaural distance ($f > 3$ kHz), head acts as a reflecting surface for sound coming from the side resulting in pressure gain as high as 6 dB than the pressure in absence of head. ILDs sensitivity is low when sound wavelength is much greater than the interaural distance. ILDs can increase from 0 dB to up to 20 dB for source in front to one side of the head [52]. It has also been shown that ILD vary considerably for sound sources close to head (<0.5

m) and frequencies less than 500 Hz [62, 63].

ITD and ILD together can not determine the source location if source is positioned on surface with equally possible ITD and ILD, popularly known as cone of confusion. As a result, confusion error results mainly in front-back confusions as well as incorrect elevation localization for fixed listening position [22]. Up-down confusions were also observed, although much less frequent than than front-back confusions [64]. Spectral cues in general are used for discrimination between front-back and elevated sources.

2.2.2 Spectral Cues

Listeners use monoaural spectral cues (SC) to resolve the ambiguity of sound source location on cone of confusion. Pinnae cues result in unique spectral peaks and notches in HRTFs for different source direction and are therefore, the most studied cues for localization. Several studies on sound localization reveal that the spectral modifications contribute significantly for both elevation localization and front-back confusions [65–67]. Additionally, SC cues are potentially useful for discriminating the sources in the front and back in horizontal plane. Spectral content in the 10-16 kHz region for frontal sources are amplified at least 10-15 dB more than the sources in the behind. This is mainly due to the pinnae being directed towards front, resulting in frontal sources being more emphasized. Hebrank *et al.* [68] carried out experiments to investigate the localization based on different frequency bands of source spectra. It has been observed that frequency range of 4 to 16 kHz are necessary of accurate source localization, which means that sound source should be broadband. SC cues for frontal sources were accompanied by 1-octave bandwidth notch with low cut-off frequency between 4 and 8 kHz. The exact location and depth of notch depends on the pinnae shape, size and of course, source direction. Notch for elevated sources were observed to be less emphasized than the frontal sources in horizontal plane. It was observed that spectral peaks around 8 kHz is an important

cue of elevated sources.

2.2.3 Individualized HRTFs

Cues due to the pinnae reflections are unique owing to the different pinnae structures in each individual and therefore, pose a special problem in binaural synthesis using headphones. Hence, human auditory perception is strongly dependent on the anthropometry of individual ear and head, especially on the unique pinnae shape of every individual. Using non-individualized HRTFs may result in front-back confusions, elevation localization errors and in-head localization, which are attributed to the considerable variation of high frequency SC across human subjects. HRTFs can also be individualized based on statistical methods using PCA analysis [69, 70], anthropometric data [71, 72], subjective tuning from a large existing HRTFs database [73–75]. Due to the impracticality of measuring individual HRTFs in an anechoic room on human subjects, generic HRTFs are widely used in binaural synthesis based on recording from a manikin of average anthropometric data. However, it has been found that use of generic HRTFs is fundamentally limited by the absence of individual attributes and only partially improves the front-back confusions and in-head localization [76]. Recently, there also have been several works on fast measurement of individual HRTFs [50, 77–80], which can substantially improve the binaural playback performance in terms of localization errors and front-back confusions.

2.2.4 Other Cues

In addition, dynamic cues due to the head motion and visual cues help in minimizing the front-back confusions and large localization errors [81]. Since most of the HRTF measurements are carried out in anechoic chamber as free-field, they are far from natural when played back over the headphones. In this context, room reflections are very important for source localization of distant sources in a reverberant room. Use of artificial or real room reverberation can thus, help in synthesizing externalized

sound image in headphone listening even with non-individualized HRTFs [82–84]. It has been observed in recent research that motion cues dominate pinnae cues in resolving the front-back reversals as well as enhancing externalization [63, 85, 86]. Dynamic cues can be generated with the help of head-tracking device attached to a headphone by continuously adapting to the head movements.

2.3 Binaural Synthesis over Headphones

As discussed above, listener makes use of following cues from the received binaural signals in natural listening:

- Interaural and spectral cues contained in individualized HRTFs for accurate sound source localization.
- Early room reflections and reverberations for out of head localization as well as naturalness of the sound.
- Dynamic cues using head movement eliminates the ambiguity of source positions on cone of confusion, especially in the absence of individualized HRTFs.

The main purpose of binaural synthesis is to replicate natural listening experience over headphones by reconstructing the binaural signals at listener's ears. Therefore, individualized BRTF must be recorded/synthesized before playing over headphones. In addition, binaural synthesis should also incorporate the head movement effect for natural synthesis of the virtual/augmented reality scene. One main advantage of binaural reproduction over headphones is that each channel is separately sent to each ear and it doesn't interfere with the room reflections as against loudspeakers, where cross-talk cancellation is needed.

However, headphones influence the intended desired spectrum at listener's ear and must be compensated. Headphones influence is described by the sound transmission from headphones transducer to a point in the ear canal and commonly known

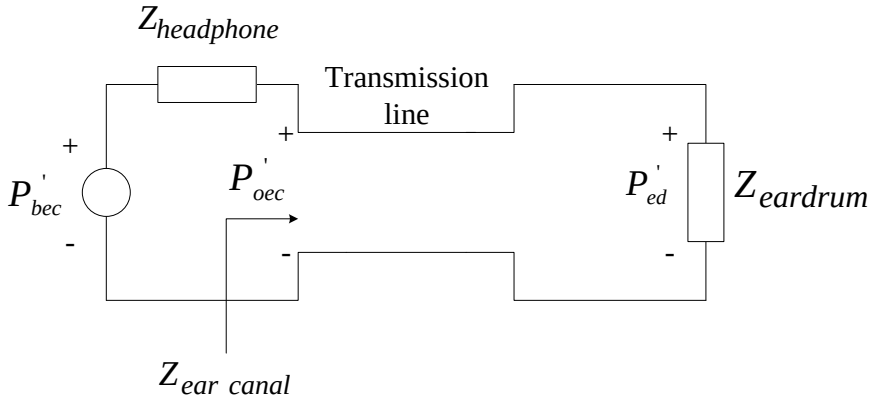


Figure 2.3: Headphone sound transmission model-adapted from [3]

as HPTF. Similar to the three definitions of HRTF, HPTF can also be measured either at the ear drum or at entrance of the blocked or open ear canal as:

$$HPTF_{ear\,drum}(w, ear) = \frac{P'_{ed}}{E'}, \tag{2.9}$$

$$HPTF_{blocked\,ear\,canal}(w, ear) = \frac{P'_{bec}}{E'}, \tag{2.10}$$

$$HPTF_{open\,ear\,canal}(w, ear) = \frac{P'_{oec}}{E'}, \tag{2.11}$$

where

$P'_{ed} = Sound\,pressure\,at\,ear\,drum\,from\,headphones;$

$P'_{bec} = Sound\,pressure\,at\,blocked\,ear\,canal\,entrance\,from\,headphones;$

$P'_{oec} = Sound\,pressure\,at\,open\,ear\,canal\,entrance\,from\,headphones;$

$E' = Transmitted\,signal\,at\,the\,headphone\,terminal;$

Symbol ($'$) represent the sound transmission due to headphones. Sound transmission from headphone to ear drum can also be split similar to the transmission in free-field in terms of HPTFs at blocked or open ear canal entrance (see Figure 2.3) as:

$$\frac{P'_{ed}}{E'} = \frac{P'_{bec}}{E'}\,\frac{P'_{oec}}{P'_{bec}}\,\frac{P'_{ed}}{P'_{oec}}, \tag{2.12}$$

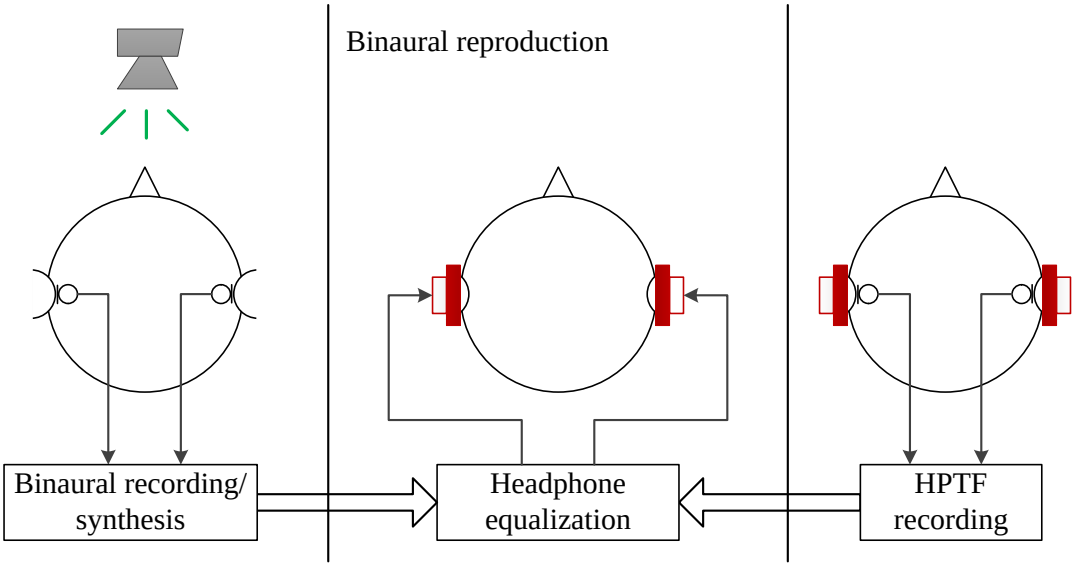


Figure 2.4: Binaural recording/synthesis and reproduction over headphones

$$\frac{P'_{ed}}{E'} = \frac{P'_{oec}}{E'} \frac{P'_{ed}}{P'_{oec}}, \tag{2.13}$$

where P'_{ed}/P'_{oec} is only dependent on the ear-canal transfer function and is equal to the pressure ratio P_{ed}/P_{oec} for the free-field radiation case. Thus, HPTFs can be measured at either of the three points in external ear (ear drum, blocked or open ear canal entrance). It should be noted that HPTF depends on subject and headphone as the acoustic propagation includes the reflections due to the pinnae and headphones. Unlike HRTFs, HPTFs doesn't depend on distance or direction of the virtual source, but it does correspond to a known position of the headphone transducer relative to ear. Furthermore, HPTF also depends on the type of headphones and their fittings. Therefore, it is important to apply a individual headphone equalization filter to negate the effect of HPTF on binaural signals and ensure that headphone does not degrade the spatial perception of the intended sound.

2.4 Headphone Equalization

Binaural reproduction comprises of two steps: recording/synthesis of binaural signals and rendering of binaural signals over headphones, as shown in Figure 2.4. Binaural signals are either recorded directly at listener's ears or synthesized by filtering the monoaural audio stream with the recorded HRTF/BRTF set. The binaural signals are then rendered over headphones after headphone equalization. The main goal of binaural reproduction is to ensure sound pressure at ear drum is same as in case of free-field case i.e., P_{ed}/P_c. It is already discussed that measurement can be done using the following three recording methods:

1. Ear drum,
2. Blocked ear canal entrance, and
3. Open ear-canal entrance

Free-field measurements using above three methods contain all the spatial information and thus, any of the them can be appropriately used to render natural listening over headphones. Based on the three recording methods, headphone equalization filter is estimated for respective HPTFs as ([87], (2.1-2.5), (2.9-2.13)):

$$W_{ear\,drum} = \frac{[P_{ed}/P_c]}{[HRTF \centerdot HPTF]_{ear\,drum}} = \frac{1}{HPTF_{ear\,drum}}. \tag{2.14}$$

$$\begin{aligned} W_{blocked\,ear\,canal} &= \frac{[P_{ed}/P_c]}{[HRTF \centerdot HPTF]_{blocked\,ear\,canal}} \\ &= \frac{1}{HPTF_{blocked\,ear\,canal}} \frac{Z_{ear\,canal} + Z_{headphone}}{Z_{ear\,canal} + Z_{radiation}}. \end{aligned} \tag{2.15}$$

$$W_{open\,ear\,canal} = \frac{[P_{ed}/P_c]}{[HRTF \centerdot HPTF]_{open\,ear\,canal}} = \frac{1}{HPTF_{open\,ear\,canal}}. \tag{2.16}$$

It is assumed here that both HRTF and HPTF are recorded at the same position in the external ear using microphones with flat frequency response. For recording

method 1 and 3, headphone equalization is simply the inverse of HPTF. However, headphone equalization becomes more complicated for recording method 2 due to the presence of additional term in (2.15) depending on the acoustical impedance $Z_{radiation}$ and $Z_{headphone}$ at the entrance of ear canal (Figure 2.1, Figure 2.3). In practice, this term has to be unity for perfect compensation or can be measured and equalized separately. For recording on human subjects, first recording method may be inconvenient because of the practical problem of measuring at the ear drum using probe microphones, while second recording method has the disadvantage that it doesn't allows listener to hear the sound and require additional compensation. Second method is particularly useful for artificial or dummy head when no ear canal is needed as this contain less individual information as against other two methods. In this thesis, we consider the measurement at open ear canal entrance to be consistent with the human subjects and dummy head recordings using miniature microphones.

2.5 Conclusions

In this chapter, an overview of binaural technology, which includes recording/synthesis of binaural signals and rendering over headphones, was introduced. Based on the past studies, it is observed that natural listening over headphones can be attained when both individualized BRTFs and individualized headphone equalization are employed in rendering. Use of non-individualized HRTFs result in large localization errors in frontal as well elevated sources and sound coloration even though interaural cues are correct. HRTFs are highly idiosyncratic and possess spectral high frequency notches due to pinnae reflections, which must be preserved in the binaural reproduction. Furthermore, room reverberations also play an important role in externalization of sound source. Interestingly, it has also been found that, when room reflections are present in the simulated sound sources, individualized and non-individualized recording doesn't have any significant perceptual differences. HPTF

also has spectral features similar to the HRTF depending on the pinnae as well as headphone reflections. Inter-individual variations in HPTFs have found to be more than 10 dB above 6 kHz. Therefore, non-individual headphone equalization can degrade the spatial perception resulting in sound coloration and localization errors. In addition, conventional equalization methods, which uses regularization techniques for inversion of HPTF may also result in pre-ringing or phasing artifacts resulting in unnatural sound. In this thesis, we present adaptive equalization techniques for individual compensation of HPTF so as to ensure virtual sounds are reproduced alike real sounds, especially in an augmented reality environment.

Chapter 3

Wave Field Synthesis using Loudspeaker Arrays

In this chapter[1], we give a detailed survey of WFS theory and the mathematical formulations derived in the literature in Section 3.1 and Section 3.2. Artifacts due to practical approximations and their solutions are also described with appropriate sound field plots (Section 3.4 and Section 3.4). Section 3.5 presents various WFS system developed in recent years. Section 3.6 concludes the chapter with future trends of WFS in commercial as well as home entertainment systems.

3.1 Wave Field Synthesis: An Overview

The main objective in developing a WFS reproduction system is to obtain driving signals (loudspeaker signals), where the primary source signals are processed using wave propagation theory [13, 88]. A typical WFS system can be divided into two subsystems: a synthesis system function and an analysis system function, as shown in Figure 3.1. Source signals are filtered, delayed, and weighted to compute all

[1] Part of this work is published in

R. Ranjan and W.S. Gan, "Wave Field Synthesis: The Future of Spatial Audio," *IEEE Potentials*, vol. 32, pp. 17-23, Mar 2013.

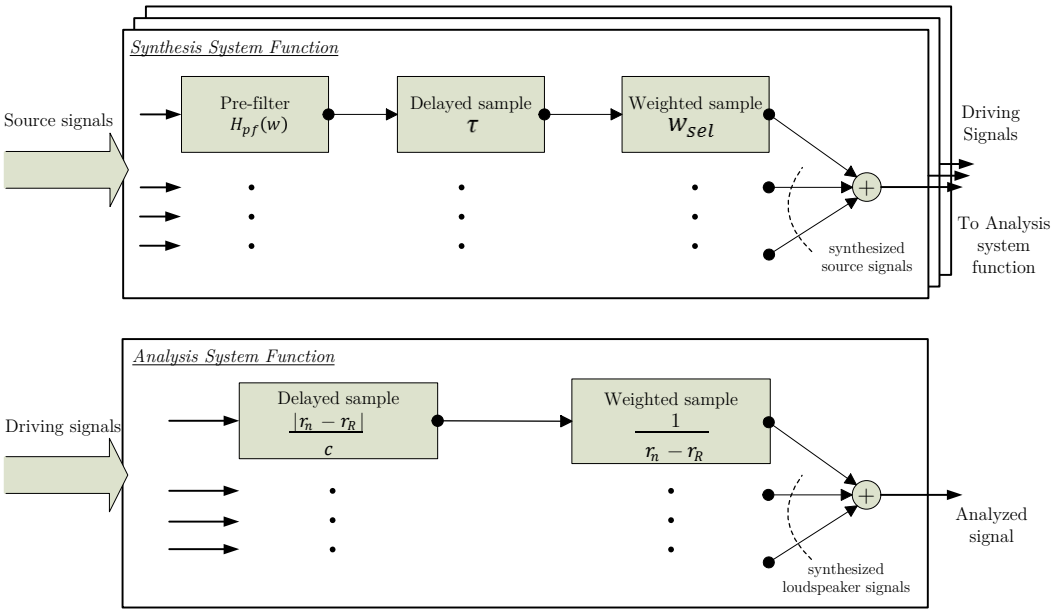


Figure 3.1: Block Diagram of a WFS reproduction system

the driving signals. This process of computing driving signals from primary source signals is termed as synthesis system function. These driving signals act as inputs to the loudspeaker array. The subsequent process of the WFS system is termed as analysis system function, which analyses the reproduced sound field at different listening positions by computing the sound pressure signals due to contributions from weighted and delayed driving signals. Detailed explanation of the two subsystems is presented in the following sections.

3.2 Principle of Wave Field Synthesis

Fundamental theory of WFS has been derived from the concept of Huygens principle [89]. This principle states that a spherical wave front (radiated by a primary source) is formed by continuous infinite secondary sources, the source strength of which determines the successive wave front and so on, as shown in Figure 3.2(a). In WFS, these secondary sources are replaced by loudspeaker arrays, which eventually reproduce the replica of the original sound field retaining the physical properties (temporal and spatial) of sound waves as shown in Figure 3.2(b) [89–92].

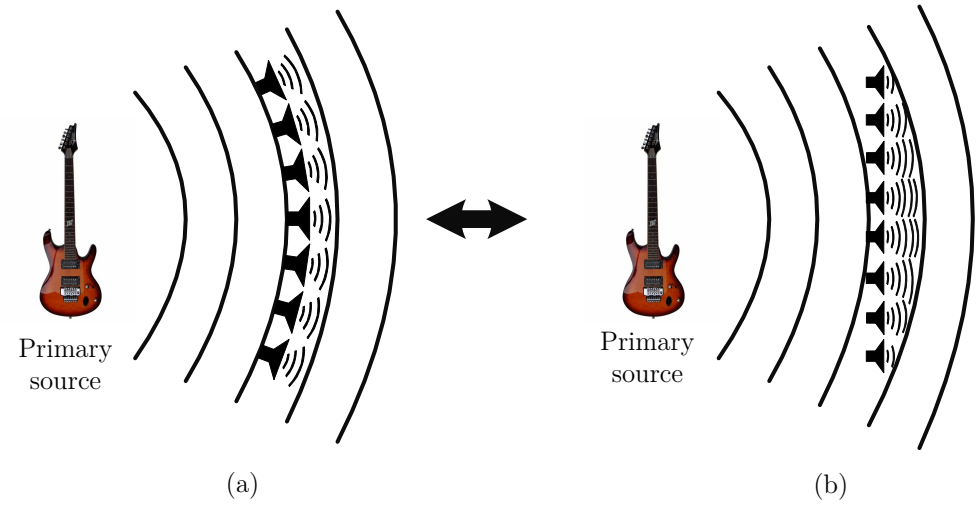


Figure 3.2: (a) Huygens Principle Realization (b) Rayleigh representation of Huygens Principle

The Kirchhoff-Helmholtz (KIH) integral (3.1) forms the mathematical basis for WFS, which applies the Huygens principle [93]. It states that sound field at a listener point, L can be calculated if the pressure and pressure gradient (due to primary or virtual sources) at boundary of the source free volume, V enclosed by a surface, S are known (Figure 3.3). In other words, a three dimensional enclosed volume is surrounded by infinite number of monopole and dipole secondary sources on its surface, which in turn reproduce the original sound field.

$$P(\overrightarrow{r}, w) = \frac{1}{4\pi} \oint \left[P(\overrightarrow{r_s}, w) \frac{\partial}{\partial n} \left(\frac{e^{-jk|\overrightarrow{r} - \overrightarrow{r_s}|}}{|\overrightarrow{r} - \overrightarrow{r_s}|} \right) + \frac{\partial P(\overrightarrow{r_s}, w)}{\partial n} \frac{e^{-jk|\overrightarrow{r} - \overrightarrow{r_s}|}}{|\overrightarrow{r} - \overrightarrow{r_s}|} \right] dS \ , \quad (3.1)$$

where k is the wave number. As shown in (3.1), KIH integral consists of two integrals, first term represents sound field due to dipole secondary sources, while second term accounts for monopole secondary sources sound field. A monopole sound source radiates sound equally in every direction with inversely proportional to the distance and its source strength is given by the pressure gradient. A dipole sound source comprises of two monopole source but in opposite phase and separated by a very small distance as compared to sound wavelength and is represented by a

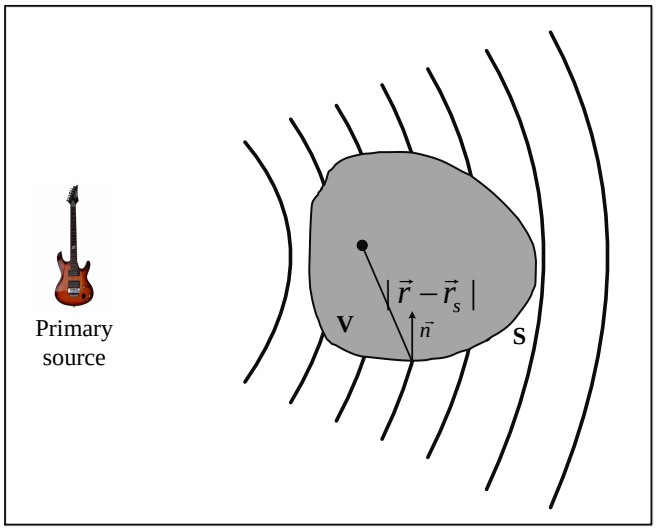


Figure 3.3: General Sound scene in an enclosed volume (Figure adapted from [4])

spherically divergent wave. Its source strength is given by net pressure at boundary of the surface, S. KIH integral (3.1) is derived using Green's theorem and then solving the wave equation [93]. Two assumptions that have been made to arrive at the KIH integral are in-homogeneous media (like air) and free field condition for wave propagation. Rayleigh proposed two modifications to the KIH integral for it to be applied in real scenarios:

- Practically, it is impossible to have an infinite continuous array of loudspeakers on the surface of enclosed volume. It is shown by Rayleigh [21] that the surface can be degenerated to a plane of loudspeakers ($z = z1$) separating the listening area from the source area, as shown in Figure 3.4. Later, it was proposed that it is also possible to move virtual sources in front of the loudspeaker array [94].

- In the KIH integral, the combined monopole and dipole secondary sources cancel out the undesirable wave fields propagating outside the enclosed surface but sums up inside the enclosed space. To correctly reproduce sound field inside an enclosed volume V, either monopole or dipole can be eliminated from the KIH integral. Above process results in non-zero sound field outside the listening volume.

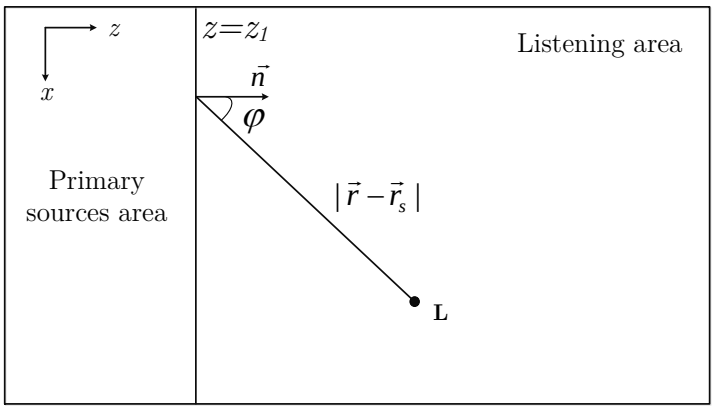


Figure 3.4: General Sound scene in an enclosed volume with degenerated surface to a plane (Figure adapted from [4])

The above propositions led to the introduction of two famous Rayleigh integrals I & II, which state that sound pressure at any point on one side of the loudspeaker plane can be synthesized from sources on the other side of the loudspeaker plane. The KIH integral (3.1) is simplified to derive the two Rayleigh integrals [93], I for monopoles (3.2) & II for dipoles (3.3) as :

Rayleigh Integral I (Monopoles):

$$P(\overrightarrow{r}, w) = \rho_0 c \frac{jk}{4\pi} \iint \left(V_n(\overrightarrow{r_s}, w) \frac{e^{-jk|\overrightarrow{r}-\overrightarrow{r_s}|}}{|\overrightarrow{r}-\overrightarrow{r_s}|} \right) dS\,, \tag{3.2}$$

Rayleigh Integral II (Dipoles):

$$P(\overrightarrow{r}, w) = \frac{jk}{4\pi} \iint \left[P(\overrightarrow{r_s}, w) \frac{1 + jk\,|\overrightarrow{r}-\overrightarrow{r_s}|}{2\pi\,|\overrightarrow{r}-\overrightarrow{r_s}|} \cos\varphi \frac{e^{-jk|\overrightarrow{r}-\overrightarrow{r_s}|}}{|\overrightarrow{r}-\overrightarrow{r_s}|} \right] dS\,, \tag{3.3}$$

where ρ_0 is the air density, c is the sound speed in air and V_n is the particle velocity in the direction of $\overrightarrow{n}$.

Rayleigh integral I use monopole loudspeakers as secondary sources with pressure gradient as signal strength (of primary source) at surface of the plane. Similarly, the Rayleigh II integral uses dipole loudspeakers as secondary source with pressure as signal strength. Rayleigh integral I have been mainly used in the literature for derivation of WFS driving signals with monopole secondary sources and we will also stick to the same definition in this thesis.

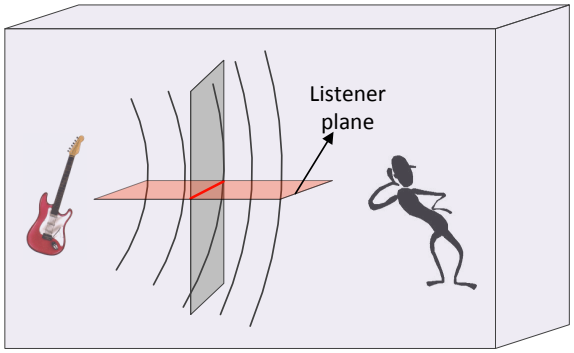


Figure 3.5: WFS in practical scenario

3.3 Practical Approximations for Wave Field Synthesis

Berkhout [13, 88, 95] introduced the concept of a virtual source, i.e., a source which is perceived by a listener when the sound is reproduced. There are two real-life scenarios where sound reproduction is employed: (a) real or primary sources exist, for example, in live performances; and (b) primary sources signals are recorded for reproduction in future, like in cinemas, TV or recorded events. In the former scenario, virtual source coincides with the real source and in the latter, WFS aims to reproduce virtual source as close as possible to real source. Furthermore, two approximations have been proposed to practically realize the WFS reproduction system:

1. *Firstly*, the loudspeaker plane is reduced to line array configuration on the horizontal plane, which is due to the infeasibility of covering the entire vertical plane with loudspeakers.

2. *Secondly*, since loudspeakers cannot be infinite in numbers and are always discrete, continuous line array is reduced to finite discrete array with uniform spacing between the loudspeakers.

Above approximations are shown in Figure 3.5 and Figure 3.6, which describes the 2D reproduction of sound field that can be perfectly reproduced on the listener

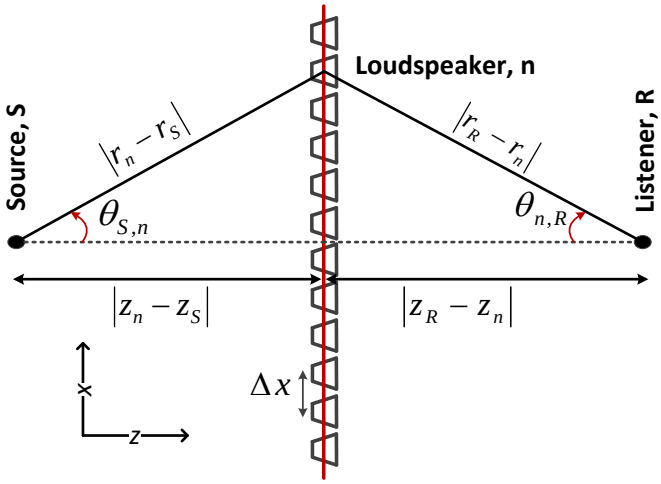


Figure 3.6: Geometry used in WFS formulations

plane.

Based on the above approximation i.e. discretization and reduction to a linear array, one dimensional form of Rayleigh integral I (3.2) is derived [89, 94] as

$$P(R,w) = \frac{1}{4\pi} \sum_{n=1}^{N} D_{WFS}(n,w) \frac{e^{-jk|r_n - r_R|}}{|r_n - r_R|} \triangle x \,, \tag{3.4}$$

where $P(R,w)$ is the synthesized sound pressure at point R in the listening area (Figure 3.6). r_n and r_R are the position vectors of the n^{th} loudspeaker and the listener position, R, respectively. The notation D_{WFS} represents the driving signal pressures function for loudspeaker n and is the principal function of a WFS system. N is the total number of loudspeakers for a given length of array. The geometry for (3.4) is shown in Figure 3.6. Δx is the loudspeaker spacing of the array. The exponential part in (3.4) is the three dimensional Green's function, which represents the radiation of a secondary monopole point source [93]. In time domain, sound pressure at any listener point, R can be simply calculated by summing the weighted and delayed contribution from each of the loudspeakers as:

$$P(R,t) = \frac{1}{4\pi} \sum_{n=1}^{N} \frac{1}{|r_n - r_R|} d_{WFS}(n, t - \frac{|r_n - r_R|}{c}) \triangle x \,. \tag{3.5}$$

The sound field for a monopole point source, S in free-field at listener position, R

is given by

$$P(R, w) = S(w)\,\frac{e^{-jk|r_R - r_S|}}{|r_R - r_S|}\,. \tag{3.6}$$

The solution for driving signal at n^{th} loudspeaker is obtained by solving (3.4) with sound field of a monopole point source in free-field (3.6), using stationary phase approximation [89] as

$$D_{WFS}(n, w) = S(w)\sqrt{\frac{jk}{2\pi}}\sqrt{\frac{|Z_R - Z_n|}{|Z_R - Z_S|}}\cos\theta_{S,n}\frac{e^{-jk|r_n - r_S|}}{\sqrt{|r_n - r_S|}}\,. \tag{3.7}$$

The planar-to-linear array reduction introduces inaccuracies in the sound pressure attenuation, which is compensated by a pre-filter with a 3 dB per octave attenuation and a phase shift of 45^0 [96]. Third term $\sqrt{|Z_R - Z_n| \,/\, |Z_R - Z_S|}$ in (3.7) is the correction factor and also accounts for the 2D approximation. As evident, it depends on the source-receiver distance and source-listener distance along the z-plane. Thus, the driving signal is weakly dependent on receiver position (in fact on the receiver line, Z_R) and as a result, synthesized sound pressure field is correct only on the reference receiver line (also called 'sweet line' by de Vries [48]), which is used in the computation of driving signal. The second last term $\cos\theta_{S,n}$ is a weighting factor depending on the source incidence angle. The last term of (3.7) represents the natural wave propagation and amplitude decay of source, S to loudspeaker, n. The geometry for the equations is shown in Figure 3.6.

Spors [97, 98] further modified the equation for driving signal to introduce a window function to suppress the undesired reflections from those loudspeakers, whose normal component does not match with local propagation direction of virtual wave field so as to minimize the error in the reproduced wave field. The window function mutes these loudspeakers reproducing undesired reflections. In addition, Spors derived separate driving signals for plane wave source and spherical point source. Details can be found in his work [97].

From (3.7) it is clear that driving signals can be calculated efficiently in time domain by pre-filtering the source signal, follows by weighted and delayed sample of the filtered source signal as:

$$d_{WFS}(n,t) = s(t) * h_{pf}(t) * \delta\left(t-\tau\right) w_{sel}\,, \tag{3.8}$$

where,

$$\begin{aligned} s(t) &= F^{-1}[S(w)] \\ h_{pf}(t) &= F^{-1}\left[\sqrt{\tfrac{jk}{2\pi}}\right] \\ \tau &= \tfrac{|r_n - r_S|}{c} \\ w_{sel} &= \sqrt{\tfrac{|Z_R - Z_n|}{|Z_R - Z_S|}}\cos\theta_{S,n}\tfrac{1}{\sqrt{|r_n - r_S|}} \end{aligned}$$

Symbol $*$ represents the convolution between two timain domain sequence.

Synthesis system function, implements driving signal using (3.8) requires source parameters (source signals, positions, orientations etc.) and loudspeaker parameters (loudspeaker positions, orientation, number of loudspeakers etc.) as inputs. Equation (3.8) computes the driving signal rendered for a single source, S. To compute driving signals for more than one source, (3.8) is repeated for each of the sources and summed together for each of the loudspeakers as shown in Figure 3.1. Similarly, analysis system function implements the synthesized sound signal at a listener position, R and time, t using (3.5) requires driving signals, listener position and loudspeaker parameters as input for analyzing sound field in the listener space. Analysis system function can be used to analyze the quality of the physical sound field either by listening to the reproduced sound signals at different listener positions or by visualizing the sound field in the entire listening area. Objectively, sound field quality can be estimated using spectral distortion (SD) between frequency response of WFS virtual source and an ideal point source for different listener postions. Spectral distortions are mainly due to the spatial aliasing, resulting in spectral peaks and dips.

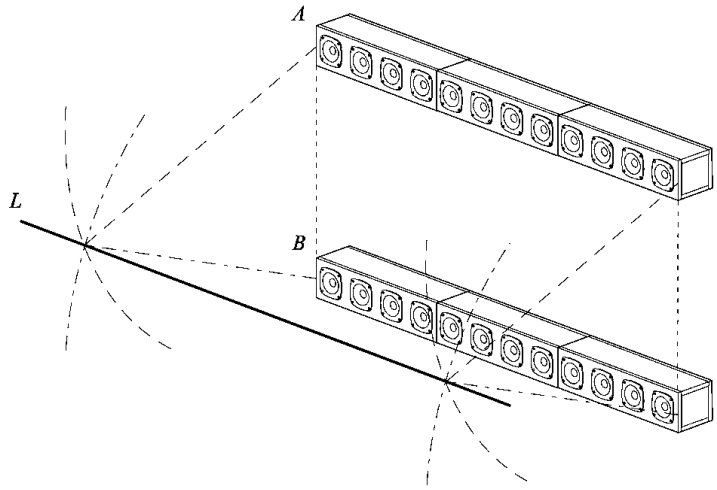


Figure 3.7: Example of a multiple line array in WFS (Figure extracted from [5])

3.4 Practical Constraints and Solutions

As discussed in the previous section, it is not practically realizable and cost effective to place loudspeakers everywhere in a closed space. Moreover, computational power of a typical WFS processing engine is largely dependent upon the number of loudspeakers and complexity of auditory scenes.

WFS formulations for driving signal equation work well only for the reproduction on horizontal plane (listener plane) because of the approximation to line array. Since the two ears are located on the horizontal plane, it would be sufficient to assume that the sound perceived will be natural to us. Reproduced sound field is accurate only at the sweet line and thus, resulting in amplitude error, which is measured as deviation from sound pressure on sweet line (in dB) [91, 99]. Thus, we need to choose a fixed reference line ($Z = Z_R$) such as to minimize the overall amplitude errors in the listening area. It is usually chosen in the center of the listening area, but Sonke et al. [100] showed that a reference receiver contour can also be chosen to minimize the overall amplitude errors, instead of using the conventional reference line. Because of the 2D reproduction on the horizontal plane, virtual sources are not correctly perceived in the vertical plane. But with the advent of 3D audio-

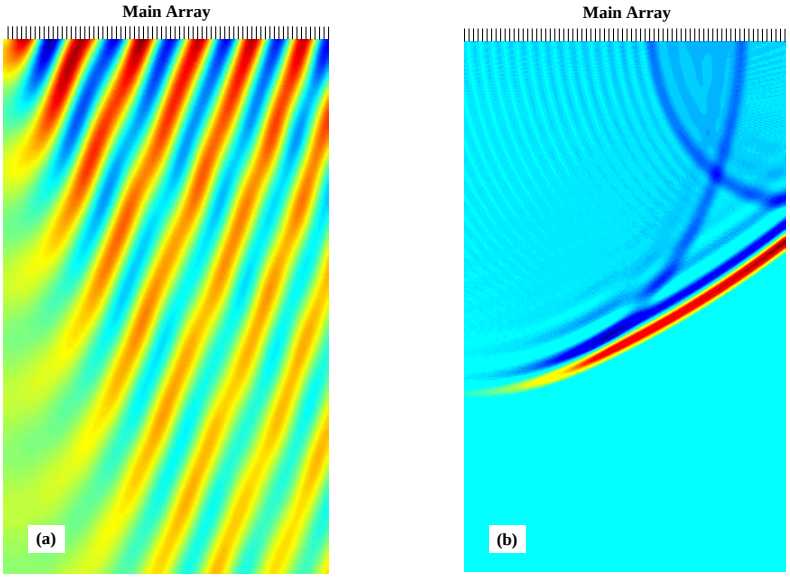


Figure 3.8: Truncation Effects for WFS reproduction of finite array with Δx = 0.1 m (a) Monochromatic plane wave signal with frequency 800 Hz (b) Low pass filtered pulse with cut-off frequency = 1,500 Hz

visual contents, like gaming videos and 3D movies, where elevation perception is of utmost importance, it is required to find a solution that emulates the 3D plane reproduction. Recently, Montag [5] proposed a multiple array lines of loudspeakers in vertical plane to extend the traditional WFS to 3D reproduction (Figure 3.7).

Mathematically, we can derive driving signal for any arbitrary configuration of closely spaced loudspeaker array, but in reality it is near to impossible to have a inter-speaker spacing less than 1 cm meaning WFS reproduces true sound field only up to a corner frequency. This is due to the reduction of an infinite continuous line array to a finite discrete array and suffers perceptual quality degradation to some extent. As a result of reduction to finite continuous line array, diffraction effects, and additional trailing waves (also called 'shadow signals [48]) are observed in the sound field derived using analysis equation. Figure 3.8 shows the diffraction effects for two test signals, monochromatic plane wave signal and a low pass filtered pulse. Virtual source is located in far left corner behind the loudspeaker array. We can clearly observe the shadow waves in left side of the Figure 3.8(a) as a result of diffraction effects and in accordance with the diffraction ray theory. In addition,

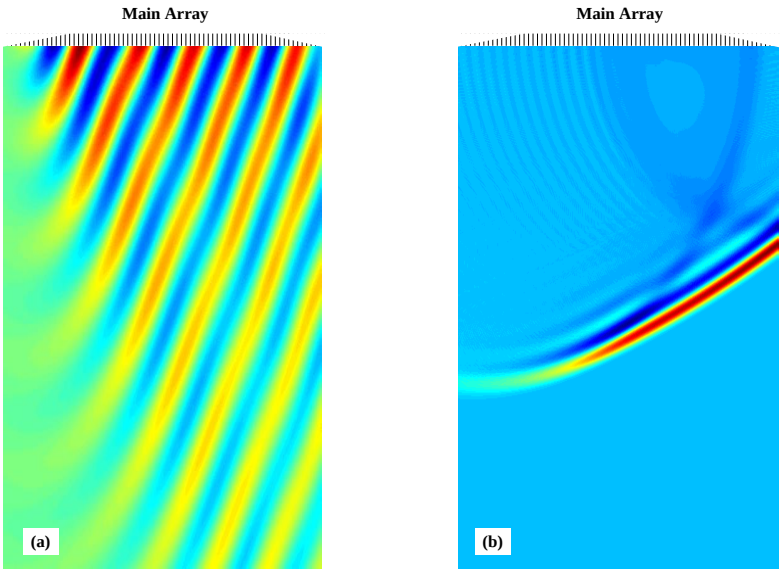


Figure 3.9: Same as Figure 3.8 with tapering applied at the extremes of the loudspeaker array

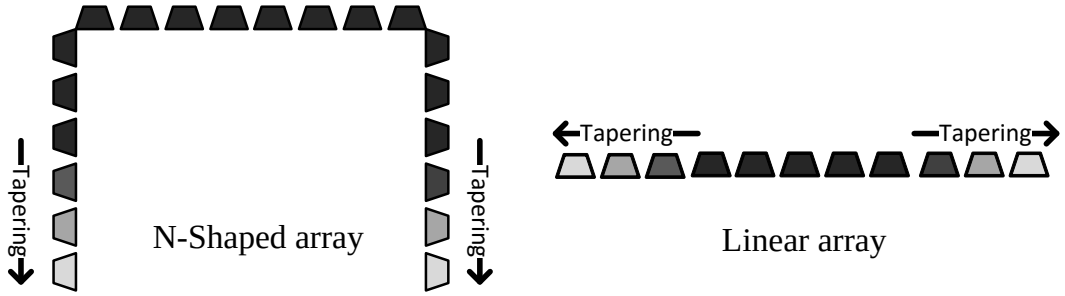


Figure 3.10: Tapering applied to N-shaped array and linear array

we can see the circular trailing waves due to edge loudspeakers (Figure 3.8(b)). This effect is also known as truncation effect which originates from loudspeakers at the extremes of array. Perceptually, this can cause slight coloration effect and echo perception depending on the time difference between desired response and shadow signals [48, 91]. In addition, visible source area for listener, which is defined by extent of the loudspeaker array, reduces due to the finite length of the loudspeaker array.

Tapering is a technique used to smoothen the truncation effects, where loudspeakers positioned on the edges are given less weighting but at the cost of reduced

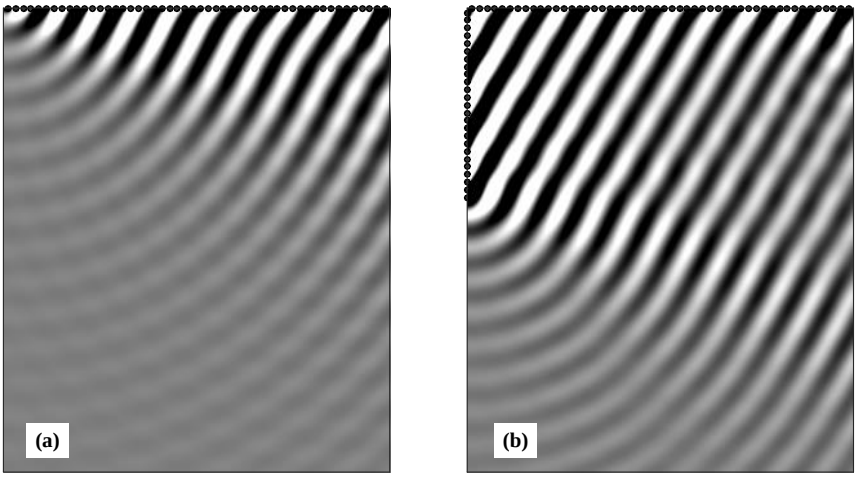


Figure 3.11: Enlarged listening area using additional side array with linear array for Monochromatic plane wave signal with frequency 800 Hz, $\Delta x = 0.1$ m (a) Linear array only in the front (b) Additional side array with linear array in the front

effective array length. One-sided cosine window is used to apply the tapering as shown in Figure 3.9. The effect of tapering can be observed from Figure 3.9 as trailing waves are smoothened but at the cost of reduced reproduced area.

Effective array length can be increased by using N-shape arrays, with tapering applied at the two extremes, like shown in Figure 3.10. Effect of side-arrays with tapering applied as shown in Figure 3.10 is discussed in [91]. As shown in Figure 3.11, we can observe the enlarged reproduction area using an additional side array.

Finite continuous array is reduced to finite discrete array by applying sampling in spatial domain resulting in spatial aliasing, which is similar to aliasing in the frequency domain. It is easier to analyze the aliasing artifacts in the wave number domain [89, 99], which is obtained by taking Fourier transform of sound signals in spatial domain. However, WFS is correctly achieved only up to a corner frequency known as 'spatial aliasing frequency'. In spite of the inaccurate synthesis of physical sound field, it has been found that a reasonable amount of deviation from aliasing criteria does not significantly degrade the perceptual quality [101]. This is mainly due to the human auditory mechanism that allows some tolerable level of distortion

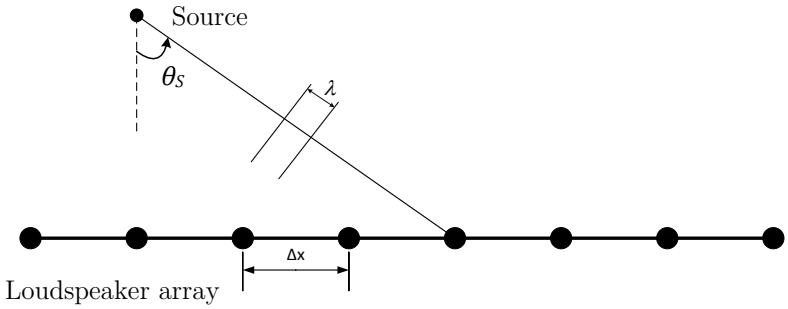


Figure 3.12: Spatial sampling of the loudspeaker array for a source far away

due to spatial sampling. Most WFS systems use a loudspeaker spacing of 10 to 20 cm [48, 102, 103]. According to sampling theorem, in spatial sampling, spatial aliasing does not occur if loudspeaker spacing is less than the minimum apparent wavelength component along the loudspeaker array, i.e.

$$\triangle x \leq \frac{1}{2}\frac{\lambda}{\sin\theta_S}, \tag{3.9}$$

where θ_S is the source incidence angle located far away from the loudspeaker i.e. acting like a plane wave source and λ is the source wavelength, as shown in Figure 3.12. Thus, from (3.9) spatial aliasing will occur only above spatial aliasing frequency f_{al} :

$$f_{al} = \frac{c}{2\triangle x \sin\theta_S}. \tag{3.10}$$

Figure 3.13 shows the spatial aliasing effects for a loudspeaker spacing of 0.2m with source situated 1m left of the array and 3 m behind the array. Spatial aliasing frequency using (3.10) is 850 Hz. Thus, we can clearly observe aliasing artifacts as destructive interference with plane wave coming symmetrical opposite direction. Perceptually, spatial aliasing can result in timbre and sound colorations.

A number of methods have been proposed in literature to minimize the spatial aliasing effects. Spatial bandwidth reduction uses the notion of directive sound

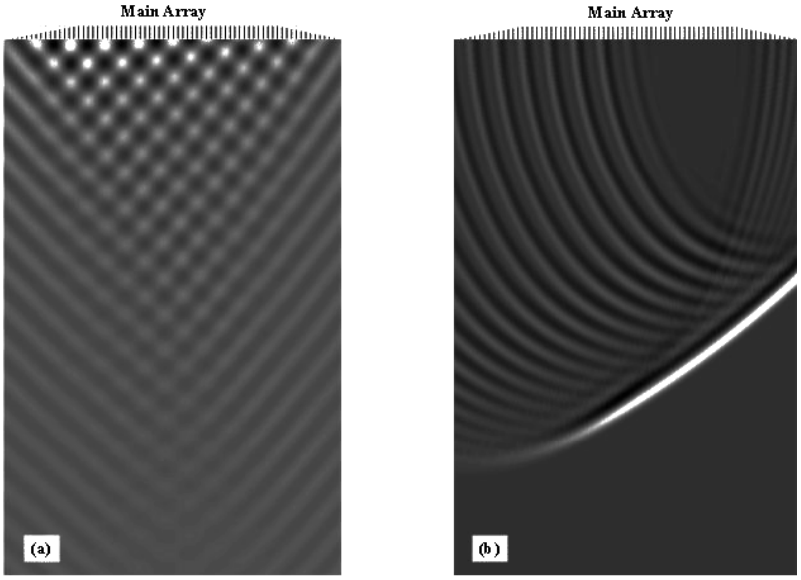


Figure 3.13: Plane wave reproduction with spatial aliasing, $\Delta x = 0.2$m (a) Monochromatic plane wave f = 1,500 Hz (b) Low pass filtered pulse with cut-off frequency = 2,200 HZ

sources to reduce the interference between loudspeakers [94, 99]. Another method is to randomize the high-frequency content over the loudspeaker array and thus, reduce the periodicity in spatial aliasing artifacts [101, 104]. In recent years, researchers have analyzed the spatial aliasing artifacts by deriving several aliasing criterions, which not only depend on spacing between loudspeakers but also on source directivity and listener positions. In one of the recent works by Corteel [105], spatial aliasing frequency is increased with the help of dynamic selection of sub-part of the loudspeaker array to target reproduction within a preferred listening area.

Another limitation of Rayleigh theory, which states that source (non-focused source) can only exist behind the loudspeaker array, has been resolved by the introduction of focused sources [106]. A focused source can be perceived in front of the array, i.e., in the listener space [107, 108]. The only constraint with focused source reproduction is that listener area is reduced and the listener is not permitted to sit in between the array and the focused source [48]. Both focused and non-focused sources are crucial in recreating the immersive sound field around the listener. Listener can feel the depth of the source, but it requires entire listener space to be

surrounded by closed configuration of loudspeaker arrays.

All loudspeakers in practice possess some directivity pattern in contrast to the ideal monopole secondary sources. This implies that conventional driving signal equation holds only for the ideal monopole conditions. De Vries [48, 90, 109] derived that driving signal for a linear array can be adapted to loudspeakers with arbitrary directivity characteristics. It should also be noted that traditional equations for WFS assumes free-field conditions, and room reflections must be accounted while dealing with real room simulations. Mirror image source model is commonly used for the analysis of room reflections [48, 89].

3.5 Evolution of Wave Field Synthesis

Since the introduction of WFS by Berkhout [13] in 1988, WFS has come a long way over the last two decades and are now playing a vital role in spatial audio reproduction technology. Berkhout started the research on WFS based system at Delft University and laid the foundation for further developments. He was supported by fellow researchers in particularly, De Vries, Vogel, Start and others in the following years [92, 110–112]. The first WFS based practical laboratory set up, which consists of 48 channels with DSP processors, was developed at Delft University in 1993 and later extended to the university's auditorium [89].

Berkhout's work was followed up by many other prominent research groups and many WFS based set ups were installed in various places, including cinemas, lecture halls and concert halls. Till late 1990s, most of the research was carried out at universities, mainly focused on developing mathematical formulations of WFS equations and also practical measurements based on various configurations (linear, circular, rectangular etc.) of loudspeaker arrays [113]. As a result of the increased interests and participations in WFS, several research groups and R&D labs collaborated to standardize a WFS format, which led to the start of European Union

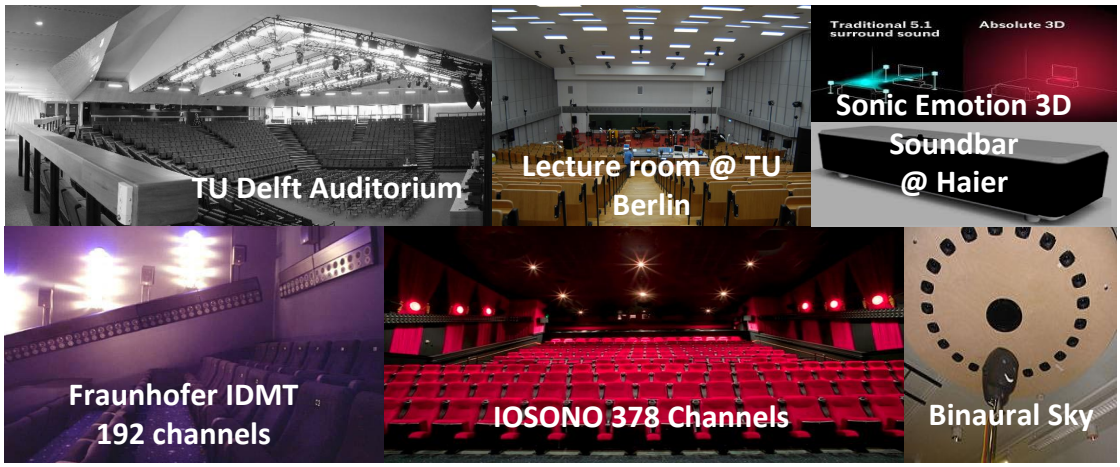


Figure 3.14: A look at various WFS developments

(EU) Information Society Technologies (IST) "CARROUSO" ("Creating, Assessing and Rendering in Real Time of High Quality Audio-Visual Environment in MPEG-4 Context") [114] project in 2001 and completed by 2003. The main goal of the CARROUSO project was to develop a new technology, which can record, encode, transmit and reproduce the sound field recorded at virtual or remote place. This project prompted researchers, as well as commercial market based on spatial audio in different parts of the world to focus on WFS with the goal of creating potential applications in spatial audio systems, and can potentially replace multichannel surround sound systems placed in cinemas, live events or home theater systems. The successful completion of CARROUSO project led to the emergence of two new companies, IOSONO [115] and Sonic Emotion [116], which are aimed to provide services and solutions for installations of 3D audio systems based on WFS. IOSONO is spin-off from Fraunhofer Institute for Digital Media Technology in 2004 and bought by Barco Audio Technologies in 2014, while Sonic Emotion is co-founded by Renato Pellegrini. Both of these companies have played a significant role in the success of CARROUSO project. They are now the major providers of WFS based products for consumer market, as well as research applications in spatial audio systems.

Recently, IOSONO has launched a spatial audio processor to control any kind of loudspeaker arrangement, room geometry and listener numbers. Sonic Emotion

has manufactured 3D audio chips based on their own patented technology employing WFS, psychoacoustics and others. In 2011, Haier launched a 3D sound bar using this chip, which claims to create the unique sound experience that can replace the current home theatre systems. In 2008, Spors [97] and his team at Deutsche Telekom in Berlin revisited the WFS theory and also proposed modified driving equations addressing arbitrarily shaped loudspeaker arrays for three-dimensional sound reproduction. They also installed practical WFS set up of 56 channels of circular loudspeaker array. Furthermore, they have developed a generic spatial audio renderer framework for real time audio processing, which is very useful for sound reproduction in real time [117]. This versatile software allows the rendering of several rendering modules like WFS, binaural, Ambisonics, virtual amplitude based panning, etc. Researchers at IRT, Germany developed a novel system known as the 'Binaural sky', which uses WFS technology for binaural sound reproduction [118]. The latter system consists of overhead circular array of loudspeakers and synthesizes focused sources using head tracking system. Figure 3.14 shows various WFS set ups installed at various universities and auditoriums.

3.6 Future Trends and Conclusions

In the last few years, WFS is increasingly becoming more popular in commercial deployment. WFS based reproduction systems are now readily accepted as the most optimal way of reproducing spatial sound. Several companies have already started installation of WFS in public places. Recently, Game of life foundation developed the world's first transportable WFS setup and demonstrated at Amsterdam in 2011. Similar set up was earlier demonstrated at the 124^{th} AES convention on the eve of 20 years of WFS in 2008 [119]. Till now, we have mainly seen large scale installations of WFS in large public places. Everyone now appreciates the immersive environment reproduced by WFS based system in such places. In recent years, researches have

started to look into scaled down version of WFS to target small group listening. Some small-scale WFS applications include virtual reality, 3D gaming environments, and video-conferencing [102, 120]. With the capability of reproducing virtual content very far (outside the room) as well as near to the listener, WFS overcomes the feeling of being in a particular environment making it suitable for realizing WFS systems in home entertainment systems in the future. However, a major hurdle for the use of WFS technology in such small-scale applications is that these systems are too space-consuming for use in every place and require minimal number of loudspeakers for enlarged sweet area. Since a typical WFS system requires large and costly set up of loudspeaker arrays, it is still an open research problem to devise a trade-off between number of loudspeakers and size of sweet spot area. Corteel [105] from Sonic Emotion have recently proposed a new methodology to employ fewer loudspeakers while increasing the spatial aliasing frequency using the focused sound reproduction in a 'preferred listening area'. Loudspeaker placement and room acoustics are the two main problems for WFS to be adopted in small scale applications. For the former, recently flat panel speakers were explored for WFS [121, 122] and might be better suited for places with space constraint, especially in home reproduction. Acoustics of the room can also affect the intended sound field in listening area and may introduce coloration. Different equalization techniques have been applied to control the sound field in listening area by compensating for the room reflections [123–126]. Additionally, for WFS to enter into our homes, all the recording and encoding should be carried out such that it can be mapped to any WFS speaker layouts (as explained in CARROUSO project) before transmitting them. Only then, we will be able to take full advantages of WFS in immersive 3D sound reproduction.

In this chapter, we gave an overview of the principle of WFS, and presented some of the key research work and commercial products over the past two decades. Today, WFS has emerged as one of the key spatial audio technology in the professional installations and is being widely adopted. We also highlighted some practical limita-

tions and technical challenges of WFS-based sound reproduction systems. Approximations employed to KIH integral for practical realizations of WFS put constraints on the synthesized sound field quality, visible area, number of loudspeakers, array length, and array configuration. Various solutions have been proposed in the literature to compensate for the practical limitations of WFS reproduction systems. In this thesis, we aim to develop a WFS based setup to be used as home entertainment system. A hybrid WFS setup is presented in Chapter 6 by combining WFS and binaural synthesis over headphones for immersive audio reproduction. The main aim of this hybrid setup is to create an immersive experience around the listener using virtual WFS over headphones for rear and side auditory scenes, while physical WFS array is used for frontal auditory scene reproduction. In the next chapter, we introduce a natural augmented reality (NAR) headset and adaptive headphone equalization techniques for natural listening in augmented reality scenarios.

Chapter 4

Natural Listening Over Headphones in Augmented Reality using Adaptive Filtering

In this chapter[1], we introduce a natural augmented reality (NAR) headset and proposed adaptive equalization techniques for individual headphone compensation for augmented reality headphones. This chapter is structured as follows: Section 4.1 gives a brief introduction and some of the related works on augmented reality headphones. Section 4.2 evaluates the effects of different types of headphones on the direct sound spectrum. Section 4.3 introduces natural listening techniques using the proposed NAR headset. Adaptive equalization methods for reproducing virtual sources, with and without the presence of external real signals are presented in the subsection 4.3.4 and subsection 4.3.5, followed by the subjective test results in Section 4.4. Finally, Section 4.5 concludes the chapter highlighting key results of

[1] This work has been published in

1. R. Ranjan, W.S. Gan, and C. Yong-Kim, "Applying Active Noise Control Technique for Augmented Reality Headphones," *in Proceedings of Internoise*, Melbourne, 2014.

2. R. Ranjan and W. S. Gan, "Natural Listening over Headphones in Augmented Reality Using Adaptive Filtering Techniques," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1988-2002, 2015.

this work.

4.1 Introduction

Augmented reality (AR) is changing the way we live in real world by adding a virtual layer to our sense of sight, smell, sound, taste, and touch along with real-world senses to give us an enriched experience. With the advent of wearable devices, such as VR Gear and Oculus Rift, sensors like microphones and cameras that capture our surroundings with global positioning system, AR technology provides sensory dimensions to the user to navigate more effectively in the real and virtual world. An AR system is defined with three main characteristics, namely, superimposition of virtual objects onto physical world, ability to interact in real-time, and projection in three dimensional space [127]. AR devices are currently used in several application areas like assistive navigation for the disabled [128], augmented reality gaming [129], medical [130], audio-video conferencing [131, 132], binaural hearing aids, and audio guides in museum or tourist places [133]. So far, visual information is predominantly used in AR-enabled devices to provide additional guidance or information to the user. Spatial sound is also being incorporated into AR devices to provide auditory cues of virtual and real objects in the listener's space via headphones [131]. These spatial cues can be used to alert listener of imminent danger/obstacle in a certain direction; add to the realism in gaming; and give a feeling of being there in the augmented environment. The ultimate goal of deploying spatial sound via headphones in AR devices is to create the impression that virtual sounds are coming from the physical world. At the same time, virtual sources should merge with the real sources in a transparent manner, enabling awareness to the real sources.

There have been several works in recent years in an attempt to playback spatial audio in AR based headphones, as well as existing commercial headphones. Haptic audio (sound that is felt rather than heard) is applied in headphones to enhance the user experience [134]. Bone conduction headset enabling hear-through augmented

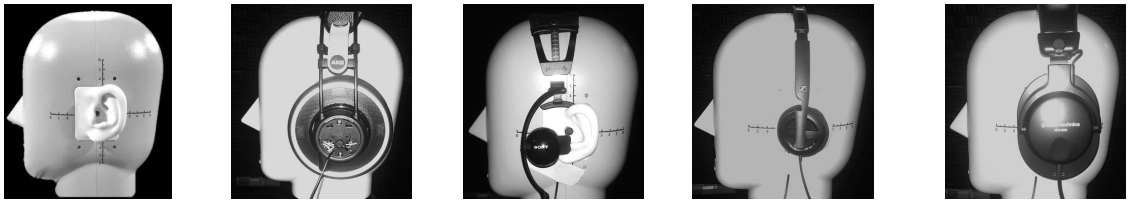
reality gives comparable performance to speaker array based system [135]. An augmented reality audio (ARA) headset has been introduced by Härmä et al. [136] using an in-ear headphones with binaural microphones to assist the listener with pseudo-acoustic scenes. In another work, the same authors further developed an ARA mixer [137, 138] to be used with headset for equalizing and mixing the virtual objects with real environments. The main problem addressed in the ARA headset is the blockage of natural sounds coming from outside and reaching ear drum due to the in-earphone structure. Thus, their goal is to reproduce natural sounds unaltered with the help of binaural microphones to capture, process, and playback so as to make the ARA headset acoustically transparent. However, the direct sound leakage cannot be completely avoided and earphone repositioning might also affect the reproduced sound quality. Schobben and Aarts [139, 140] proposed a headphone based 3-D sound reproduction system with binaural microphones positioned inside the ear cup near ear opening using active noise control (ANC) based calibration method. Filtered-x least mean square (FxLMS) adaptive filtering algorithm is used to achieve sound reproduction close to the 5.1 multichannel loudspeaker setup. The key problem being solved here is the large localization errors for most listeners due to non-individualized equalization of headphones. ANC is used to calibrate the system for every individual to identify loudspeakers' transfer function at listeners' ears before playing 5.1 multichannel virtual auditory scenes through headphones. Therefore, the primary challenge in AR based headphones is to reproduce sound as close to natural as possible so that augmented audio environment presented are well externalized with no front-back confusions. Most importantly, virtual audio objects/scenes are seamlessly augmented in the real environment.

In this chapter, we present a natural augmented reality (NAR) headset with two pairs of binaural microphones to achieve natural listening experience using online adaptive filtering. An open type headphone structure is chosen over closed in-ear type headphones mainly because of two reasons: (1) its open-cup design allows ex-

ternal sound to pass through without much attenuation, resulting in a more natural listening experience; (2) open ear canal resonance, which is more natural compared to the blocked ear canal resonance of the in-ear headphones. With the use of sensing microphones installed in the headphone structure and by applying real-time adaptive training, virtual sources are reproduced as close as possible to real sources and thus, adding realism to the augmented reality environment (ARE). Modified versions of the filtered-x normalized least mean square (FxNLMS) algorithm are proposed in order to improve the slower convergence rate and steady state performance of the conventional FxNLMS. One of the main objectives is to ensure virtual sound objects become part of the real auditory space as augmented space. Therefore, the proposed approach is extended for the case when both real and virtual sources are to be mixed together, such that signal due to external sources does not interfere with the convergence process of the FxNLMS. The main advantage of applying FxNLMS technique here is that it adapts to the individualized head-related transfer functions (HRTF), while compensating for the individual headphone transfer function (HPTF), as it alters the desired spectrum at listeners' ears. Thus, adaptive process ensures the NAR headset is individualized to a listener and virtual sources are reproduced alike real sources. Using dummy head measurements, it is found that the proposed approach is able to closely match the natural sound reproduction with faster convergence rate. The proposed method is also found to be equally effective in the presence of external sounds. Subjective study based on individualized binaural impulse responses (BRIRs) is conducted to validate the proposed approach and assess whether listeners can distinguish between real and virtual sounds.

4.2 Headphones Effect on Direct Sound Spectrum

In an ARE, a user wearing a NAR headset must not feel isolated from the surroundings. The choice of headphones is crucial in designing a NAR headset as it



(a) Reference (ref) Without headphone

(b) Open Circumaural (CA)

(c) Open personal field speaker (PFS)

(d) Open Supra-aural (SA)

(e) Closed circumaural (CA)

Figure 4.1: Bruel & Kjaer dummy head with different type of headphones used for the HRTF measurements.

should allow external sounds to pass unblocked and reach listeners' ear drum in a natural manner. We have tested four different types of commercial headphones to evaluate their effects on direct sound source spectrum. Figure 4.1 (b)-(e) show the four types of headphones, namely, open circumaural (CA), open supra-aural (SA), open personal field speaker (PFS) and closed circumaural (CA), which are worn on the dummy head for measurement. The open PFS headphones are completely open with external drivers facing towards the pinna from the frontal direction, as shown in Figure 4.1 (c).

Effects of the aforementioned headphones on the direct sound source spectrum for different azimuths are measured and shown in Figure 4.2 along with the direct sound spectrum as reference (Ref) HRTF measured without the headphones. Measurements were conducted in an anechoic chamber at NTU, Singapore using the Bruel & Kjaer 4128D head and torso simulator. Exponential sine sweep signal with sampling frequency of 44.1 kHz was played from sound source placed at 1.4 m away from the center of dummy head and recorded at the binaural microphones located at ear drum of the dummy head. All the HRTFs are one-third octave smoothed to decrease perceptually redundant fluctuations, especially in high frequencies [141]. As shown in Figure 4.2, headphones act as passive low pass filters and that is why the difference between the headphone modified spectrum and direct sound spectrum is observed only in high frequencies above 1.5 kHz, except for the closed CA headphones, which attenuates up to 10-15 dB is observed below 1.5 kHz. One common

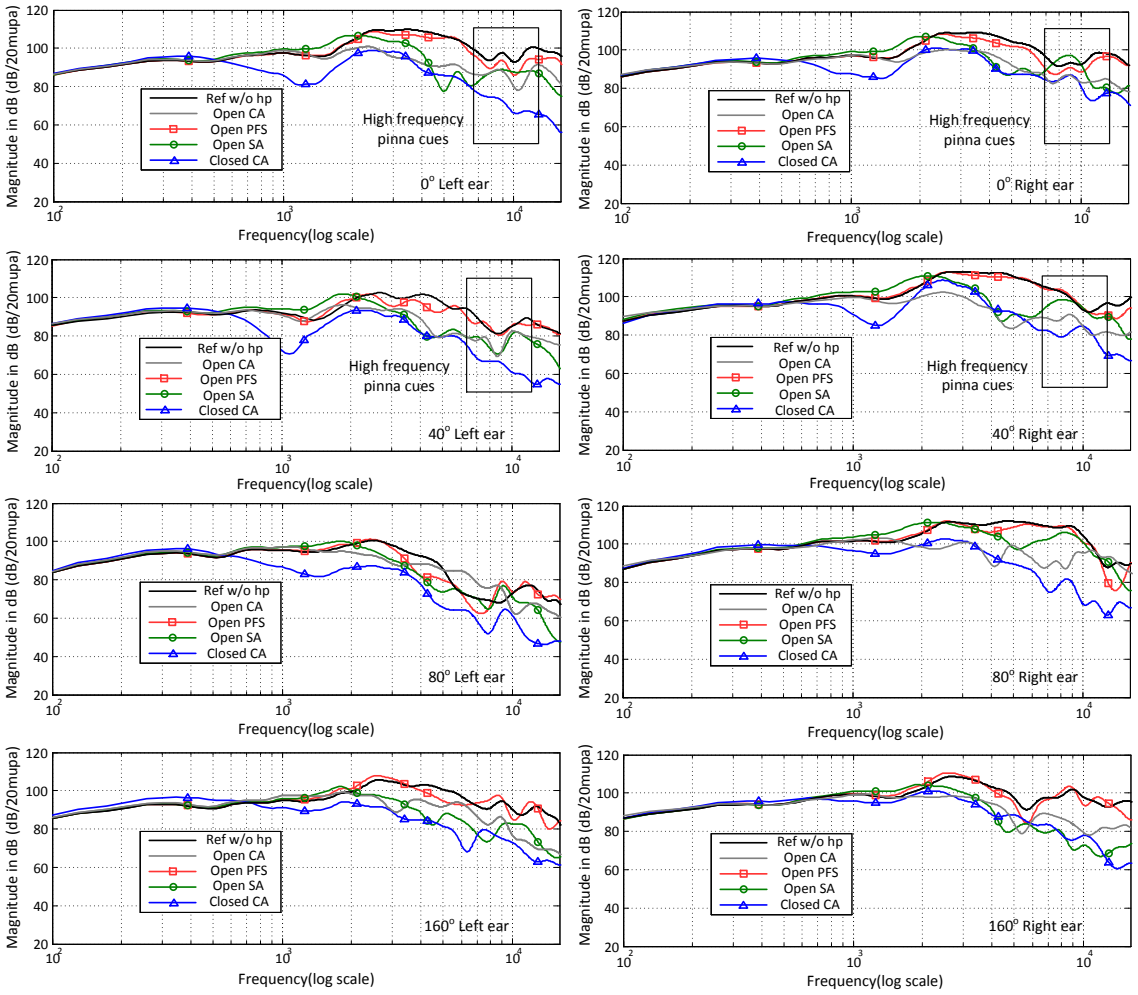


Figure 4.2: Effects of four types of headphones on the direct sound source spectrum at ear drum of the dummy head

observation from the HRTF plots in Figure 4.2 is that the closed CA headphones attenuate direct sound the most as compared to other open headphones, while open PFS headphones is the most acoustically transparent headphone for all azimuths. For the closed CA headphones, attenuation up to approximately 30 dB is observed for most of the azimuths. This leads to significant coloring of the direct sound source spectrum. The open SA headphones and open CA headphones possess average headphone attenuation of roughly up to 10-15 dB in high frequencies. Other important aspect that needs to be observed is the high frequency pinnae specific notches, which is particularly essential for the frontal localization [68, 142] as well as elevation [143]. These notches for the open CA and open PFS headphones are

consistent with the reference HRTF, but mismatch/absence of the notch positions is observed for the open SA and closed CA headphones. For the open SA headphones, it might be due to the fact that headphones are resting on the ear, suppressing the reflections due to pinna. For the closed CA headphones, strong passive isolation by the headphones structure possibly leads to reduction/disappearance of notches. To summarize, closed headphones are not suitable for AR based application due to its strong isolation property and coloration of the sound spectrum when no signal processing is allowed for natural listening of external sounds.

Similar observations were also reported in [60] on the influence of headphones to localize loudspeaker source. In particular, it was found that the localization accuracy degraded only slightly due to wearing of headphones as compared to open-ear listening. It was also found that listeners used head rotation more frequently as additional cue to assist in localization when headphones were worn. However, large ILD errors due to high frequency loss will result in audible coloration, as well as dulling of the sound. Spectral and ILD distortions were less pronounced for headphones with "more" open design. Therefore, it is suggested that care must be taken while selecting headphones for ARE. In practice, absolute acoustic transparency cannot be achieved, but headphones characteristics can be modified through signal processing techniques and/or passive techniques to achieve realistic impression of physical sound sources and environments. The following sections outline the adaptive signal processing techniques to achieve natural listening through headphones.

4.3 Natural Listening via Natural Augmented Reality Headset based on Adaptive Filtering

Natural listening using headphones require sounds to be reproduced as natural as possible. For AR based scenarios, we would need both real and virtual sounds to be perceived in the same way such that one cannot distinguish between the

two. In addition, a realistic fusion of virtual sound objects with the real auditory environment is also desired for ARE. We thus, divide our analysis into three possible practical scenarios, which will be presented in the subsequent sub-sections:

Case I: Only real source present

Case II: Only virtual source present

Case III: Virtual source in the presence of real source/surroundings

In the following sub-sections, we present the proposed NAR headset structure followed by adaptive signal processing techniques to achieve natural listening in ARE. Our main focus in this work is on the adaptive algorithms to create virtual auditory events that are engrossed with the real environment, giving an immersive experience to the listener. Case I scenario with only real external sources, may not need any additional processing if using an open type headphones. Hence, the focus of this chapter is mainly on Case II and Case III, where virtual sources are needed to be reproduced naturally to the listener, without and with the presence of the real sound sources, respectively.

4.3.1 Proposed headset Structure

The proposed NAR headset structure and the prototype constructed using AKG K702 open CA studio headphones is shown in Figure 4.3 (a). The open CA headphones is preferred over the other two types of open headphones for ease of microphones placement in our prototype. There are two microphones attached on each side of the headphones ear cup. As shown in the figure, internal microphone, denoted by m_{int}, is positioned very near to ear opening, whereas external microphone, denoted by m_{ext}, is positioned just outside the headphones ear cup. The main purpose of internal microphone (also known as the error microphone) m_{int} is to adapt to the desired virtual sound field measured at listeners' ears. External microphone (or reference microphone) m_{ext} is used to capture the external sounds. Besides these,

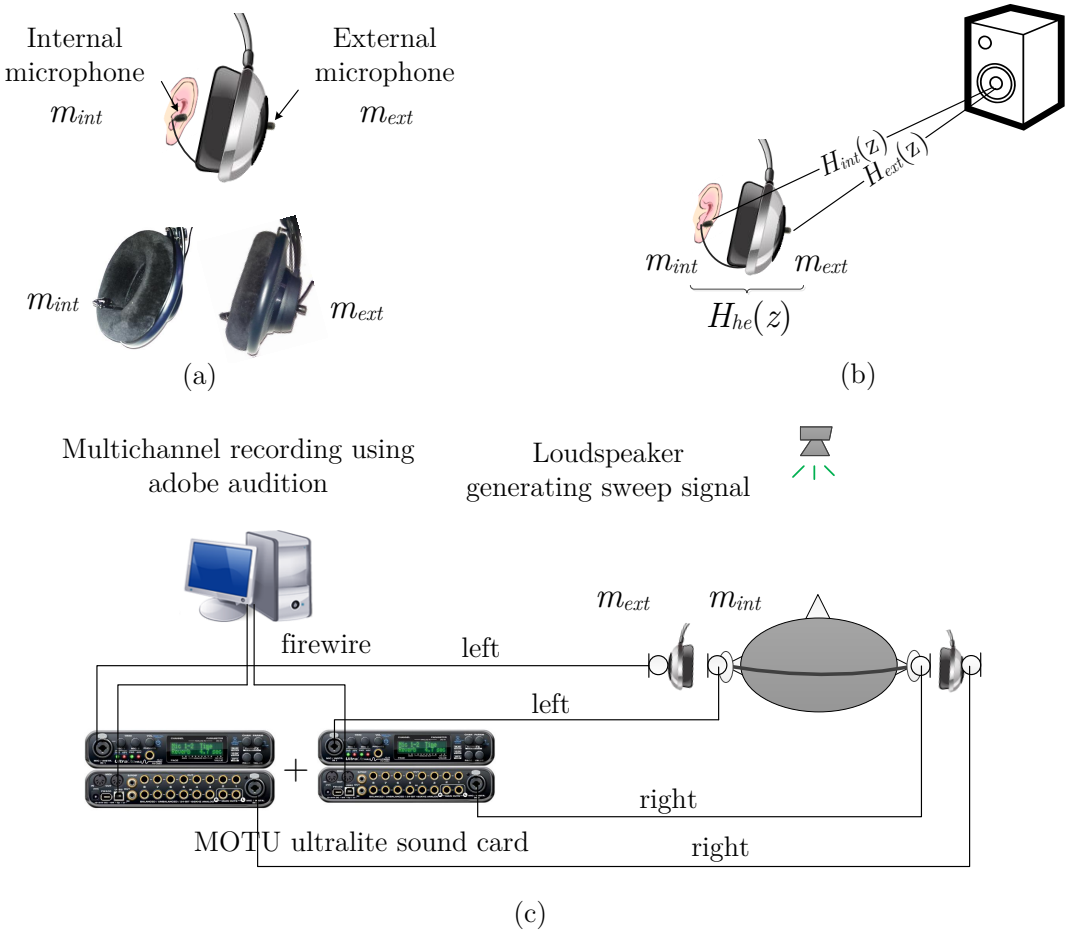


Figure 4.3: (a) Proposed NAR headset structure (top) and prototype using open CA headphones with two microphones (below) (b) Headphone modified transfer functions (HMTF) (c) Transfer functions measurement set up

both pairs of microphones are also used for off-line measurements of the transfer functions modified due to the presence of headphones. These transfer functions are used in the binaural reproduction of virtual sources and to be stored in our own personalized HRTF database for different listening environments. In the next subsection, the characteristics of the headphone modified transfer functions (HMTFs) measured at the two microphone positions are discussed.

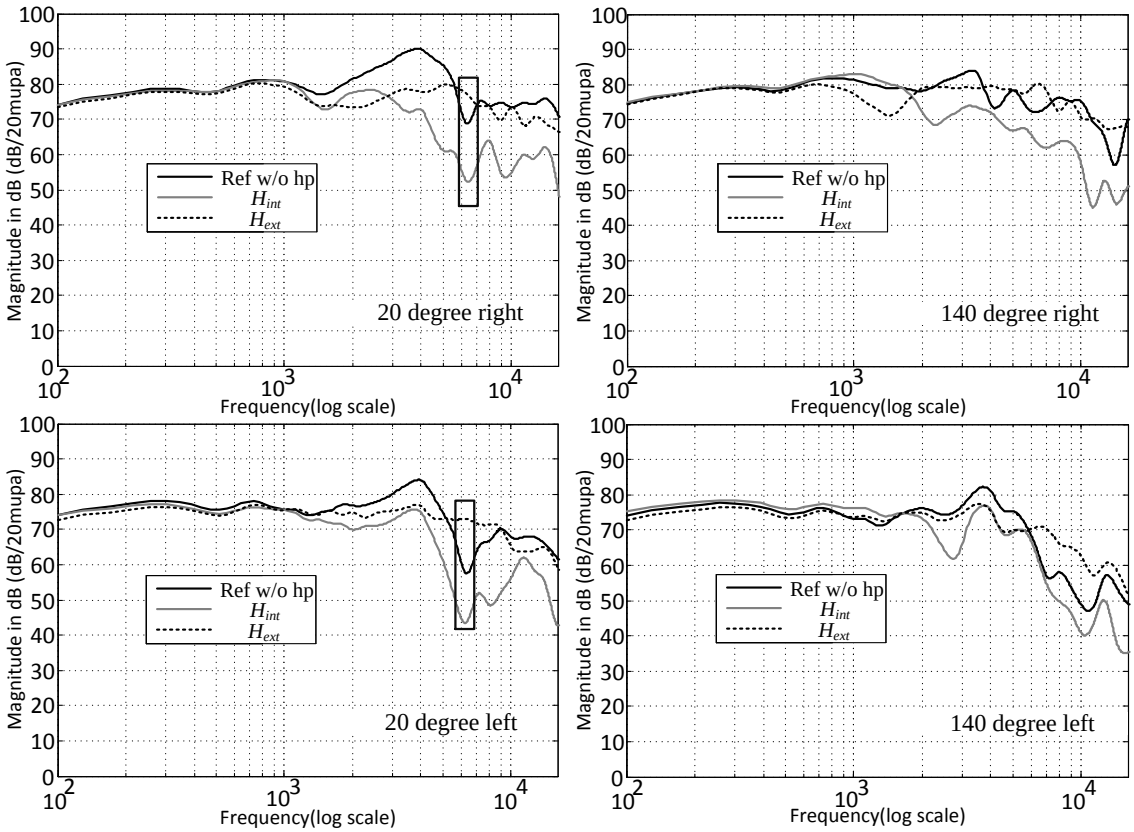


Figure 4.4: Measured modified transfer functions ($H_{int}(z)$and $H_{ext}(z)$) for two azimuths (Top: Ipsilateral ear; Bottom: Contralateral ear)

4.3.2 HMTF measurements and observations

Figure 4.3 shows the measurement setup for HMTFs at the two microphone positions using the NAR headset prototype. Four miniature AKG C417 microphones are used in the measurements having mostly flat response in the frequency range 20-20000Hz. The two HMTFs denoted as $H_{int}(z)$ and $H_{ext}(z)$, (modified due to the passive headphones structure) are measured on the dummy head using the two pair of microphones (See Figure 4.3 (b)). $H_{int}(z)$ represents the transfer function similar to HRTF to account for the sound propagation from source to ear entrance but modified by the presence of headphones, while $H_{ext}(z)$ accounts for the sound propagation from source to the just outside the ear cup.

It should be noted that since $H_{ext}(z)$ is measured just outside the ear cup (~2 cm away from the pinna), its spectrum/impulse response will contain all the indi-

vidual related characteristics and environment without pinna and shell reflections. Spectrums of the two HMTFs for two azimuths are shown in Figure 4.4. In addition, a reference HRTF measured without headphones at the ear canal entrance is also shown for comparison with the two HMTFs. One noticeable observation is that spectrum of $H_{ext}(z)$ is closer to the direct sound spectrum than that of $H_{int}(z)$ with little or no high frequency loss. It should also be noted that sound pressure reaching at internal microphone is modified due to the presence of headphone shell and thus, resulting in slightly lower energy as compared to that of external microphone. Furthermore, there are no pinnae specific frontal notches (especially for the frontal azimuths) observed in $H_{ext}(z)$ as compared to the reference HRTF as well as $H_{int}(z)$. Therefore, the spectrum of $H_{ext}(z)$ is much smoother with only smaller peaks and notches due to the absence of reflections within the shell and the pinnae. In contrast, $H_{int}(z)$ have sharper peaks and notches compared to $H_{ext}(z)$. This prior information in $H_{ext}(z)$ (environment, torso, head related characteristics) can help in estimating the signal accurately at listeners' ears. As will be shown later in the Chapter, $H_{ext}(z)$ is very useful in improving the performance of adaptive equalizer methods presented. In addition, external signals received at m_{ext} are also used to further estimate the real signals at m_{int} adaptively. We now present the analysis for three practical scenarios in the following three sub-sections.

4.3.3 Case I: Only real source present

In this scenario, only real sound sources are present, which is what we experience in day-to-day listening. But in this case, it is required to hear the sounds coming from the environment and external sources, while wearing the NAR headset. This scenario is depicted in Figure 4.5 using an open ear cup headphone along with its corresponding signal flow block diagram representation. Natural sound from real source, $r(n)$ propagates through air and reaches the listeners' ear after passing through the ear cup. Thus, $h_{int}(n)$ (corresponding impulse response of HMTF,

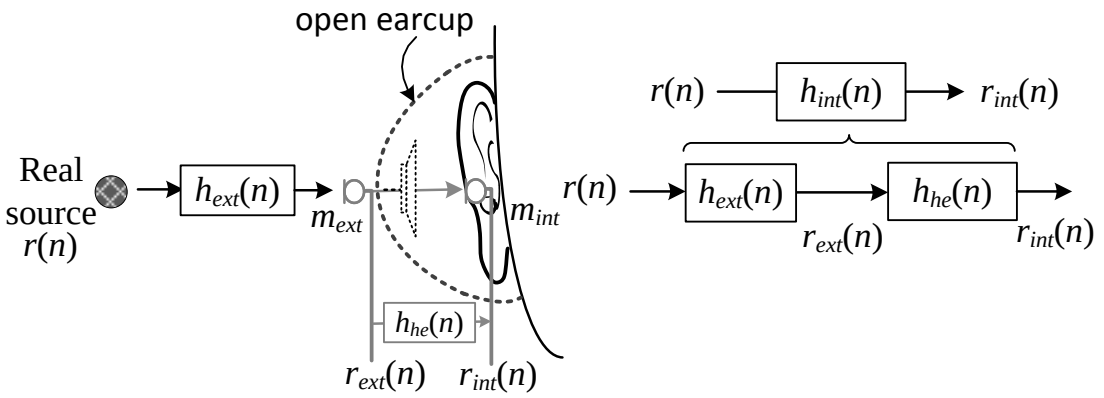


Figure 4.5: Case I: Only real source present scenario and corresponding signal flow block diagram

$H_{int}(z)$) accounts for the natural sound propagation in air from source to the listeners' ear. Alternatively, external sound source propagation can also be seen as real source signal, $r_{ext}(n)$ just outside the ear cup (m_{ext}) passed through a passive headphone-effect filter with impulse response $h_{he}(n)$, accounting for the ear cup effect and pinnae reflections, before reaching listeners' ear. Hence, $H_{he}(z)$ is a transfer function from m_{ext} to m_{int}:

$$H_{he}(z) = \frac{H_{int}(z)}{H_{ext}(z)}. \tag{4.1}$$

For an acoustically transparent headphone, (4.1) does not contain the headphone-effect but is just the free-field transfer function between the two microphone positions without headphones. Ideally, a completely open headphones is best suited for AR based applications as there is no need for any additional processing for natural listening of external sounds, as discussed in Section 4.2. On the other extreme, completely closed headphones can also be used in the NAR headset by capturing the external sounds, process and play back from emitter but at the cost of increased computational load and modified natural sounds content. But this would also mean giving more control to the listener and external sounds can be turned off if external sounds are noisy or unwanted. However, almost all commercial headphones available in the market are in between these two extremes and thus, may require some active

and/or passive techniques to make it closer to perceptually acoustically transparent. Schobben and Aarts showed in [139, 140] that high frequency attenuation in open headphones can be partially compensated by replacing the headphone ear pads by an acoustically transparent foam type material. Making closed headphones acoustically transparent is not straightforward due to direct sound leakage of headphones resulting in comb-filtering effect and very low latency requirement.Fortunately, direct sound reaching eardrum itself is delayed slightly (especially for closed headphones) and it has been found experimentally that with delay of 0.5-3 msec and average attenuation of 12-17 dB for speech, piano and drum signals can still result in inaudible comb-filtering effects [144]. Härmä *et. al.* [136] developed a generic equalization method to compensate for the closed in-ear type headphones isolation by capturing the external sound using an external microphone and playing back through earphone after filtering through an equalization filter. Similar to [136], with the help of external microphone in our NAR headset, high frequencies can be boosted to compensate for the headphones isolation. In our studies, it was found that headphone attenuation depends on source content (i.e., frequency) as well as azimuths (source incoming direction). Therefore, it needs to be investigated if a real-time compensation based on the two microphones signals in NAR headset would be more appropriate for headphone isolation compensation. Real-time compensation for headphones (especially open headphones) would be really hard as direct sound will reach ear drum faster as compared to closed headphones. With today's advanced ADC/DAC conversion latency in the range of tens of microseconds, overall latency with processing is still achievable within the inaudible range of 0.5-3 msec. In addition, attenuation for open headphones mainly occur above 6 kHz with attenuation of 10-15 dB, compensation may only be applied in higher frequencies. It is the subject of further research on perceptual comb-filtering effect only due to high frequencies. Furthermore, detailed perceptual analysis of the headphones effect in long-term listening needs to be studied and listeners can be expected to become

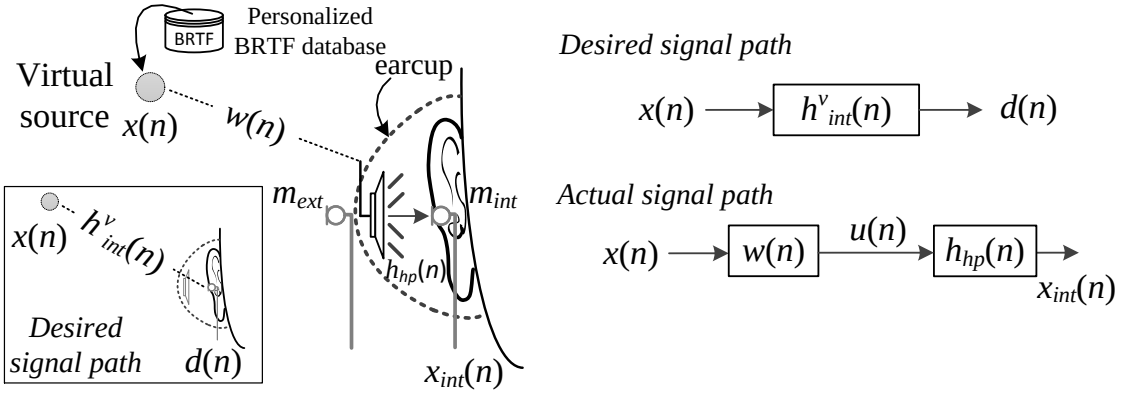


Figure 4.6: Case II: Only Virtual source present scenario and corresponding signal flow block diagram

accustomed to a somewhat modified direct sound spectrum based on past studies [145]. In this work, we primarily focus on the adaptive equalization methods for personalized virtual sound source reproduction over headphones presented in the next two subsections.

4.3.4 Case II: Only virtual source present

To enable natural listening via headphones for virtual sounds, it is required to reproduce exact replica of real signals at listener's ears as in natural listening for external sources. To achieve this, desired binaural room impulse response (BRIR), $h^v_{int}(n)$ (superscript v represents virtual sound reproduction), which must be measured for each and every individual in the same environment as listener's, are required to create an illusion that virtual source is perceived similar to real sources. In addition, $h^v_{int}(n)$ must be equalized to compensate for the individual HPTF, an electroacoustical transfer function measured at the listeners' ears as impulse response $h_{hp}(n)$. HPTFs are also unique for every individual and modify the intended sound to be reproduced in an undesired manner. In a recent study on the effects of headphone compensation in binaural synthesis by Brinkman et al. in [146], it has been found that only individualized headphone compensation is able to completely eliminate the audible high frequency ringing effects as against non-individual

and generic headphone compensation. In short, both individualized desired transfer function ($H^v_{int}(z)$) and individualized headphone equalization are necessary for the NAR headset to accurately replicate the physical sound spectrum virtually.

Figure 4.6 shows the scenario with only virtual source present along with the corresponding signal flow block diagram. A virtual source can be placed anywhere in the virtual auditory environment by convolving monoaural virtual signal, $x(n)$ with the desired BRIR $h^v_{int}(n)$ based on the intended position (direction, distance) of virtual source resulting in desired signal $d(n)$ at m_{int}. With the NAR headset, the virtual source signal, $x(n)$ is first convolved with an equalization filter, $w(n)$, to compute secondary source signal, $u(n)$ and subsequently, passed through the inherent HPTF filter,$h_{hp}(n)$, before reaching listeners' ear. $w(n)$ is estimated as convolution of $h^v_{int}(n)$ and inverse filter of $h_{hp}(n)$ such that the virtually synthesized signal, $x_{int}(n)$ at m_{int} approaches the desired signal, $d(n)$. In this sub-section, we are mainly focusing on the individual binaural headphone compensation assuming that individualized set of BRTFs measured in the listener environment are available in the database. The proposed NAR headset has an advantage over most of the current headphones in the market. This is due to individualized headphone equalization, which is possible because of the two internal microphones attached as $h_{hp}(n)$ can be measured and compensated for every individual. Usually, the headphone equalization requires inversion of the HPTF, which need not necessarily exist and regularization techniques are used to avoid large boosts [147]. But regularization can also convert a causal minimum-phase inverse filter into one with non-minimum phase characteristics, which can create audible distortions like pre-ringing effects [148]. Another widely used alternative and generally the most effective approach is to use adaptive algorithm like, filtered-x least mean square algorithm (FxLMS), where an estimate of $h_{hp}(n)$ is placed in reference signal path for weight update to ensure convergence and stability [149]. This type of adaptive process is termed as adaptive equalization, since equalization filter, $w(n)$ is adapted to any time-varying

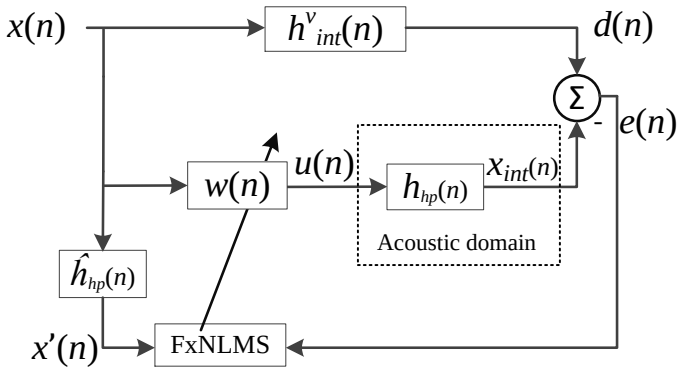


Figure 4.7: Conventional FxNLMS Block Diagram for virtual source reproduction

changes in the individual HPTF due to headphone repositioning or even change of listener. Therefore, fast convergence and minimum steady state mean square error (SS-MSE) of the adaptive process is very crucial for the performance of NAR headset. In addition, minimum spectral distortion (SD) is required to ensure similar spectral variation between the desired and estimated secondary path transfer function, while preserving the pinnae cues crucial for localization. Fast convergence will also ensure that virtual signal captured by the error microphone (m_{int}) converges to the desired signal as quickly as possible. FxLMS usually suffers from slow convergence and can be improved by its normalized version (FxNLMS). Figure 4.7 shows the block diagram of conventional FxNLMS for virtual source reproduction. In the case of adaptive equalization presented in this chapter, signals are electrically subtracted unlike the FxNLMS algorithm used in conventional ANC applications of acoustic duct and ANC headset, where primary signal is acoustically summed at the error microphone. The FxNLMS algorithm is expressed below:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu \frac{\mathbf{x}'(n)}{\|\mathbf{x}'(n)\|^2} e(n) \ , \tag{4.2}$$

where $\mathbf{w}(n)$ is the coefficient vector of $w(n)$ with length L, and $\mathbf{x}'(n) = [x'(n)\, x'(n-1) \ldots x'(n-L+1)]^T$ is the set of current and past $(L-1)$ samples of filtered reference signal $x^{'}(n)$ at time n.$\| \cdot \|^2$ represents the norm-2 of the vector. Optimum value of the equalization filter $w(n)$ is achieved when the expectation of the squared error,

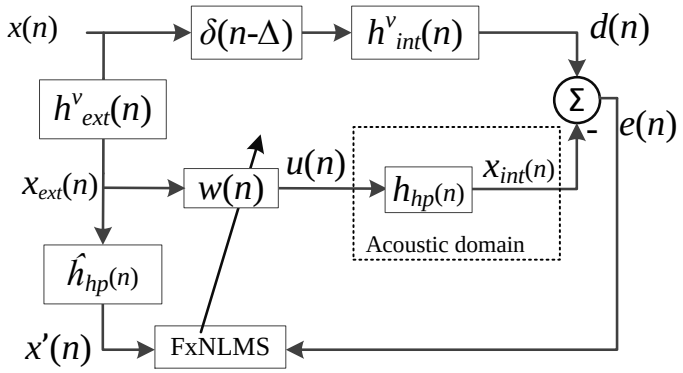


Figure 4.8: Block Diagram of modified FxNLMS for virtual source reproduction

$e(n)$ approaches zero and can be found as:

$$W^o(z) = \frac{H^v_{int}(z)}{H_{hp}(z)} \, . \tag{4.3}$$

It should be noted that the required number of filter taps for the equalization filter can be large because of the fact that desired signal, $d(n)$, may contain the additional delay due to the distance and room acoustics stored in the desired impulse response, $h^v_{int}(n)$. As larger filter taps will improve the accuracy of the adaptive process by closely following the desired signal, it also slows down the convergence rate at the same time. Since fast convergence being one of the stringent requirements for our system performance with the SS-MSE, we propose a modified version of FxNLMS, as shown in Figure 4.8. The secondary path of the adaptive process is modified by including an additional filter, $h^v_{ext}(n)$, and a forward delay (Δ) is also introduced in the primary path to take into account for the overall delay (A/D, D/A, processing) of the secondary path. As discussed in the subsection 4.3.2, transfer function measured just outside the ear cup, $H^v_{ext}(z)$, contains all the spatial information of the human torso, head, as well as environments without the pinnae and headphone shell reflections. By using this prior-information in estimating the desired signal, the adaptive process is simplified with shorter adaptive filter length and subsequently, faster convergence. As compared to the conventional FxNLMS approach, virtual signal is first pre-filtered with $h^v_{ext}(n)$ before passing to the equalization filter $w(n)$.

Using this approach, we are trying to emulate the natural listening process by using an estimate of virtual signal at m_{ext}, i.e.$x_{ext}(n)$, to reproduce the replica of real sound alike at listeners' ear, as shown in Figure 4.8. Equalization filter weights will be optimum when square of the residual error is minimized:

$$E\left[e^2(n)\right] = 0 \implies d(n) - x_{int}(n) = 0, \tag{4.4}$$

where $d(n)$ is defined as:

$$d(n) = h^v_{int}(n) * x(n - \Delta). \tag{4.5}$$

Substituting (4.5) into (4.4) and transforming into Z domain:

$$X(z)H^v_{int}(z)z^{-\Delta} - X(z)H^v_{ext}(z)W^o(z)H_{hp}(z) = 0. \tag{4.6}$$

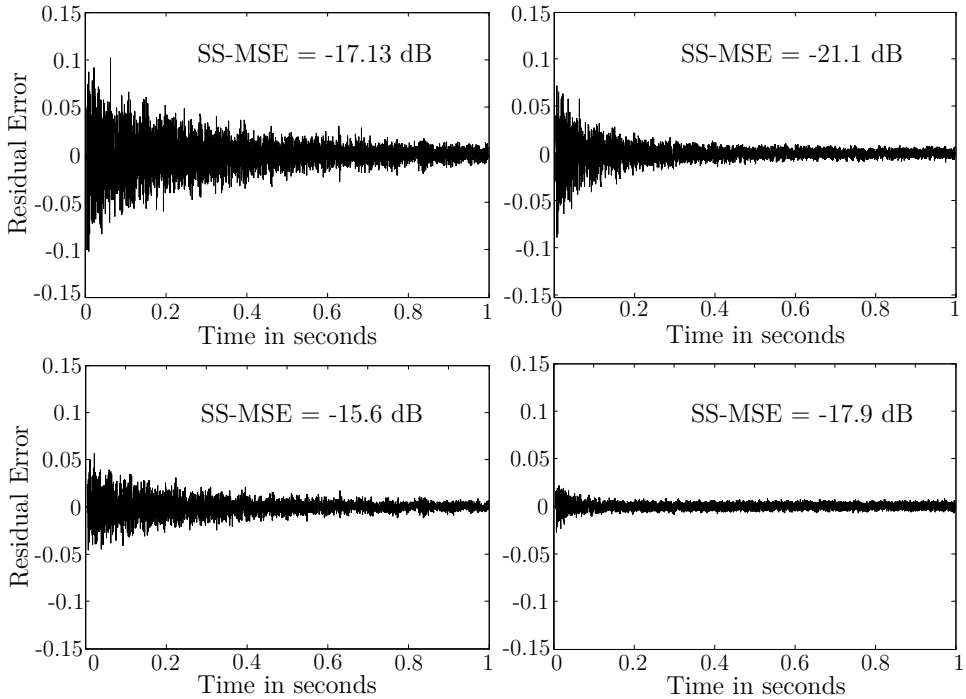
By simplifying (4.6), the optimum equalization filter can be expressed as:

$$W^o(z) = \frac{H^v_{int}(z)z^{-\Delta}}{H^v_{ext}(z)H_{hp}(z)} = \frac{H^v_{he}(z)z^{-\Delta}}{H_{hp}(z)}, \tag{4.7}$$

where $H^v_{he}(z)$ is the headphone effect transfer function for virtual sound reproduction denoted by the superscript, v. Therefore, the difference between the optimal solution of conventional FxNLMS in (4.3) and (4.7) is due to the filter $h^v_{ext}(n)$ and the forward delay in primary path. Delay in the primary path must be at least equal to the secondary path delay for a feed-forward adaptive filter to converge [149]. Weight update equation for the modified FxNLMS approach is expressed similarly as in (4.2), except that the filtered reference signal is now defined as follows:

$$x'(n) = \hat{h}_{hp}(n) * x_{ext}(n), \tag{4.8}$$

where $\hat{h}_{hp}(n)$ is an estimate of the secondary path transfer function (HPTF), which is estimated offline by playing a test sequence through the headset and recording the



(a) Conventional FxNLMS (b) Modified FxNLMS

Figure 4.9: Comparison between Conventional FxNLMS and Modified FxNLMS for 40^o azimuth (Top: Ipsilateral ear; Bottom: Contralateral ear)

response at internal microphone. As in ANC, FxNLMS algorithm converges with the limit of 90^o phase error between $\hat{h}_{hp}(n)$ and $h_{hp}(n)$ [149].

4.3.4.1 Case II results: Conventional FxNLMS Vs Modified FxNLMS

In this sub-section, we compare the performance of the modified FxNLMS with the conventional FxNLMS method. A white noise sequence of 1 second duration is used to estimate the adaptive filter in both methods. Length of impulse responses, $h^v_{int}(n)$ and $h^v_{ext}(n)$, are set at 1024 taps, and 256 taps are used for $h_{hp}(n)$. Longer filter length for the desired responses is required to account for the distance and reverberations. The equalization filter lengths are set at 1024 taps and 256 taps for the conventional FxNLMS and the modified FxNLMS, respectively. A step size of 0.1 is chosen for both the algorithms.

Figure 4.9 compares the performance of the two approaches. Three main performance criteria used in this chapter are the faster convergence rate, minimum

SS-MSE, and minimum SD. As shown in Figure 4.9, conventional FxNLMS suffers from slow convergence rate as expected. SS-MSE for both the approaches do not differ much from each other, as can be seen in Figure 4.9 (a) and (b). Besides, we can also objectively quantify the spectral error between reference and estimated transfer functions using a widely used objective metric [150, 151] i.e., spectral distortion (SD) score:

$$SD = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(20\log\frac{|H(f_k)|}{|\widehat{H}(f_k)|}\right)^2}\,[dB], \tag{4.9}$$

where $H(f)$ is the magnitude response of reference primary path transfer function, $\widehat{H}(f)$ is the secondary path estimated transfer function, and K is the total number of frequency samples in the observed range (100 Hz – 16 kHz). Secondary path transfer functions for conventional FxNLMS and modified FxNLMS are expressed, respectively as:

$$S_{conv}(z) = W_{conv}(z)H_{hp}(z)\,, \quad and \tag{4.10}$$

$$S_{mod}(z) = H^{v}_{ext}(z)W_{mod}(z)H_{hp}(z). \tag{4.11}$$

Figure 4.10 shows the spectral distortion scores for low frequency (below 1.5 kHz) and high frequency (above 1.5 kHz). To clearly demonstrate the difference between the two approaches, simulation is stopped after 0.2 second and SD scores using (4.9) are computed at this instance. It is clearly observed that mean spectral distortion for the conventional FxNLMS is much higher than the modified FxNLMS, especially at low frequencies. Even at higher frequencies, modified FxNLMS has higher accuracy than the conventional FxNLMS for most of the azimuths except for some source positions. It should also be noted that spectral distortion is considerably greater for ipsilateral ear than the contralateral ear for both approaches. This might be due to the pinna effects being more pronounced at the ipsilateral ear. Although, the

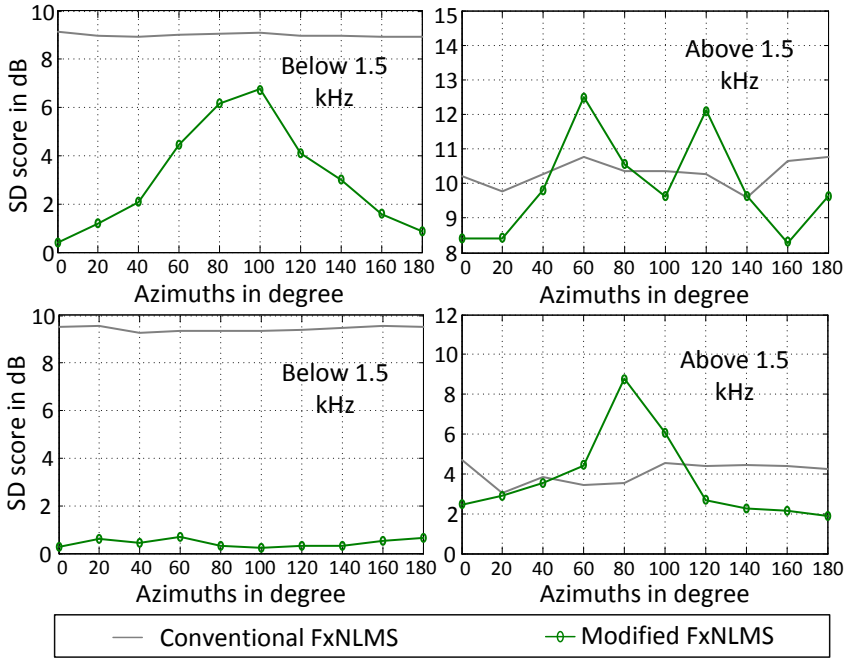


Figure 4.10: Spectral distortion comparisons for both the approaches over 0 to 180^o azimuths (Top: Ipsilateral ear; Bottom: Contralateral ear)

modified FxNLMS has faster convergence rate as well as better accuracy for most azimuths, larger spectral deviations in higher frequencies can significantly affect the NAR headset performance and may result in higher SS-MSE for some of the azimuths. Based on above observations, a hybrid adaptive equalizer (HAE) is also proposed by combining both the above approaches to obtain an optimum steady state performance of the adaptive algorithm for all azimuths.

4.3.4.2 Hybrid Adaptive Equalizer (Hybrid FxNLMS)

The conventional FxNLMS suffers from slow convergence and generally requires large filter order for equalization filter to converge to the optimum solution. The modified FxNLMS uses an additional pre-filter in secondary path, which contains most of the spatial information of the primary path. This ensures faster convergence of the adaptive process. High spectral distortions have been observed for the conventional FxNLMS in low frequency, while modified FxNLMS has relatively larger errors in high frequency regions for some source positions. The proposed hybrid

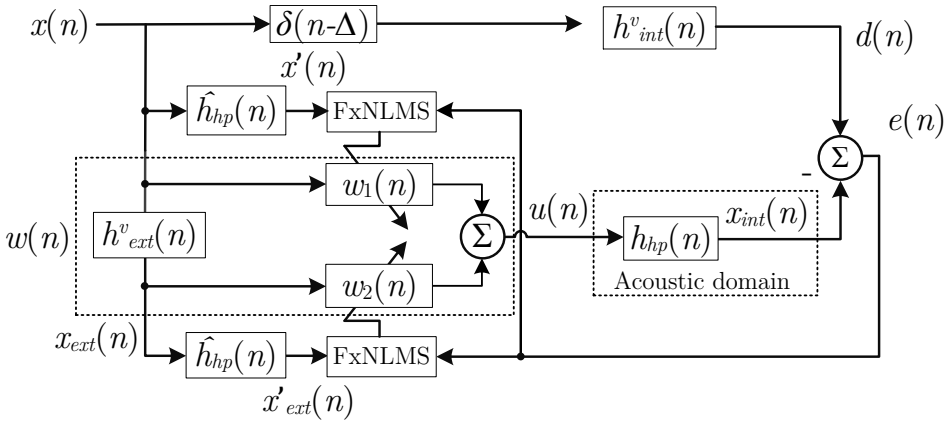


Figure 4.11: Block diagram of hybrid FxNLMS using conventional FxNLMS and modified FxNLMS algorithm

FxNLMS uses simple combination of both the conventional and modified FxNLMS structures discussed above, which can result in significant steady state performance improvements as well as fast convergence most of the times [152]. The block diagram for the hybrid FxNLMS is shown in Figure 4.11. The secondary source signal $u(n)$ is generated using outputs of both the conventional FxNLMS equalization filter $w_1(n)$ and the modified FxNLMS equalization filter $w_2(n)$. As shown in Figure 4.11, the equivalent equalization filter $w(n)$ has two reference inputs: $x(n)$ as the virtual signal, and $x_{ext}(n)$ is the virtual signal estimated at the reference microphone (m_{ext}). Filtered versions of the two reference signals $x'(n)$ and $x'_{ext}(n)$ are used to adapt the filter coefficients $w_1(n)$ and $w_2(n)$, respectively.

The secondary signal $u(n)$ is computed by the equivalent equalization filter $w(n)$, which consists of the two adaptive filters' length of L_1 and L_2, respectively, for $w_1(n)$ and $w_2(n)$ as:

$$u(n) = \mathbf{w_1}^T(n)x(n) + \mathbf{w_2}^T(n)x_{ext}(n)\,, \tag{4.12}$$

where

$$\mathbf{x}(n) = [x(n)\; x(n-1).\ldots\ldots\ldots..x(n-L_1+1)]^T\,, \tag{4.13}$$

and

$$\mathbf{x_{ext}}(n) = [x_{ext}(n)\, x_{ext}(n-1).....x_{ext}(n-L_2+1)]^T . \tag{4.14}$$

The hybrid FxNLMS algorithm for the weight update of the two filters is expressed as:

$$\mathbf{w_1}(n+1) = \mathbf{w_1}(n) + \mu \frac{\mathbf{x'}(n)}{\|\mathbf{x'}(n)\|^2} e(n) . \tag{4.15}$$

and

$$\mathbf{w_2}(n+1) = \mathbf{w_2}(n) + \mu \frac{\mathbf{x'_{ext}}(n)}{\|\mathbf{x'_{ext}}(n)\|^2} e(n) . \tag{4.16}$$

Weight update equation (4.15) corresponds to the conventional FxNLMS with only difference in the calculation of residual error signal (delayed version) as defined by (4.4) and (4.5), while weight update for $w_2(n)$ is same as the modified FxNLMS. The main purpose of the hybrid approach is to take advantage of both the adaptive processes so as to minimize the residual error. Ideal solution for $w(n)$ is derived using (4.4) and (4.5) as:

$$X(z)H^v_{int}(z)z^{-\Delta} - U(z)H_{hp}(z) = 0. \tag{4.17}$$

Taking Z transform of (4.12) and substituting into (4.17) results in

$$X(z)H^v_{int}(z)z^{-\Delta} - X(z)W^o(z)H_{hp}(z) = 0. \tag{4.18}$$

$W(z)$ is an equivalent equalization filter representation for the HAE and expressed as:

$$W(z) = W_1(z) + H^v_{ext}(z)W_2(z). \tag{4.19}$$

Thus from (4.18), the optimal solution of equivalent adaptive filter, $W^o(z)$ is similar to that of conventional FxNLMS with an additional delay term:

$$W^o(z) = \frac{H^v_{int}(z)z^{-\triangle}}{H_{hp}(z)} \ . \tag{4.20}$$

Therefore, the optimal solution of the hybrid FxNLMS can be written as linear combination of optimal solutions $W^o_1(z)$ and $W^o_2(z)$ for the conventional and modified FxNLMS approach, respectively:

$$W^o(z) = \alpha W^o_1(z) + \beta H^v_{ext}(z) W^o_2(z). \tag{4.21}$$

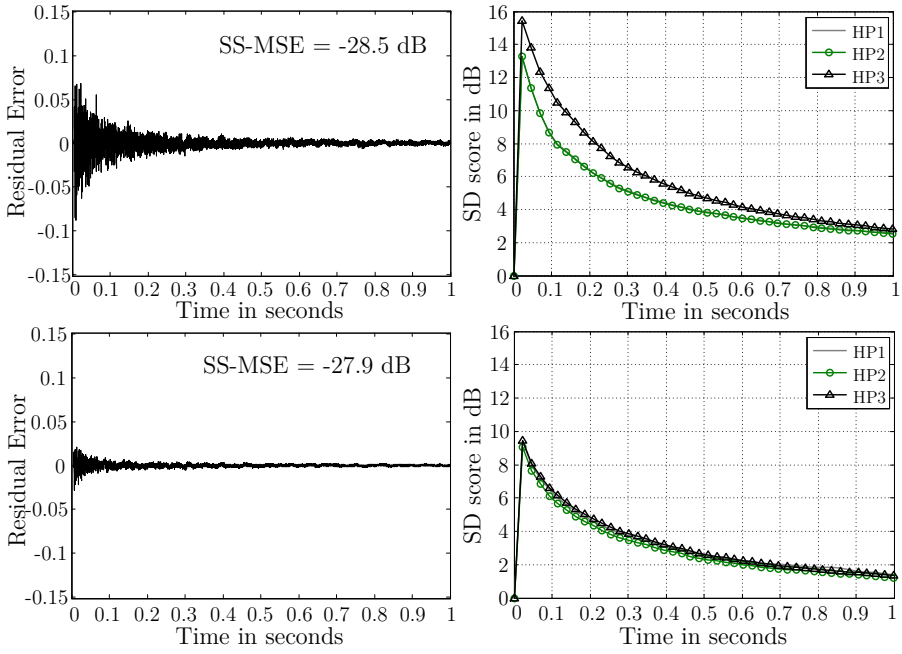
such that,

$$\alpha + \beta = 1; \qquad 0 \leq \alpha, \beta \leq 1 \tag{4.22}$$

The values of α and β are inherently determined by the adaptive algorithm such that the residual error is minimized. Next, we will discuss the performance of the presented HAE and compare the results with the conventional FxNLMS and modified FxNLMS.

4.3.4.3 Case II Results: Hybrid FxNLMS Vs Others

In this subsection, we compare the performance of the proposed hybrid FxNLMS with the conventional and modified FxNLMS algorithms. Same number of taps for the two filters, $W^o_1(z)$ and $W^o_2(z)$ are used, as in subsubsection 4.3.4.1. Figure 4.12(a) shows the residual error for the hybrid FxNLMS. Comparing the results with that of Figure 4.9, the hybrid FxNLMS performs much better as compared to conventional and modified FxNLMS with optimum MS-SSE. Moreover, its convergence rate is also much faster than the conventional FxNLMS but slightly slower than the modified FxNLMS. Spectral distortion scores versus time plots for three headphone placements (HP1-3) are shown in Figure 4.12(b). For the first two head-



(a) Residual error vs time

(b) Spectral distortion vs time for three headphone placements (HP1-3)

Figure 4.12: Hybrid adaptive equalizer performance for 40^o azimuth (Top: Ipsilateral ear; Bottom: Contralateral ear)

phone placements, headphone was slightly adjusted, while in the third placement headphone was lifted and placed back on the ears. Clearly, the proposed hybrid approach is also robust to headphone placements while its SD converges to the optimal solution in all cases. It was also found that adaptive equalization performs better than fixed headphone equalization with 7-8 dB higher error reduction. Spectral distortion scores for the three approaches are shown in Figure 4.13 computed at two time instants of 0.2 second and 1 second. Since conventional FxNLMS has the slowest convergence rate, the residual error cannot be completely converged after 0.2 second and results in larger spectral distortion. On the other hand, hybrid approach has the least spectral distortion, as shown in the Figure 4.13. For longer noise sequence of 1 second, when steady state performance is reached, relatively larger spectral differences is observed between conventional and modified FxNLMS approach, whereas the hybrid FxNLMS has the best overall performance, as shown in the right side plots of Figure 4.13. Mean steady state error attenuation for the

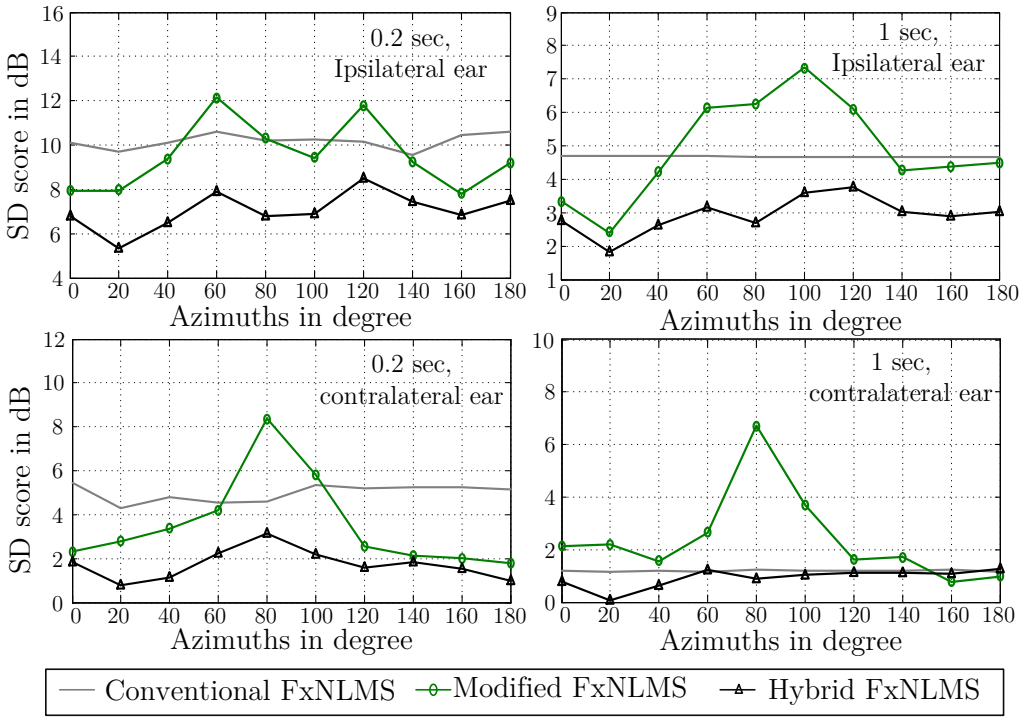


Figure 4.13: Spectral distortion score comparisons: Hybrid FxNLMS versus others (Top: Ipsilateral ear; Bottom: Contralateral ear)

hybrid FxNLMS across all azimuths was found to be around 25 dB and 28 dB for the ipsilateral and contralateral ears, respectively. Finally, we show the SD scores for elevated sources in Figure 4.14. Target impulse responses for the adaptive equalization were measured at $0^o, 15^o, 30^o$ and 45^oelevated positions for 4 different azimuth positions ($0^o, 40^o, 80^o$ and 120^o). As shown, mean SD score for the ipsilateral ear for all the angles is within 5 dB, while it is less than 2 dB for the contralateral ear. Spectral distortion is clearly more pronounced for the ipsilateral ear, especially when source is one side of the dummy head and directly facing the ipsilateral ear (See Figure 4.14 (b) for 80^o azimuth). Due to the fast converging speed of the hybrid FxNLMS, estimated responses are closely tallied to the desired response for most of the source positions with optimum SS-MSE and SD. The high frequency pinnae cues, which are primary cues for the sources in the front as well as elevation, have also been preserved in the virtually synthesized responses.

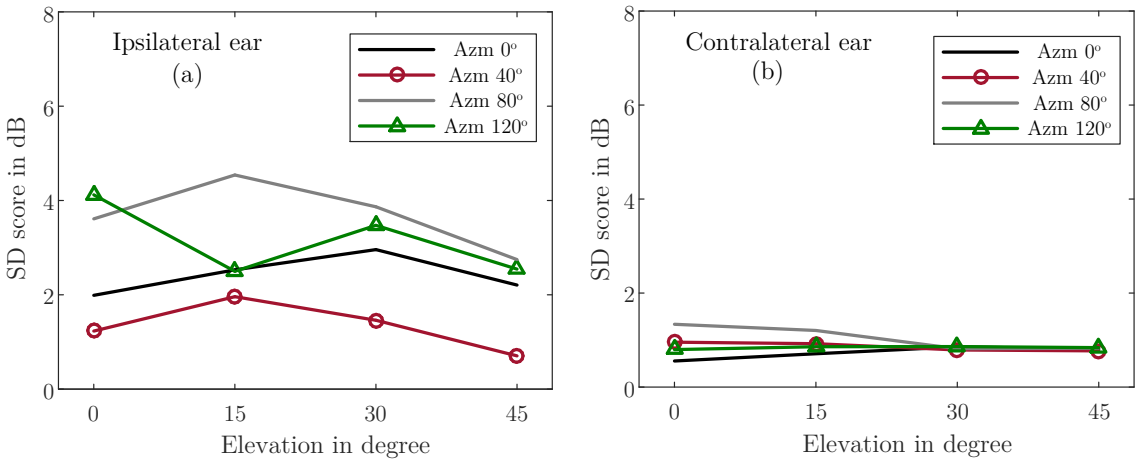


Figure 4.14: Spectral distortion score for Hybrid FxNLMS for elevated sources

4.3.5 Case III: Both virtual and real source present (Augmented reality): HAE with online adaptive estimation

In this section, we present the most general case for augmented reality with virtual and real sounds being intermixed. As explained earlier, augmented reality requires virtual sources to be coherently fused with the real source and surroundings such as to create an illusion of virtual sources being perceived as one of the real sound sources. Figure 4.15 shows this scenario along with the corresponding signal flow block diagram. In an ideal case with virtually no headphones present, virtual source after passing through the target response is acoustically added with real signals reaching the listeners' ear, as shown in Figure 4.15. But the HPTF colors the intended sound spectrum and thus, virtual signal must be equalized before playing back through the headphones. We presented the HAE for virtual source reproduction in the previous section, ensuring that there is no difference between the real and virtual source signals. In this scenario with NAR headset, real signal is also captured by the internal error microphone (m_{int}) simultaneously with the synthesized virtual signal, as shown in the block diagram of Figure 4.15. In addition to the external sounds, leakage signal, $l(n)$ from inside of headset is also captured by the external microphone, with $h_{le}(n)$ as the headphone leakage impulse response

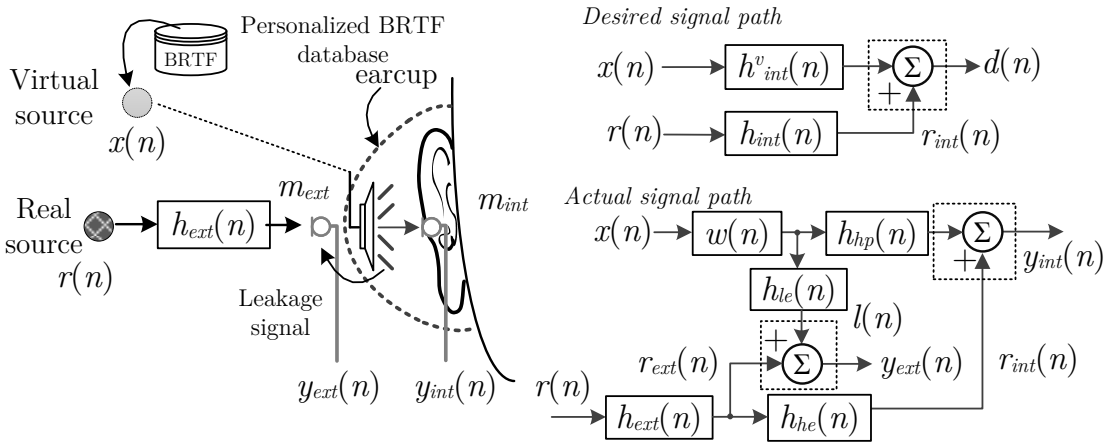
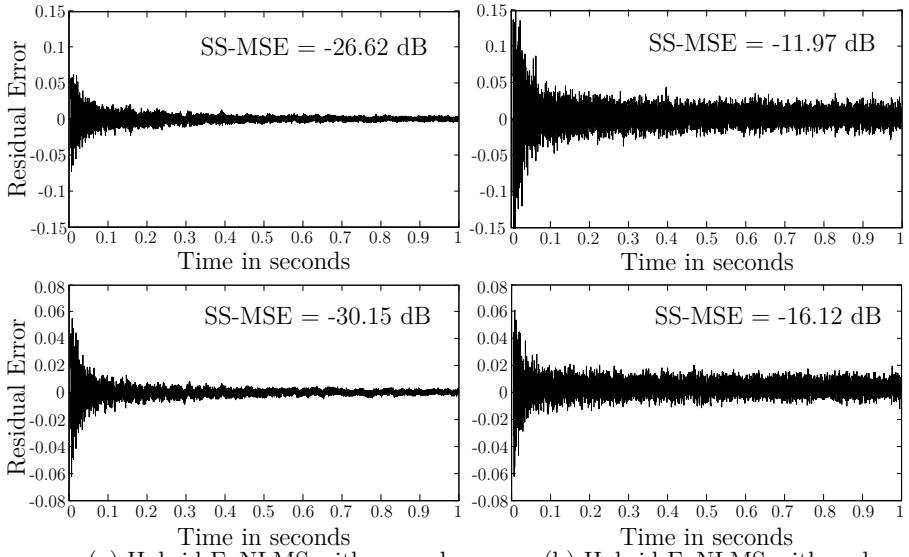


Figure 4.15: Case III: Both virtual and real source present scenario and corresponding signal flow block diagram

measured at m_{ext}. Leakage signal is assumed to be negligible with more than 20 dB attenuation to playback level at ear opening. However, the real signal $r_{int}(n)$ can hamper the adaptive equalization for $w(n)$ by acting as interference to the system. Figure 4.16 shows the plots for residual error of the hybrid FxNLMS with and without real source present. Two uncorrelated random noise sequences are used in simulation for the virtual and real source. Clearly, due to the presence of real signals, the hybrid FxNLMS cannot reach the optimum solution and subsequently, resulting in roughly 14 dB lesser reduction in steady state than the case with no real source. Therefore, the effect of real signal must be removed from the adaptive process; otherwise it might result in large steady state error depending on the energy and nature of external signals.

In augmented reality, both real and virtual sounds are equally crucial for an immersive experience and therefore, either of them must not interfere with each other to reproduce a natural superposition. In this respect, the acquired real signal $r_{int}(n)$ must be removed from the adaptive process of $w(n)$. There are two ways to compute an estimate of the signal $r_{int}(n)$ using the real signal received at m_{ext}, i.e. $r_{ext}(n)$:

1. With the help of pre-computed $h_{he}(n)$: As explained in subsection 4.3.3, $h_{he}(n)$ represents the headphone-effect transfer function from m_{ext} to m_{int}, an exact



(a) Hybrid FxNLMS with no real source present

(b) Hybrid FxNLMS with real source present

Figure 4.16: Residual error plots for hybrid FxNLMS with and without real source. Virtual source is positioned at 0^o azimuth, while real sound is coming from 40^o azimuth and added to virtually reproduced signal at m_{int}. (Top: Ipsilateral ear; Bottom: Contralateral ear)

estimate of $r_{int}(n)$ is computed from $r_{ext}(n)$ as:

$$r_{int}(n) = r_{ext}(n) * h_{he}(n). \tag{4.23}$$

But in practice, the precise location of external sound is not known and hence, a filter averaged over entire azimuths has to be used instead of the exact $h_{he}(n)$ as:

$$r_{int}(n) = r_{ext}(n) * h_{he,avg}(n). \tag{4.24}$$

The headphone-effect transfer function $h_{he}(n)$ is computed as off-line adaptive estimation using the FxNLMS algorithm.

2. Using an online adaptive process to estimate $r_{int}(n)$ from $r_{ext}(n)$: It has been observed that $h_{he}(n)$ varies considerably with head movements. Therefore, online adaptive estimation of $h_{he}(n)$ can give better estimate of $r_{int}(n)$ instead of using an average filter $h_{he,avg}(n)$.

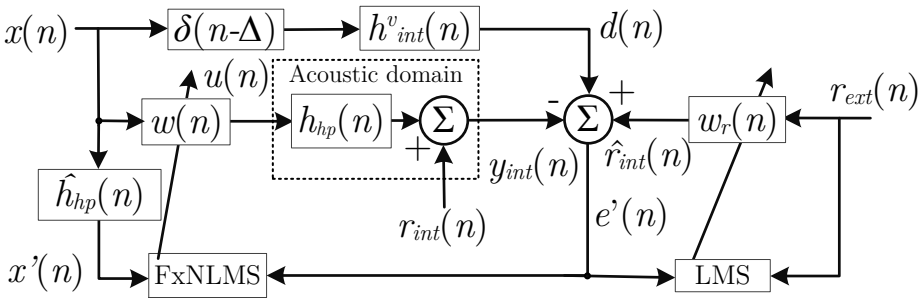


Figure 4.17: Block diagram of hybrid adaptive equalizer with online adaptive estimation of $r_{int}(n)$

We further, extend the hybrid FxNLMS with online adaptive estimation of real signal, as shown in Figure 4.17. The adaptive equalization filter $w(n)$ is the equivalent representation of the HAE given by (4.19). As discussed in the previous section, equalization filter $w(n)$ comprises of two adaptive filters $w_1(n)$ and $w_2(n)$ corresponding to the conventional FxNLMS and modified FxNLMS algorithms, respectively. As shown in Figure 4.17, $w_r(n)$ is adapted to generate an estimate of $r_{int}(n)$, $\widehat{r}_{int}(n)$ and added to $d(n)$ from which the acoustically superimposed signal $y_{int}(n)$ is subtracted. After $\widehat{r}_{int}(n)$ has converged, we obtain the residual error signal as:

$$e'(n) = \{d(n) + \widehat{r}_{int}(n)\} - y_{int}(n)\,, \tag{4.25}$$

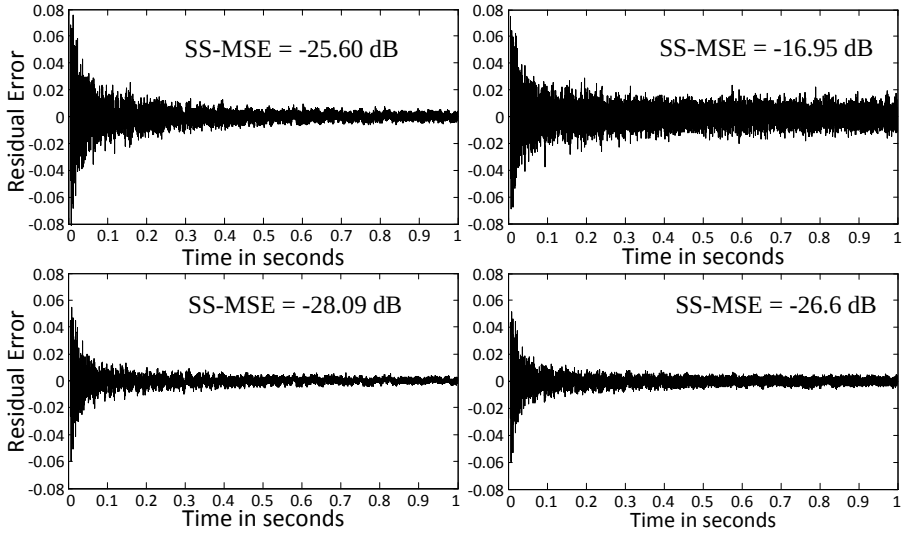
where $y_{int}(n)$ is defined as

$$y_{int}(n) = r_{int}(n) + x_{int}(n). \tag{4.26}$$

Substituting (4.26) into (4.25) and re-arranging,

$$e'(n) = \{d(n) - x_{int}(n)\} + \{-r_{int}(n) + \widehat{r}_{int}(n)\}\,. \tag{4.27}$$

Hence, the residual error signal consists of two separate error signals. The first term in RHS of (4.27) is the error signal for hybrid adaptive process defined by $e(n)$ in (4.4), while the second term is the negative error signal due to the online



(a) Hybrid FxNLMS with adaptive estimation

(b) Hybrid FxNLMS with off-line estimation

Figure 4.18: Results for hybrid FxNLMS with and without adaptive estimation. Simulation set up is kept same as in (Figure 4.16). (Top: Ipsilateral ear; Bottom: Contralateral ear)

adaptive estimation of $\widehat{r}_{int}(n)$. The optimum solution of adaptive estimation process is derived when $\widehat{r}_{int}(n) = r_{int}(n)$, or

$$r_{ext}(n) * w_r(n) = r_{int}(n). \tag{4.28}$$

Taking the z-transform of (4.28) , the optimum control filter $w_r(n)$ is expressed as

$$W_r^o(z) = \frac{R_{int}(z)}{R_{ext}(z)} = \frac{H_{int}(z)}{H_{ext}(z)} = H_{he}(z). \tag{4.29}$$

Thus, optimum control filter is simply the headphone-effect impulse response. Weight update equation for the two control filters in HAE is defined as in (4.15) and (4.16), whereas weight update equation for the control filter $w_r(n)$ is defined using the LMS algorithm as

$$\mathbf{w_r}(n+1) = \mathbf{w_r}(n) - \mu_r \, \mathbf{r_{ext}}(n) e'(n). \tag{4.30}$$

Note that negative sign in the weight update equation (4.30) is due to the way

error signal is defined in (4.27). Figure 4.18 shows the performance comparison of the HAE with and without adaptive estimation. Clearly, with the proposed adaptive estimation, the performance of the hybrid FxNLMS is very close to the one without any real source present, as observed in Figure 4.16(a) and Figure 4.18(a). With off-line estimation i.e., using an average filter, the steady state error increases especially for the ipsilateral ear. However, the approach with offline estimation still performs much better than the one without any estimation (See Figure 4.16(b) and Figure 4.18(b)). A perceptual validation of the hybrid FxNLMS is carried out via subjective study, which is explained next in the next section.

4.4 Listening Test

The goal of the NAR headset is to reproduce augmented reality contents such that users cannot distinguish whether the sounds are coming from physical sources/environments or from the NAR headset. A listening test was conducted to subjectively validate the proposed HAE approach using individualized BRIRs. Three main research questions were asked in following listening tests:

- **Naturalness:** Does virtual sound perceive natural?
- **Sound similarity:** Does virtual sound perceive similar to the real source i.e., sound coming from physical speakers?
- **Source position similarity:** How close is the virtual source position in 3D space as compared to real source?

The setup used for the listening test is shown in Figure 4.19. Listener wearing the NAR headset prototype is surrounded by 7 Genelec 1030A loudspeakers. Five of the speakers are in azimuth plane (3 in the front and 2 in the rear), while two speakers are elevated at 30^o in the front hemisphere. All the loudspeakers were positioned at a distance of 1.2 m away from the center of listener's head. Two MOTU Ultralite

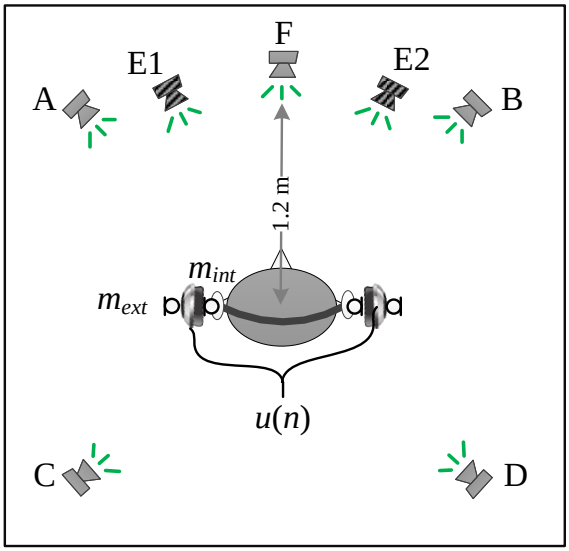


Figure 4.19: Listening test setup (: Elevated speaker; : Azimuth speaker)

soundcards were used to interface with the 7 loudspeakers, 2 channels of AKG K702 headphones and 4 AKG C417 microphones. Three different listening sets were carried out as follows:

- **SET *1***: Perceptual similarity test between speaker and headphone playback of a male speech signal.
- **SET *2***: Perceptual similarity test between real and virtual mixing of two male speech signals.
- **SET *3***: Perceptual similarity test between real and virtual superposition of a speech signal with ambient sound.

Individualized BRIRs were measured for each of the seven speakers at both the binaural microphones' positions (m_{int} and m_{ext}) attached through the NAR headset, as shown in Figure 4.19. Head tracker was mounted on the NAR headset to help subjects to maintain still head position during measurement process. Subjects were asked to repeat the measurement if they moved their head more than 5^o in any of the three degrees of freedom. Individual HPTFs were compensated for with the measured individualized BRIRs. White noise sequence was used to train the

adaptive filters offline using the proposed hybrid FxNLMS presented in this work. In each of the sets, listeners were presented with a pair of stimuli, one of them is played over physical speakers, while other can be either played over headphone as virtual source or physical speaker serving as hidden anchor.

In SET *1*, a 4 second male speech signal was used to evaluate the similarity between real and virtual playback. Speech signal is played through all the seven loudspeakers for real playback, while same speech signal is convolved with the headphone equalized filters for virtual playback for the left and right ears using (4.12). Hybrid FxNLMS for Case II presented in subsection 4.3.4 was used to obtain the equalized filters. The virtually synthesized secondary source signal, $u(n)$ for both the ears was subsequently played back over headphones, as shown in Figure 4.19. Two additional pairs were included in this set as hidden anchors with both the stimuli played over speakers, resulting in a total of 9 test pairs.

In SET *2*, a scenario is created, where two persons are having a conversation. Thus, two male speech signals (each around 3.5 seconds long) were played back from two different directions one after another, thereby merging the two signals. Three pairs of loudspeaker configurations were chosen for the playback, namely, front left-front right (A-B), rear left-rear right (C-D), and elevated left-elevated right (E1-E2). For real playback, the two speech signals are played through each of the 3 loudspeaker pairs. For virtual playback, first speech was played through speaker, while second speech was played over headphones, and vice-versa. Thus, two virtually synthesized tracks were computed for each set of three pairs, in which one of them is played over speaker and the other is rendered virtually over headphones, while keeping the order of speech signals fixed. Thus, a total of 8 virtual signals were used in SET *2* including two hidden anchor pairs of both real sounds.

In SET *3*, a male speech signal is superimposed onto ambient sounds of length around 6 seconds. In this scenario, two configurations are chosen for the superposition of speech signal. In the first configuration, ambient signal is played over two

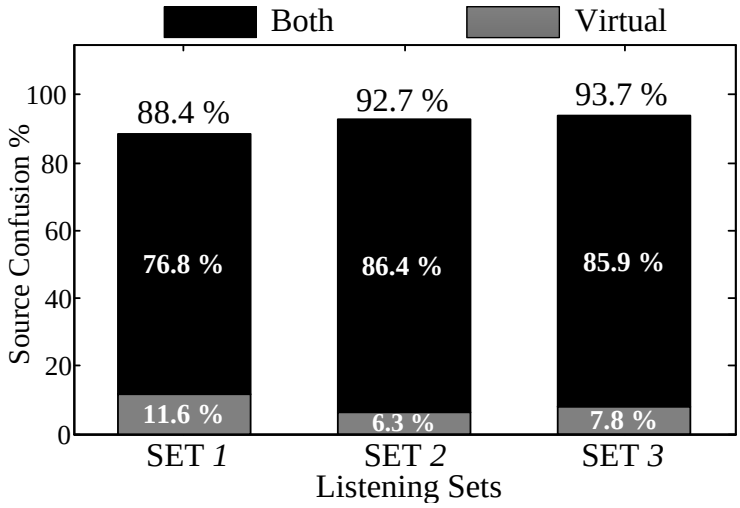


Figure 4.20: Source confusion % for the three listening sets

frontal loudspeakers A and B, while all four surrounding loudspeakers in horizontal plane (A, B, C and D) were chosen for the ambient signal playback in second configuration. Speech signal was played from front loudspeaker position, F for both real and virtual playback. For real playback, both the speech and ambient sound track is played through the loudspeakers. For virtual playback, the proposed hybrid FxNLMS with adaptive estimation of real ambient signals presented for Case III in subsection 4.3.5, was used to obtain the headphone equalized filters. The pre-recorded ambient signals ($r_{int}(n)$ and $r_{ext}(n)$) were used to remove the effect of real signals from the hybrid adaptive equalizer. The speech signal was then convolved with the equalized filter, and played back over the headphones simultaneously with the real ambient signals playing from the surrounding loudspeakers. An additional pair was constructed for each configuration with equalized filters computed using the hybrid FxNLMS in Case II with no real source present. The main objective here is to evaluate whether the adaptive equalization in the presence of external sounds (i.e., Case III) performs as good as with no external source present (i.e., Case II). Thus, there were total of 5 pairs used in this listening set including one hidden anchor.

4.4.1 Listening test results

The listening test was conducted in a small quiet room with reverberation time of around 80 milliseconds. In all the listening sets, BRIRs were truncated to 50 milliseconds so as to include all the early reflections and most part of the late reverberations of the listening room. Order of the pairs in each listening set was randomized. Listeners were asked three questions for each randomly assigned pair. First question asked to subjects was to identify which of the two sounds are real, .i.e. coming from physical speaker. They were given the option of either choosing one of the two sounds or "both" if they perceive both sounds as natural. Similar subjective rating was used in [153] to do the pairwise comparison of two audio samples. Secondly, they were asked to rate the similarity of the two sounds on a scale of 0-10 from "*completely different*" to "*same*". The main purpose here is to quantify the difference between real and virtual sounds, if any. Finally, they were also specifically asked to rate the proximity of the two sounds in 3D space on a scale of 0-10 from "*very different*" to "*same*". The subjective ratings of the last two questions were decided based on some of the past works on A/B pairwise test to study the perceptual similarity of audio signals [154, 155]. These tests were mainly conducted for evaluation of blind source separation or different audio coding algorithms. There were a total of 22 pairs of audio tracks used in the listening test (9 pairs in SET *1*, 8 pairs in SET *2* and 5 pairs in SET *3*). All the participants in the listening test were given training prior to the actual listening test to learn what is real source and what is virtual source. During the training, they listened to both the real (i.e., sound coming from physical speaker) and virtual (sound coming from headphones emulating the real source) source for different stimuli used in the three listening sets. Listener's head movements were also restricted during the entire duration of the listening test. A total of 18 subjects participated in the listening test comprising of 3 females and 15 male subjects. Two subjects were discarded as the similarity ratings for the hidden anchors with identical stimuli were given score less

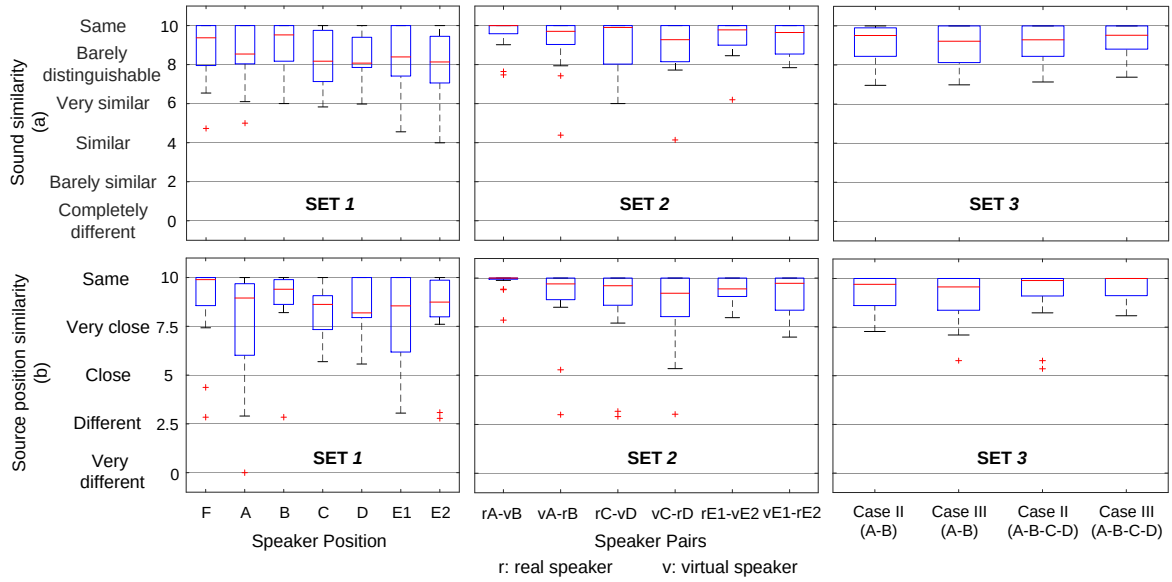


Figure 4.21: Box plot showing subjective scores for sound similarity and source position similarity

than 8. We will now discuss the listening test results based on the three research questions we want to answer in this study.

Naturalness: Naturalness of the virtual source is evaluated based on response of the first question, where subjects were asked to select the real source among the two sounds or both if they perceived both sounds to be natural. Here, the term 'natural' refers to the case when subject is listening to the physical objects existing in the real world. Source confusion (i.e., virtual source is being confused with real source) is used as a measure of naturalness of the virtual source compared to sound coming from speakers. Source confusion can occur in two of the three possible scenarios: (1) subjects chose virtual source as real instead of real source, and (2) subjects perceived both virtual and real sounds as natural and marked "*both*" as the response. Thus, if virtual sound was reproduced very close to real sound, it was expected to have a very high percentage of responses for the "*both*" option. Figure 4.20 shows the source confusion for the three listening sets estimated in percentage as sum of the two scenarios. As shown, for all the three listening sets, in more than 75% of

cases, subjects identified both sounds as real implying virtual sound perceived natural. On the other hand percentage responses for the first scenario was very low, where subjects might have found real source colored and they chose virtual sound as real instead. Overall, very high source confusion of around 90 % is observed for all the listening sets, where subjects marked virtual sound as real. It was also interesting to note that source confusion increases further with increased number of external sources as for SET *2* and SET *3*. One way ANOVA test was conducted to test the significance of reference-test pairs in each set across subjects. It was found that there were no significant variations among loudspeaker positions in SET *1* $[F(6, 105) = 0.68, p = 0.67]$, speaker pairs in SET *2* $[F(5, 90) = 1.27, p = 0.29]$ and different configurations in SET *3* $[F(3, 60) = 0.69, p = 0.56]$.

Sound similarity: Figure 4.21(a) shows boxplots for subjective ratings of the perceptual similarity between real and virtual sounds for all the three sets. Center line in the box represents the median value, while edges of the box are 25 and 75 percentiles responses. Top and bottom lines represent the extreme subject responses, while outliers are shown in red. In SET *1*, most of the subjects found virtual sound highly similar to the real sound source with median of subjective ratings lying in the range 8-10 for all the loudspeaker positions. However, fewer subjects could easily perceive difference between the two sounds. Using the one way ANOVA test, differences in mean scores for different loudspeaker positions across all subjects were found to be insignificant $[F(6, 105) = 0.98, p = 0.44]$. Similar to the source confusion, sound similarity further increased with increased number of sources and even in the presence of ambient sounds. Almost all the subjects rated the two sounds as highly similar with mean subjective ratings of 9.19 and 9.13 for SET *2* and SET *3*, respectively. Different loudspeaker pairs in SET *2* were also found to have insignificant variations in their mean scores $[F(5, 90) = 0.82, p = 0.54]$.

In addition, no significant differences were found between the two adaptive equalization methods used in SET *3* with $[F(1,30) = 0.05, p = 0.83]$ and $[F(1,30) = 0.21, p = 0.65]$ for the 2 and 4 ambient channels, respectively.

Source position similarity: Subjects were asked to compare the position of the two sounds and rate them based on their proximity in 3D space in terms of direction, distance and height. Figure 4.21(b) shows the boxplot for the subjective ratings of source position similarity. Rating of "*very close*" indicates that the two sounds presented are very close to each other in 3D space, while rating of "*very different*" meant one of the sources may be located in completely different position possibly due to front-back confusions or even in-head localizations. This can be observed for SET *1* in Figure 4.21(b) with couple of subjects giving "*different*" score. In general, source position similarities were observed very high with mean rating of 8.26, i.e. the two sounds are perceived very close to each other in 3D space. Similar to the previous two attributes, source position similarity increases further with increased number of sources and mean subjective ratings were found to be 9.1 and 9.3 for SET *2* and SET *3*, respectively, implying close proximity of the two sounds. However, few outliers were also observed with rating of "*different*" for SET *2*. One way ANOVA results for the effect of reference-test pairs in each set across subjects showed that no significant variations were observed among loudspeaker positions in SET *1* $[F(6,105) = 0.89, p = 0.51]$ and speaker pairs in SET *2* $[F(5,90) = 1.46, p = 0.21]$. Furthermore, adaptive equalization for Case II and III have no significant differences in their ratings with $[F(1,30) = 0.4, p = 0.53]$ and $[F(1,30) = 0.85, p = 0.36]$, respectively for 2 and 4 ambient channels, which indicates that proposed adaptive equalizer performs equally well, even in the presence of external sounds.

Table 4.1 summarizes the mean subjective ratings with their 99 % confidence interval for sound similarity and source position similarity. Clearly, virtual sources were

Table 4.1: Mean subjective scores along with their 99 percentile intervals for the three listening sets

Attribute	**SET *1***	**SET *2***	**SET *3***
Sound similarity	8.44 (7.89 - 9.02)	9.19 (8.59 - 9.79)	9.13 (8.61 - 9.65)
Source position similarity	8.26 (7.33-9.19)	9.09 (8.32-9.87)	9.28 (8.77-9.78)

found to be highly similar (barely distinguishable), as well as very close to the real sources in 3D space using the NAR headset. It was also interesting to find correlation among the three sound source attributes studied above. High similarity between the two sounds also meant that they are very close to each other in 3D space and vice-versa for most subjects, but very close proximity in space doesn't always mean they are highly similar as reported by few subjects. In addition, naturalness of the virtual sound does not necessarily imply that the two sounds are highly similar or are very close to each other in 3D space. Nevertheless, high sound similarity and source position similarity indeed resulted in virtual source being identified as real.

4.5 Conclusion

In this chapter, we presented a new approach in reproducing natural listening in augmented reality headsets based on adaptive filtering techniques. The proposed NAR headset structure consists of an open ear cup with pairs of internal and external microphones. Based on the study of different headphones isolation characteristics, it was found that headphones with open design are more suitable for AR related applications, as they allow direct external sound to reach listeners' ear without much attenuation. However, for closed-end headphones or less open headphones, additional processing should be applied to compensate for the headphone isolation using the same sensing units. Based on the amplitude/spectral difference between the two microphone signals, a pair of compensation filters can be applied to make the headsets acoustically transparent. This has been identified as an extension of the

current prototype. For virtual source reproduction via binaural synthesis, individual headphone equalization is applied using adaptive algorithms to compensate for the HPTFs. Modified FxNLMS is proposed with additional spatial filter introduced in the secondary path to improve the convergence rate. However, it is observed that the modified FxNLMS is not able to entirely adapt to the desired response in high frequencies for some of the source positions, whereas conventional FxNLMS suffers from spectral distortion in low frequencies. Hence, we proposed a hybrid FxNLMS to combine the two approaches for optimal performance. Using simulation results, it was found that the hybrid FxNLMS is superior to both approaches with mean square steady state error reduction of more than 25 dB for most of the source positions tested. This implies that virtual sound is reproduced perceptually similar as in direct natural listening. Hybrid FxNLMS is further extended with adaptive estimation of external sounds, as they might interfere with the convergence process. Therefore, with the help of hybrid adaptive equalizer, NAR headset can be individually equalized even in the presence of noisy environments. Listening test was conducted to evaluate perceptual similarities between physical speaker playback and virtual headphone playback. Very high source confusion % was observed, which indicates that virtual source sounds quite natural. Subjects could not differentiate between real and virtual sounds and their positions in 3D space were also in very close vicinity. Moreover, perceptual similarity between real and virtual sounds further increased in an augmented scenario with both real and virtual sources present. In the next chapter, we address some of the practical limitations of the NAR headset and proposed extensions of the adaptive equalization techniques presented in this chapter.

Chapter 5

Practical Limitations, Solutions and Extensions of Natural Augmented Reality Headset

In the previous chapter, we presented NAR headset for augmented reality applications using adaptive filtering techniques. With the help of adaptive headphone equalization techniques, NAR headset can be used to create a sense of natural listening experience in an ARE, where listener can interact with virtual objects while being continuously aware of the real acoustic environment. In this chapter, we address issues related to practical limitations of the NAR headset. Furthermore, we present extensions of the NAR headset for some of the limitations of proposed adaptive equalization techniques presented in the previous chapter.

This chapter is organized as follows: Section 5.1 gives an brief overview of the practical limitations of the NAR headset and the proposed adaptive equalization techniques. Section 5.2 presents fast BRIR acquisition in both static and dynamic scenarios. In Section 5.3, we extend the adaptive equalization technique used in NAR headset for any non-stationary virtual signals. Section 5.4 presents fast detection and estimation HPTF. In Section 5.5, we address the causality issue in online adaptive estimation of external signals and Section 5.6 concludes the chapter with key results from the proposed extensions of the headset.

5.1 Current Practical Limitations Overview

Below is the list of assumptions and limitations of the NAR headset, which are critical for its practical implementations:

L1. It was assumed in subsection 4.3.4 that desired transfer functions ($h^v_{int}(n)$ and $h^v_{ext}(n)$) are measured in the same environment as listeners' external environment ($h_{int}(n)$ and $h_{ext}(n)$) to ensure that virtual sources are perceived similar to real sources. This assumption is particularly important for an ARE, where both virtual and real sources are present and reproduced virtual source must contain the temporal and spatial characteristics similar to that of real source.

L2. Individualized desired transfer functions measurements must be used in virtual sound synthesis to ensure presented augmented auditory environment are well externalized with no front-back confusions. The proposed algorithm in Chapter 4 only compensates for the individual HPTF such that reproduced ear spectrum matches the desired individual sound spectrum.

L3. It has been shown that proposed hybrid adaptive equalizer performs well when adaptive process is trained with stationary broadband white noise source signal $x(n)$. However, there is no such restriction on signals in real life and adaptive equalization should work for any type of source signals.

L4. Accurate estimation of secondary path HPTF model $\widehat{h}_{hp}(n)$ is critical for performance of adaptive equalization in NAR headset, especially for significant changes in HPTF when headset is refitted or even when different listener uses the headset.

L5. The proposed HAE with adaptive estimation of real signals does not guarantee causality for all source directions and may result in incorrect adaptive equalization of NAR headset.

In this chapter, we will address the above five limitations and introduce measures to overcome the practical limitations. In the next section, we address the first two issues (L1 and L2) regarding acquisition of BRIRs using NAR headset.

5.2 BRIRs acquisition using NAR headset

One of the primary challenge of the NAR headset is to create a seamless integration of virtual and real sources in an ARE. For this to happen, virtual source must be reproduced such that it becomes part of the listener's environment. The two most essential attributes required for natural listening in an ARE are:

- Listener's environment characteristics especially, the early reflections and late room reverberations: This information is obtained via measurement of room impulse responses (RIR). It has been widely studied that room reflections and reverberations play an important role in externalization and naturalness of the source.

- Individual related spectral cues due to the head, torso, and most importantly, pinnae: HRIRs contain these spectral cues along with ITD and ILD cues, which are essential for accurate localization and reduction of front-back/up-down reversals.

In other words, desired responses ($h^v_{int}(n)$ and $h^v_{ext}(n)$) used to synthesize virtual source over headphones must emulate the characteristics possessed by the listener environment, as well as preserve the spectral cues, which are highly idiosyncratic. The above two attributes can also be measured together for every individual as BRIRs in the listener environment and subsequently, used in binaural synthesis for adaptive equalization via NAR headset. It should also be noted that virtual sound reproduction becomes more critical in the case of ARE, where virtual source is reproduced alongside the real source. Therefore, there is a natural comparison with

the real source and any deviation from the real auditory environment can be easily detected by the listener, which can be unnatural to the listener. Thus, the main goal here is to obtain BRIRs, which are personalized to each individual, as well as the inclusion of essential listener environment characteristics.

Acoustical measurements of individualized BRIRs or HRIRs are usually carried out using binaural microphones placed at ear canal entrance with loudspeakers positioned at different azimuths and elevations. However, these acoustical measurements are a tedious and time consuming process, which makes it difficult to use in practical scenarios with human subjects. Several other techniques have been presented in literature for customization of HRTFs to individual without measuring the acoustical transfer functions. A few statistical methods based on characterization of HRTF database using principal component analysis (PCA) were presented in [69, 70]. PCA analysis of HRTF database reveals that the individual HRTFs can be sufficiently expressed as linear combination of few orthonormal basis functions. But the parameters for PCA analysis have to be calculated for each source direction and listener. Another method of individualization is based on anthropomorphic geometric model of individual head, but is subjected to errors in head modelling and measurements [71, 72]. Among existing individualization methods, subjective tuning of parameterized or generic HRTFs is the simplest way for individualization with reasonable localization performance [73–75], but it is also a time consuming process.

With the help of our proposed NAR headset and attached binaural microphones, we can readily measure the personalized BRIRs in a given listener environment. However, to overcome the fundamental limitation of conventional method of discrete stop-and-go acoustical measurement, continuous acquisition of HRIRs using adaptive algorithm normalized least mean square (NLMS) [49, 50] is used in this work for rapid acquisition of individualized BRIRs. It can be further extended to recursive least square algorithm (RLS) [156] and other variants of NLMS like propor-

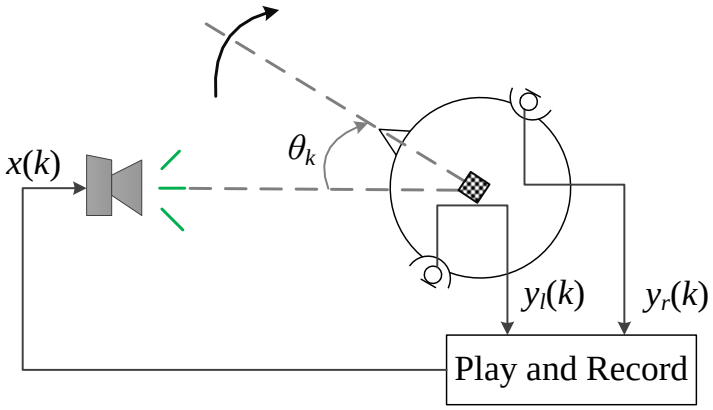


Figure 5.1: Measurement set up for single channel continuous BRIR acquisition with head-tracking

tionate normalized least mean square (PNLMS) [157], variable step-size normalized least mean square algorithm (VSS-NLMS) [158–160] and affine projection algorithm (APA) [161] for better performance but at the cost of increased computational complexity.

5.2.1 Continuous BRIR acquisition using NLMS

Typically, HRIRs are measured in an anechoic chamber on a dummy head or human subject using the binaural microphones placed at eardrum. Excitation signals are played from the loudspeaker and recorded at the ear opening. It is then repeated for multiple loudspeaker positions and/or head orientations. Impulse response is computed using linear deconvolution in frequency domain using spectral division. Depending on the resolution of desired HRIRs in different directions, measurement time of this discrete stop-and-go method is reported in order of hours. Sine sweep signals are most commonly used for estimating the HRIRs in static scenarios and have clear advantage over other excitation signals in terms of noise, time variance and non-linear distortions [162]. Since NAR headset is meant to be used for human subject in real-life scenarios, our goal is to obtain a quick, comprehensive and consistent acoustical measurements via NAR headset for dynamic scenarios. There has

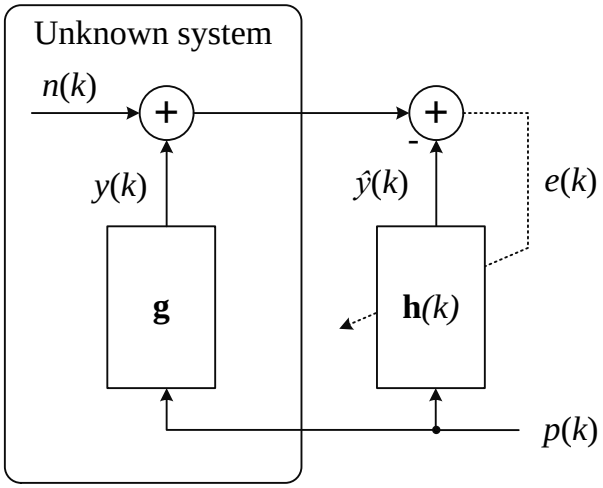


Figure 5.2: System identification using NLMS [6]

been few works in the recent past to achieve dynamic measurements using moving microphones or continuously rotating dummy heads or human subjects [78, 163, 164].

We consider here the continuous BRIR acquisition for fast measurements on human head. Figure 5.1 shows the measurement set up for single channel continuous acquisition of BRIRs. Binaural signals ($y_l(k)$ and $y_r(k)$) at listener's ear canal entrance are captured as the subject continuously rotates in the azimuth plane and fixed loudspeaker continuously plays the excitation signal, $x(k)$. Using method of continuous acquisition, one complete rotation of 360^o azimuth measurements can be completed swiftly. NLMS, which is the normalized version of least mean square algorithm (LMS), is widely used nowadays because of its simplicity and ease of implementation. It is the most commonly used in acoustic echo cancellation [165] and identification of the unknown response of time-varying system proposed by Enzer [49, 50, 163]. Block diagram of the system identification using NLMS is shown in Figure 5.2 and adaptive filter $h(k)$ is updated iteratively as:

$$\mathbf{h}(k+1) = \mathbf{h}(k) + \frac{\mu}{\|\mathbf{p}(k)\|^2}\mathbf{p}(k)e(k)\,, \tag{5.1}$$

with

$$e(k) = y(k) - \widehat{y}(k) + n(k) = (\mathbf{g} - \mathbf{h}(k))^T \mathbf{p}(k) + n(k)\,, \tag{5.2}$$

where $\mathbf{g}$ is the unknown system impulse response which is estimated by $\mathbf{h}(k)$ of length N at any time instant k and $\mathbf{p}(k)$ represents the input vector of length same adaptive filter length. $e(k)$is the residual error signal used to estimate the system response (5.1) with $n(k)$ is the undesired environmental and measurement noise. NLMS uses the normalized step size which is independent of the signal characteristics and should satisfy the criterion $0 < \mu < 2$ for stability. The optimal step size of NLMS algorithm is found to be unity in the case of noiseless condition ($n(k)$=0) [166]. However, in practice, it is usually recommended to choose smaller step-size to ensure the stability in presence of environmental noise. Antweiler [6] studied the stability of NLMS algorithm using the mismatch between system response $\mathbf{g}$ and estimated response $\mathbf{h}(k)$ as stability measure, known as distance vector $\mathbf{d}(k)$:

$$\mathbf{d}(k) = \mathbf{g} - \mathbf{h}(k)\,. \tag{5.3}$$

Under noiseless condition, using (5.2) and (5.3), weight update equation can be rewritten as:

$$\mathbf{d}(k+1) = \mathbf{d}(k) - \mu \frac{(\mathbf{d}(k))^T \mathbf{p}(k)}{\|\mathbf{p}(k)\|^2} \mathbf{p}(k)\,. \tag{5.4}$$

The second term in (5.4) can be interpreted as orthogonal projection of distance vector $\mathbf{d}(k)$ on to the input signal $\mathbf{p}(k)$. It further implies that, distance vector can be completely eliminated if N successive vectors of input sequence, i.e. $\mathbf{p}(k), \mathbf{p}(k-1), ..., \mathbf{p}(k-N+1)$ are orthogonal in N-dimension vector space [6, 167]. In other words, if these N input vectors are independent over time, distance vector will be independent of input vector and adaptive process will optimally converge after N iterations in a noiseless environment with $\mu = 1$. Therefore, choice of input excitation signal is also crucial for the convergence and optimal performance of the

NLMS algorithm. White noise sequence are usually used in acoustic echo cancellation but due to its finite length it does not guarantee orthogonality and results in non-optimal convergence speed. Perfect sequences (PSEQs) are another class of excitation signals, which are periodic repeated signals, such that their auto-correlation function becomes zero for its N shifted sequences, i.e., PSEQs with period N are orthogonal. Thus, PSEQ satisfies the requirement of optimal excitation of NLMS. Telle et al. [168] proposed a new type of PSEQ, known as perfect sweep signal. They are preferred for acoustical measurements because of the distortion free measurements even at higher amplitudes as compared to other perfect noise sequences. Perfect sweep signal is a linear sweep signal with perfectly constant amplitude spectrum such that it satisfies the orthogonality requirement of NLMS. The perfect sweep signal is continuously played over loudspeaker and its period must match the adaptive filter length N. It is constructed in the frequency domain by keeping the spectral amplitude constant and designing a linearly increasing group delay, before taking its inverse Fourier transform. One important characteristic of the perfect sweep signal is that for periodically repeated sweep signal, highest most frequency fold back to lower most frequency, which implies continuous transition between two periods, as shown in Figure 5.3.

With the assumption of linear and broadband transducers, sound propagation between loudspeaker are continuously rotating subject can be described as a slow time-varying system with impulse response, $h_{l/r}(\theta_k)$ [50] for azimuth θ_k at time instant k. Using NLMS, the weight update equation for estimating $\widehat{h}_{l/r}(\theta_k)$ is expressed as:

$$\widehat{\mathbf{h}}_{l/r}(\theta_{k+1}) = \widehat{\mathbf{h}}_{l/r}(\theta_k) + \frac{\mu}{\|\mathbf{p}(k)\|^2}\mathbf{p}(k)e_{l/r}(k)\,, \tag{5.5}$$

where $e_{l/r}(k)$ is residual error signals for left and right ears defined as:

$$e_{l/r}(k) = y_{l/r}(k) - \mathbf{p}^T(k)\widehat{\mathbf{h}}_{l/r}(\theta_k)\,, \tag{5.6}$$

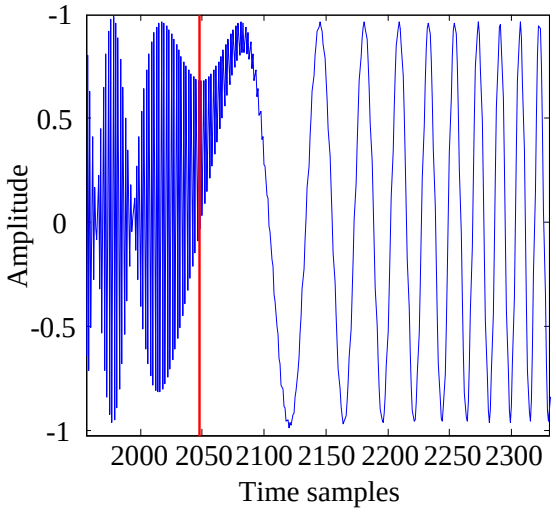


Figure 5.3: Perfect sweep signal showing continuous transition from 1 period to another. $N = 2048$ samples

and binaural signals captured by the two microphones are denoted as:

$$y_{l/r}(k) = \mathbf{p}^T(k)\mathbf{h}_{l/r}(\theta_k) + n_{l/r}(k). \tag{5.7}$$

The unknown state for estimated impulse response at time $k = 0$ is taken as vector of N zeros. Thus, for a 360^o rotation time of 20 sec and at sampling frequency of 44.1 kHz, azimuth resolution of around 10^{-4} degrees can be obtained using the above method, which implies that estimated impulse response can be considered a very good approximation of continuous HRIRs. Due to the large data to be stored in memory, care must be taken to selectively choose the desired data. It should also be noted that in estimating continuous HRIRs using (5.5), (5.6) and (5.7), it was assumed that system impulse response changes slowly with time, such that adaptive filter can converge to the desired response. In the next sub-section, we show the results for single channel BRIR acquisition for static scenario, i.e., without head rotation and follows by the dynamic scenario with head rotation.

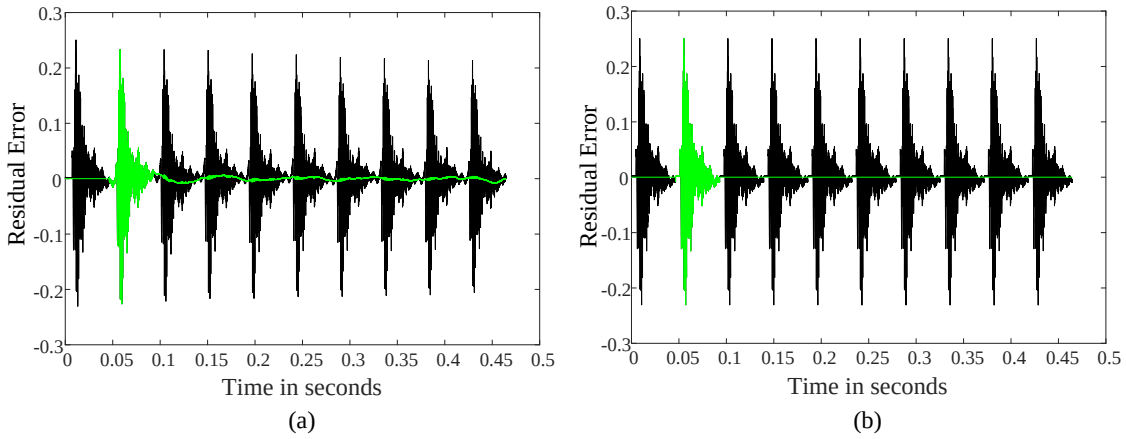


Figure 5.4: Residual errors (in green) for BRIR acquisition using NLMS for (a) Noisy captured signal (b) Noiseless signal with 10 repetitions of actual recorded signal

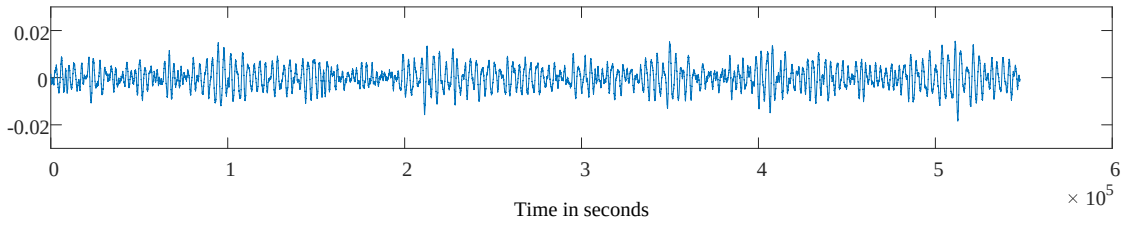


Figure 5.5: Environment noise of the measurement room recorded at subject's ear

5.2.2 Continuous BRIRs acquisition results - *Without Head Rotation*

Continuous BRIRs measurements using perfect sweep were carried out in same room used for subjective study for NAR headset with reverberation time of around 80 milliseconds. Length of the NLMS adaptive filters were set at 2048 samples (50 msec) assuming that it is sufficiently long to include all room reverberation and torso, head, ear reflections. Thus, a perfect sweep signal of period of 2048 samples, as shown in Figure 5.3, is repeatedly played 10 times from the loudspeaker located at 0^0 azimuth. First, we validate the static scenario, where head movements were restricted. Binaural signals were captured at subject's ears for fixed head position using the binaural microphones attached with the NAR headset.

We compute the residual error for two cases namely, (a) with NLMS applied on

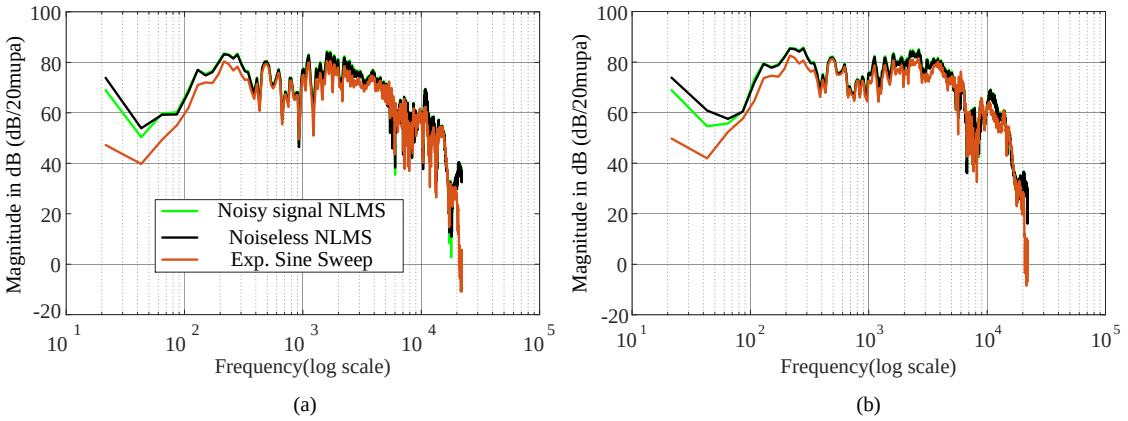


Figure 5.6: Magnitude frequency responses of estimated BRIRs compared with exponential sine sweep method (a) Left ear (b) Right ear

noisy recorded signal for 10 periods of perfect sweep signal, and (b) with NLMS applied on 10 repetitions of one period of noisy recorded signal, considered as noiseless case. Residual errors for the two cases are shown in Figure 5.4 (a) and (b). For the noisy case, it is observed that although adaptive filter converges after N iteration (plus N iterations for initialization) but residual error is not completely eliminated with gradient noise in steady state. This is due to the environment noise present in the measurement room probably generating from air-con vent, analog amplifiers and computer fan. The environment noise, which is measured at subject's ear, is shown in Figure 5.5. Clearly, noise floor is noticeable as compared to the magnitude of binaural recorded response of sweep signal with signal to noise ratio (SNR) of 28 dB and for that reason NLMS algorithm is not able to optimally converge, as shown in Figure 5.4(a). Noiseless case, shown in Figure 5.4(b), optimally converges after $2N$ iterations. Frequency response of estimated BRIRs for both the cases is shown in Figure 5.6 along with the binaural room transfer function (BRTF) computed using exponential sine sweep method as reference. Clearly, both the estimated BRTFs closely match with the reference response above 100 Hz, while in low frequencies, a boost of 10-20 dB is observed for the NLMS estimated response. In addition, noiseless case shows a boost of 2-3 dB in low frequency below 70 Hz as compared to noisy NLMS case. This low frequency boost may be due to the fact that noiseless esti-

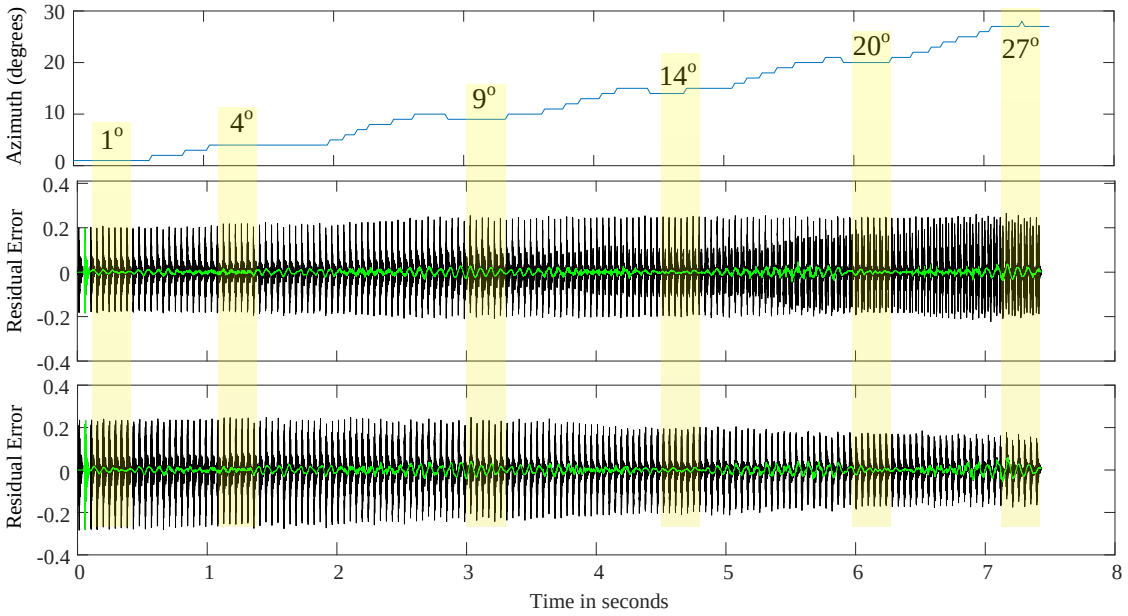


Figure 5.7: Continuous BRIR estimation results with head rotation in clockwise direction (Top: Head orientation in clockwise direction; Middle: Residual error for left ear; Bottom: Residual error for right ear)

mation is based on single shot measurement, i.e., one period of perfect sweep, while noisy case adapts with time for the varying system impulse response. Nevertheless, BRIRs estimated using NLMS method can be considered as a good approximation of the unknown system response. As shown in the results, measurement for static source position can be performed promptly in less than half a second as compared to discrete stop-and-go exponential sine sweep method, which also require repetitions.

5.2.3 Continuous BRIRs acquisition results - *With Head Rotation*

For continuous BRIR acquisition with head rotation, a head tracker is mounted on the NAR headset to track the head orientation. YEI 3-Space Sensor™ Micro USB is used as head tracker in the measurement process. One main advantage of using head tracker with NAR headset is that subject can measure the BRIRs for different directions themselves conveniently in a room environment, without any external aid. However, one important assumption of continuous acquisition using

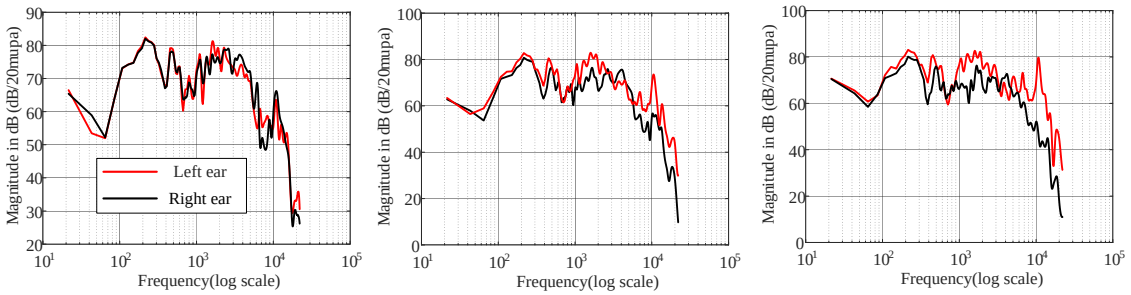


Figure 5.8: Estimated BRTFs with continuous head rotation for 3 head azimuths (Left: 0^0; Middle: 14^0 ; Right: 27^0)

NLMS that subject should rotate slowly and at constant angular speed for adaptive filter to adjust to the changes, can only be met using machine controlled rotation. When asking subjects to rotate their head in either direction, it is imperative that the rotation speed may not be constant and slow enough for NLMS algorithm to optimally converge. In our measurement process, subjects were asked to rotate slowly in either direction and stop regularly for a moment during the measurement process. In this way, we can at least expect to obtain a reliable BRIRs for those head orientations where subject stopped. Figure 5.7 shows a sample head orientation data recorded for a subject along with the residual errors obtained using NLMS adaptive algorithm. As shown in top of Figure 5.7, subject stopped at least 6 times during the entire measurement duration of 7.5 seconds. Since at these 6 positions, subjects can be assumed to be static, we can estimate BRIRs similar to the static case by taking 5-6 periods of the recorded signal (shown as yellow rectangular window in Figure 5.7). Figure 5.8 shows the estimated frequency response for 3 of the static head positions. For other intermediate head orientations, system response may not be assumed as slowly time-varying and thus, is excluded in the estimation process. However, it needs to be studied further subjectively if the intermediate BRIRs can actually be used in binaural synthesis for dynamic head movements.

5.3 Adaptive Equalization for Non-stationary Virtual Signals

In Chapter 4, we presented adaptive equalization techniques based on the FxNLMS algorithm to compensate for any change in individual HPTF over time such that natural listening can be enabled over the NAR headset. We showed that broadband stationary white noise signal performs well with the adaptive equalization of NAR headset for training of equalization filters ($w_1(n)$ and $w_2(n)$). However, in real-life virtual signals are mainly non-stationary and transient in nature, and adaptive equalization must work for all kind of signals. One of the main problems with FxNLMS is that it may become unstable due to the transients in virtual signals, which is mainly due to the normalization term used in the weight-update equation (see (4.15) and (4.16)). Alternatively, we can use FxLMS algorithm without any normalization but with slower convergence rate and higher SS-MSE, as compared to the FxNLMS algorithm.

We extend the proposed adaptive equalizer in Chapter 4 by including an online adaptive training phase using white noise signal and a playback phase of virtual signals using FxLMS, as shown in Figure 5.9. The training adaptive phase is mainly to ensure that adaptive filter is converged before starting the real-time playback of virtual signals. As shown, the training adaptive phase is same as the HAE presented for Case II in subsubsection 4.3.4.2, except for secondary path response is replaced by an estimate of it. It should be noted that training is excluded from the playback and ensures optimal convergence and MSE for virtual signals. The playback phase of virtual signal uses a copy of equalized filters from training phase and an additional adaptive filter, $w_v(n)$ for compensation of individual HPTF, as shown in Figure 5.9. $w_v(n)$ is updated based on the FxLMS algorithm as:

$$\mathbf{w}_v(n+1) = \mathbf{w}_v(n) + \mu_v \, \mathbf{v}'(n) \, e_v(n) \ , \tag{5.8}$$

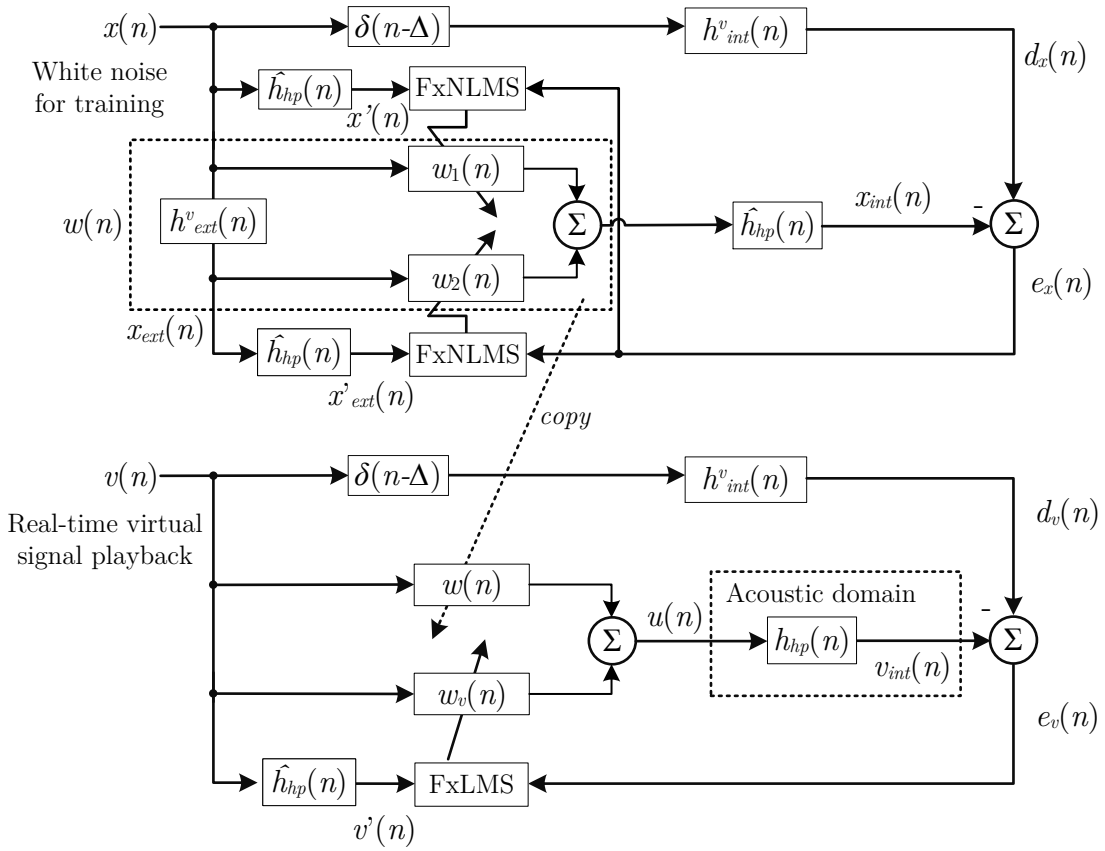


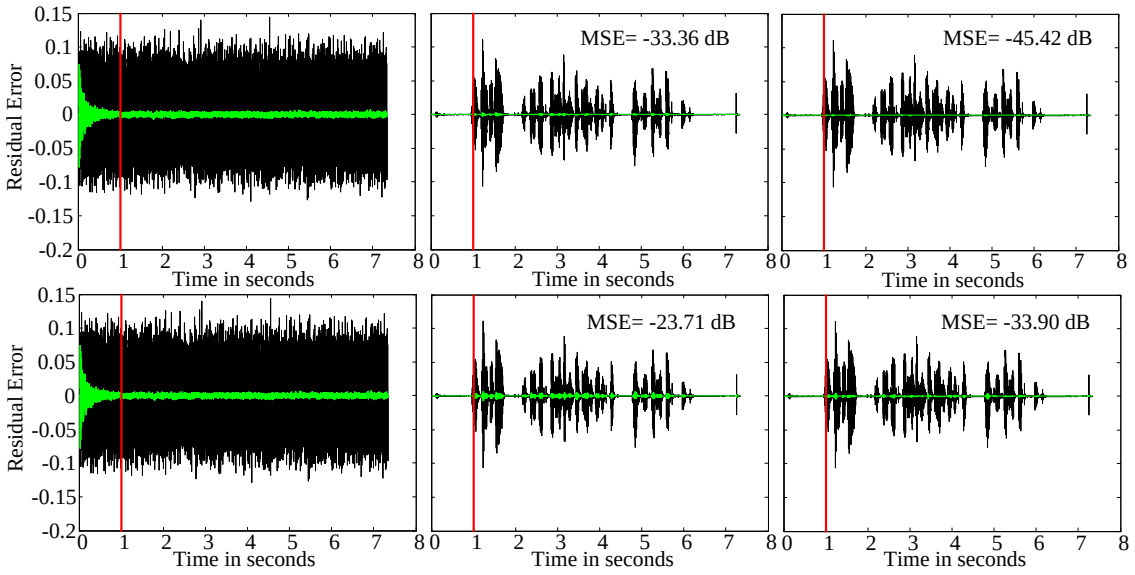
Figure 5.9: Proposed adaptive equalizer for NAR headset extended for non-stationary virtual signals

where $\mathbf{v}'(n)$ is the filtered reference signal, μ_v is the step-size different from training phase and $e_v(n)$ is the residual error signal during the playback phase. Step-size μ_v for FxLMS is decided based on the power of filtered reference signal as well as length of adaptive filter, L and secondary path delay, $\triangle$ [169]:

$$0 < \mu_v < \frac{2}{E[x'^2(n)](L + \triangle)}\,. \tag{5.9}$$

However, if the input signal is non-stationary, it was suggested by Elliott [170] that the maximum value of step size is proportional to the $\frac{1}{1.2L}$ instead of $\frac{2}{L}$. It was explained that this is mainly due to the poor conditioning of the co-variance matrix for filtered reference signal.

Results for the proposed adaptive equalization with training period of 1 sec and



(a) Training with 1 second white noise

(b) Playback with no adaptive compensation (w_v=0)

(c) Playback with adaptive compensation ($w_v \neq 0$)

Figure 5.10: Results for the proposed adaptive equalizer with training period of 1 second using white noise signal (Top: $h_{hp}(n) = \widehat{h}_{hp}(n)$; Bottom: $h_{hp}(n) \neq \widehat{h}_{hp}(n)$)

playback of a speech signal is shown in Figure 5.10.

Step size μ and μ_v for training and playback phases are set as as 0.1 and 0.4, respectively. During the training phase, both both the adaptive process is on to ensure fast convergence and optimum MSE. The training using white noise is stopped after 1 second once the adaptive filter $w(n)$ is sufficiently converged and stabilized. For the ideal case, where both secondary path and its estimate are taken same shows an improvement of approximately 12 dB error reduction due to the playback phase adaptive filter $w_v(n)$ (See top Figure 5.10(b) and (c)). For a practical case, when physical secondary path is different from its estimate due to re-positioning of headset, we observe that performance of adaptive filter degrades by 10 dB when no adaptive compensation is applied, i.e. $w_v = 0$ (See Figure 5.10(b)). After applying the adaptive compensation, an improvement of 10 dB in MSE is observed, as shown in the bottom Figure 5.10(b) and (c). The main advantage of adaptive compensation in playback phase is that it adapts according to the virtual signal characteristics, as well as robust to the error between $h_{hp}(n)$ and $\widehat{h}_{hp}(n)$. However, due to the

large difference in secondary path and its model, adaptive filter $w_v(n)$ is not able to completely compensate for the new HPTF, as increase of 11 dB in MSE is observed when HPTF changes abruptly after 1 second (See Figure 5.10(c)). In the next chapter, we present a way to detect large changes in secondary path response and propose a method to compensate for it using a fast HPTF estimation method.

5.4 Detection and Fast Estimation of Headphone Transfer Function in Natural Augmented Reality Headset

As concluded in the preceding Section 5.3, the FxLMS algorithm in adaptive equalization is robust for changes in the HPTF, although it may not be able to completely compensate for large changes in HPTF. This is mainly due to the fact that both adaptive headphone compensation (i.e. playback phase) and training phase uses an estimate of HPTF, which may deviate largely from its physical model in the event of any large changes. Therefore, it is essential to detect any large changes in HPTF (either due to re-positioning of headset or even change of the listener) and quickly find an accurate estimate of it to correct the individual headphone compensation. First, we show online detection of change in physical secondary path, as shown in Figure 5.11. The virtually synthesized secondary source signal $u(n)$ is filtered with with the HPTF estimate $\widehat{h}_{hp}(n)$ and subtracted with the signal received at error microphone:

$$e_{hp}(n) = v_{int}(n) - \widehat{v}_{int}(n) = u(n) * \left(h_{hp}(n) - \widehat{h}_{hp}(n)\right). \tag{5.10}$$

As can be seen in (5.10), we can estimate the difference in actual secondary path and its current estimate using the residual error signal between $v_{int}(n)$ and $\widehat{v}_{int}(n)$. It should be noted that error signal $e_{hp}(n)$ also depends on the characteristics of

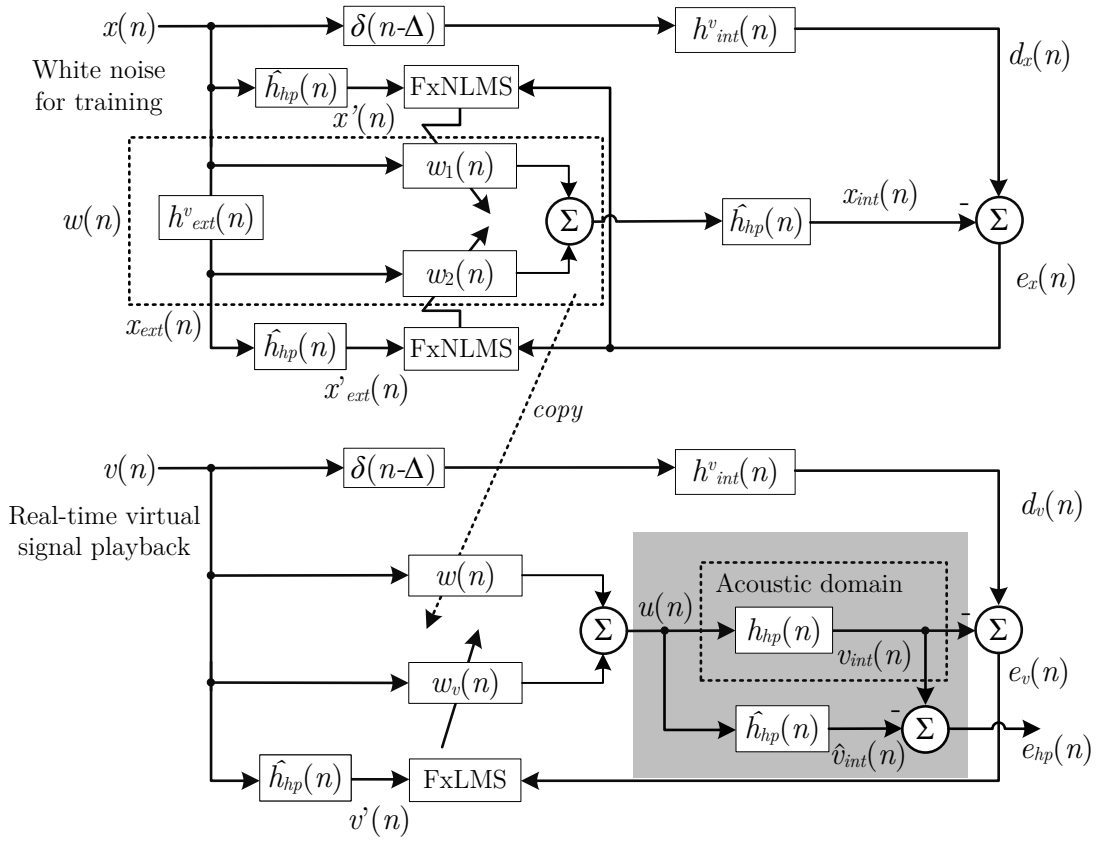


Figure 5.11: Modified block diagram of adaptive equalizer with online detection of change in HPTF (highlighted in grey box)

virtual source signal $u(n)$. We define the running average power estimate (PE) in dB of the residual error signal $e_{hp}(n)$ as:

$$PE(n)\,[dB] = \frac{1}{W}\sum_{l=1}^{W} 10\log_{10}\left[\frac{e_{hp}^2(n-W+l)}{v_{int}^2(n-W+l)}\right], \tag{5.11}$$

where W is the window size over which average power of the error signal is computed at time instant n. Ideally, if both the physical model and its estimate are same or in close agreement with each other (i.e., $h_{hp}(n) \approx \widehat{h}_{hp}(n)$) , power estimate of the error signal will be very low. On the other hand, if the two deviates from each other significantly, power estimate increases and change in HPTF can be detected as soon as $PE(n)$ crosses a threshold.

Result of the HPTF detection for the two measured model of HPTFs ($hp2$ and $hp3$) with respect to a given HPTF model ($hp1$) is shown in Figure 5.12. Headphone

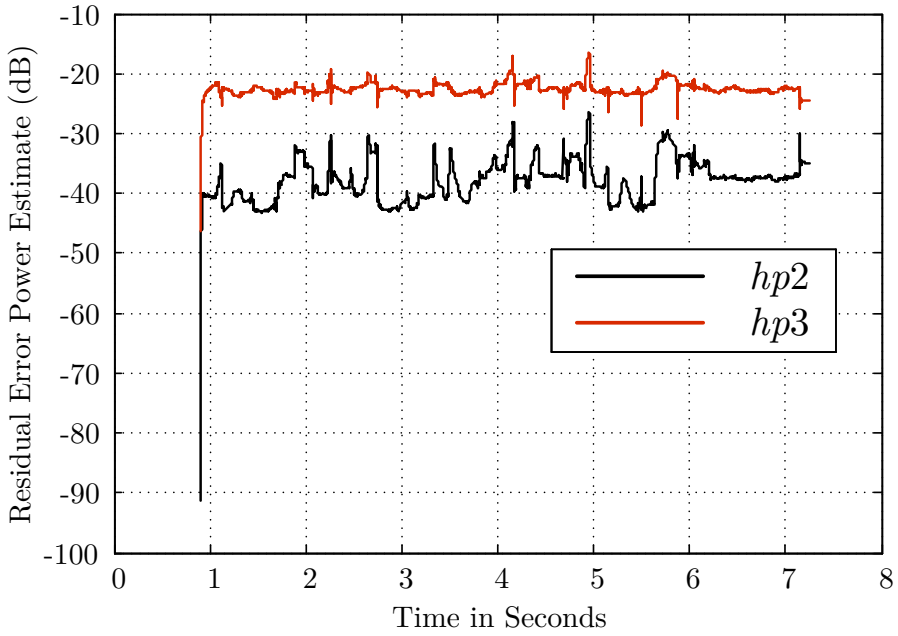


Figure 5.12: Results of HPTF detection for two measured physical model of HPTFs using (5.11). (Window size of $W = 4096$ samples i.e., around 100 msec is used)

model $hp2$ was measured by slightly adjusting the NAR headset without lifting it completely, while for model $hp3$ headset was completely lifted and placed back on the dummy head. Thus, large deviation is expected for transition $hp1$ to $hp3$ as against transition to $hp2$. For the results shown, during the first second, HPTF model $\widehat{h}_{hp}(n)$, which is equal to the physical model ($h_{hp}(n) = hp1$), is adaptively compensated simultaneously with the white noise training process and very low value of the power estimate is observed. After one second, training is stopped and $h_{hp}(n)$ is replaced by physical headphone model $hp2$ or $hp3$. Clearly, after one second, power estimate of the error signal increases with average PE of around -23 dB and -38 dB for $hp3$ and $hp2$, respectively as shown in Figure 5.12. We set the threshold for change in HPTF detection as -30 dB. With power estimate of $e_{hp}(n)$ for $hp3$ around 15 dB more than that of $hp2$ and substantially greater than the threshold, large change in HPTF is detected for $hp3$ as soon as the training phase ends after one second.

Once the change in HPTFis detected, we must update the estimate of secondary

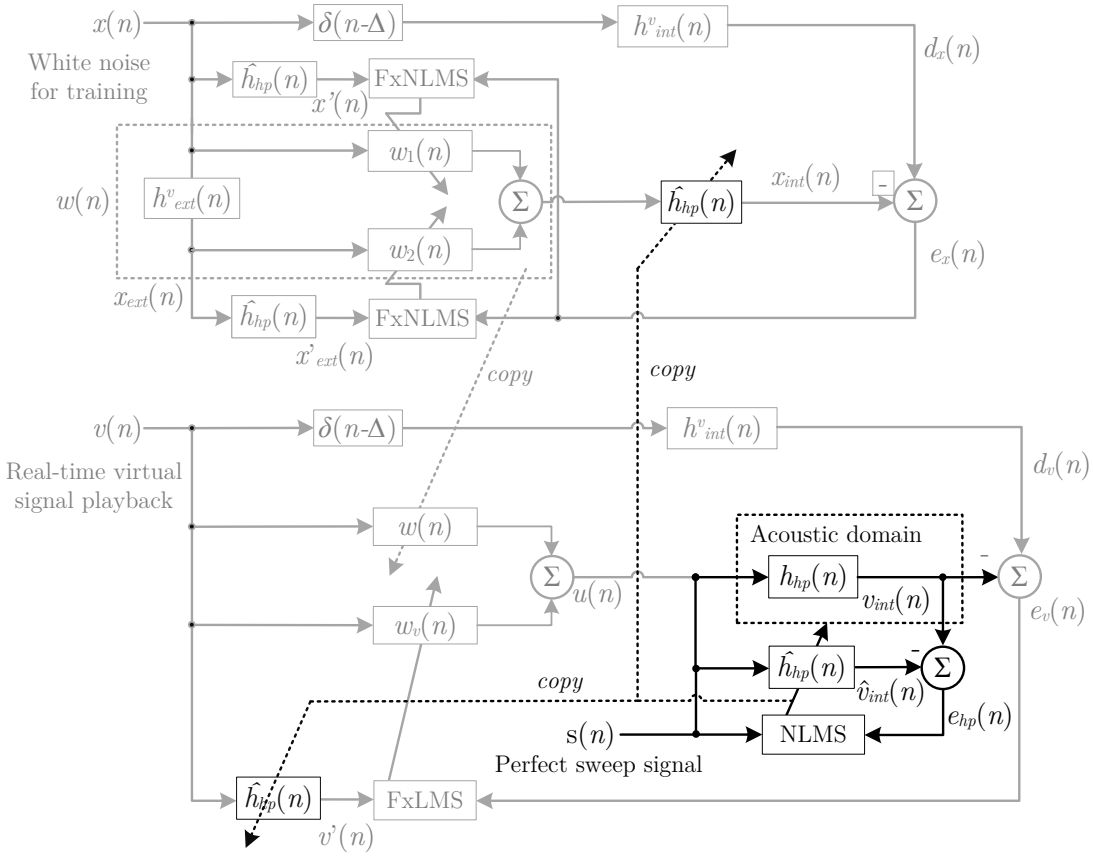


Figure 5.13: Modified block diagram of adaptive equalization of NAR headset with HPTF estimation. Training and playback phase is shown in grey colors indicating that virtual signal playback is stopped, while HPTF estimation is going on.

path with the new estimate of changed secondary path. There have been some works [171–174] in the past for online estimation of secondary path model for FxLMS algorithm by injecting a random noise to the headphones in addition to the virtual signals and computing the secondary path estimate using LMS algorithm. However, these methods were mainly applied in case of active noise control applications, where primary source of interference is also a noise signal. In contrast, NAR headset aims to adaptively compensate for any large change in HPTF and for any type of virtual source signals. In addition, playing noise signals for online adaptive estimation of HPTF may not be practical for NAR headset as it will be interfere the real-time playback of virtual sounds. Furthermore, online estimation method also slows down

the convergence process, which is very critical for the optimal performance of NAR headset. Therefore, we need a quick way of estimating HPTFs and updation of secondary path models in the adaptive equalization of the NAR headset. It was shown in Section 5.2 that unknown system response can be very quickly measured and estimated with the help of the NLMS algorithm using perfect sweep signal. In addition, NLMS using perfect sweep signal is converged just after N iterations in a noiseless case, where N is the adaptive filter length. We can thus, measure and estimate the HPTF model using NLMS in a short duration because of the following reasons:

1. Secondary path response, i.e., length of $h_{hp}(n)$ is very short due to the fact that headphone emitter and internal microphone (m_{int}) are closely placed to each other. (See Figure 4.3).

2. Because of the closed structure of NAR headset, there will be fewer reflections in the headphone impulse response.

Block diagram for the estimation of headphone response is shown in Figure 5.13. As shown, during the estimation process, the playback of virtual signal is stopped and restarted as soon as the estimation is over. Taking $N = 256$ samples as the headphone impulse response length and a perfect sweep signal comprising of 4 repetitions, we can easily estimate $\widehat{h}_{hp}(n)$ in less than 50 milliseconds at sampling frequency of 44.1 kHz ($N \times 4 = 1024$ samples). Results for the modified adaptive equalizer with HPTF estimation is shown in Figure 5.14. Headphone model $hp3$ is used in the simulations as large change in HPTF was detected using power estimate of residual error signal $e_{hp}(n)$. We consider three cases for comparisons here namely, (a) Adaptive equalization without any HPTF estimation, (b) Reference case, i.e., it is assumed that headphone model of the new secondary path is already available once change in HPTF is detected and thus, there is no need of any HPTF estimation, and (c) Adaptive equalization with HPTF estimation using NLMS after HPTF

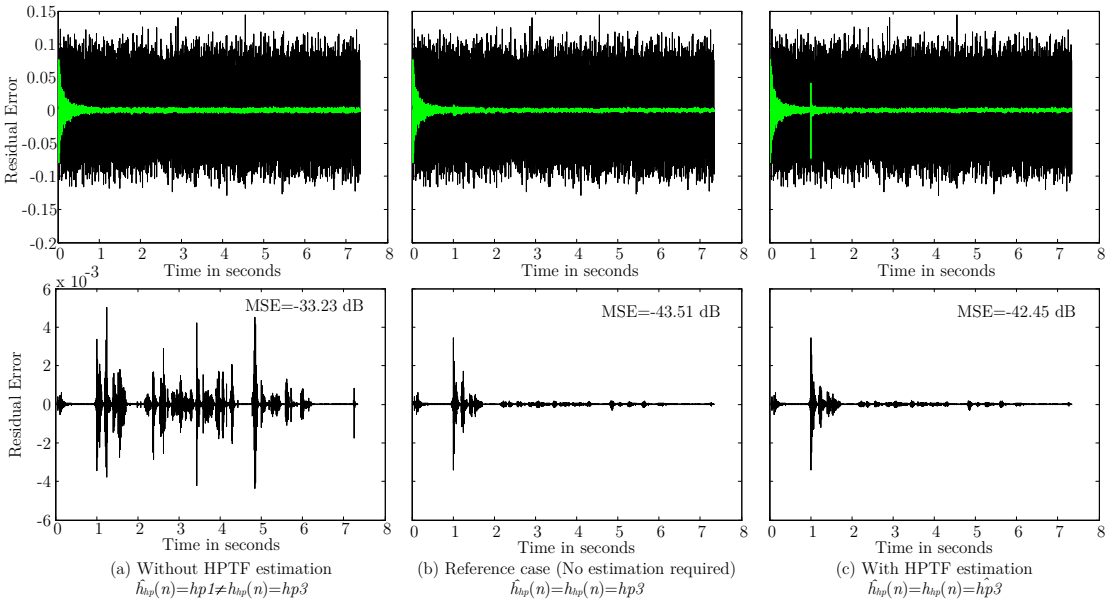


Figure 5.14: Results for the adaptive equalization of NAR headset with HPTF estimation and compared with the reference case as well as adaptive equalization without HPTF estimation (Top: Residual error plots for training phase white noise ; Bottom: Residual error plots for a virtual speech signal) Simulation is performed for azimuth: 40^o and ipsilateral ear.

detection with secondary path estimate is replaced by new HPTF estimate. Below are the different simulation settings used in Figure 5.14 at different time instants:

- **$0 < t < 1$ second:** Training phase using white noise signals with $\widehat{h}_{hp}(n) = h_{hp}(n) = hp1$.
- **1 second:** Physical secondary path headphone model is changed to $hp3$, i.e., $\widehat{h}_{hp}(n) = hp1 \neq h_{hp}(n) = hp3$.
- **$1 < t < 2$ second:**
  - Case (a): White noise training is stopped and playback of virtual continues with incorrect estimate of secondary path without HPTF estimation for the changed secondary path response $h_{hp}(n) = hp3$
  - Case (b): White noise training is continued with secondary path estimate is replaced by the exact changed physical model i.e., $\widehat{h}_{hp}(n) = h_{hp}(n) =$

$hp3$

– Case (c): White noise training as well as playback of virtual signals is stopped. HPTF estimation process is started and run for 1024 samples. Since, NLMS converges after N iterations, previous secondary path estimates are replaced by new estimate of the changed physical model, i.e., $\widehat{h}_{hp}(n) = h_{hp}(n) = \hat{hp}3$ and training phase is restarted. Once the estimation process is completed after 1024 samples, playback of virtual signal begin.

- **t > 2 second:** White noise training is stopped and playback of virtual signal begin.

Clearly, when no HPTF estimation is applied adaptive equalization suffers from higher MSE because of the incorrect estimate of $h_{hp}(n)$. When HPTF estimation is applied and used in the playback of virtual signals as well as training phase, MSE of the residual error is found be very close to the reference case with difference of around 1-2 dB, implying that secondary path has been accurately estimated. During the 2^{nd} training phase, as adaptive filter compensates for new $\widehat{h}_{hp}(n)$, MSE starts with high value and reduces to its optimum value for both the reference case, as well as adaptive equalization with HPTF estimation (See Figure 5.14(b) and (c)). For the third case, since HPTF estimation takes around 1024 samples to complete, a peak is observed in the residual error around $t = 1$ second during the this period, but quickly converges to the optimum value once an accurate estimate of $h_{hp}(n)$ is found, as shown in top Figure 5.14(c). However, this is not the case for reference case as there is no HPTF estimation involved, and training process continues to compensate for the change in HPTF. In summary, any large change in HPTF can be immediately detected using power estimate method and subsequently, can be used to trigger the HPTF estimation process, which can also be completed in a very quick time. In the next section, we further extend the above proposed adaptive

equalization for any virtual signals with adaptive estimation of external signals and also address the causality issue.

5.5 Adaptive Equalization for Non-stationary Virtual signals with Adaptive Estimation of External signals

Similar to the previous section, we further extend the adaptive equalizer for augmented reality mode presented for Case III in subsection 4.3.5 for any type of virtual signal, as shown in Figure 5.15. As compared to Figure 5.9, there is difference only in the playback phase with inclusion of adaptive estimation process of external signals. The main purpose of adaptive estimation here is to remove the interference of external signals in the adaptive compensation of playback phase. Therefore, it is very important for adaptive estimation of external signals during real-time playback of virtual signal in order to ensure its stability and convergence. However, the external sounds coming from different directions may not guarantee optimum convergence because of the causality issue between the reference signal ($r_{ext}(n)$) and error microphone signal ($r_{int}(n)$). Since in NAR headset, both the microphones are closely placed, the adaptive filter $w_r(n)$ must be of very short length to avoid stability of adaptive process. Also, in practice, real signals captured at external microphone may reach later than at internal microphone causing causality problem for the adaptive estimation process. Causality can be avoided by adding a forward delay ($\triangle_r$) to the incoming real signal at internal microphone position in the primary path of adaptive estimation, as shown in Figure 5.15. Weight update equation for the adaptive filter $w_v(n)$ remains the same as in (5.8), except for the error signal, $e(n)$ obtained due to the error difference for both virtual and real signal:

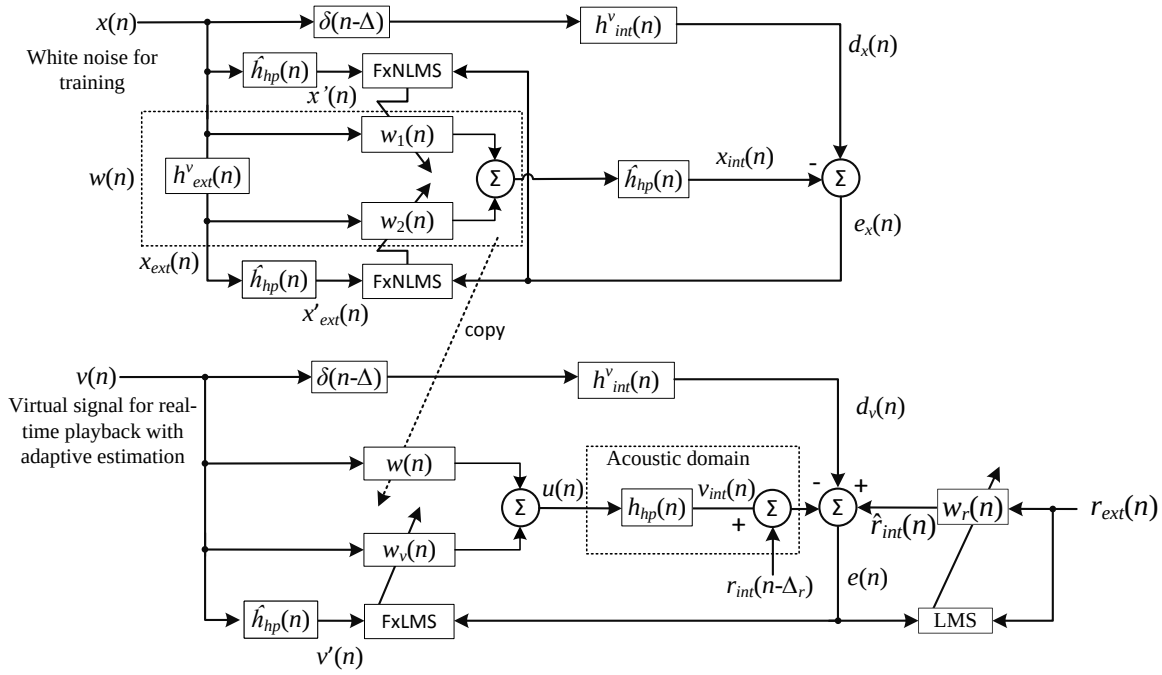
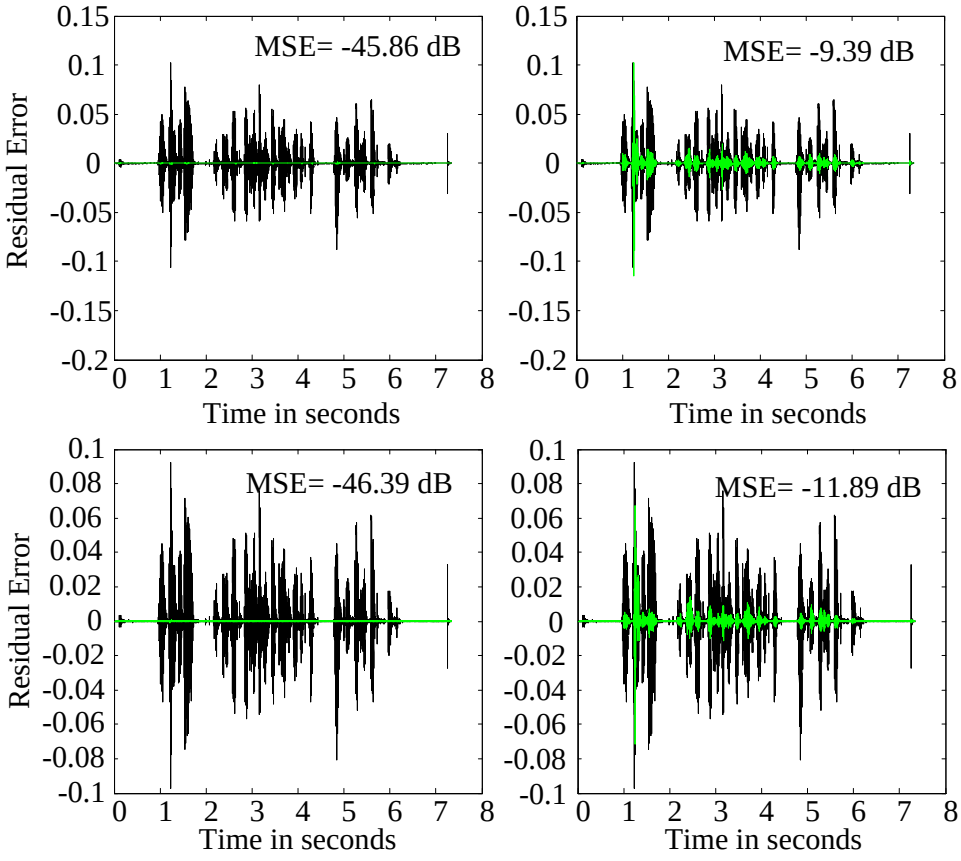


Figure 5.15: Proposed adaptive equalizer with adaptive estimation of real signals for NAR headset extended for non-stationary virtual signals

$$
\begin{aligned}
e(n) &= d_v(n) - y_{int}(n) + \widehat{r}_{int}(n) \\
&= [d_v(n) - v_{int}(n)] + [-r_{int}(n - \triangle_r) + \widehat{r}_{int}(n)] \\
&= e_v(n) + e_r(n)\,.
\end{aligned} \tag{5.12}
$$

Weight update equation for the adaptive estimation filter $w_r(n)$ is computed similar to (4.30), but with error signal defined as (5.12).

Figure 5.16 shows the residual error plots for adaptive equalizer in the presence of external signals with no adaptive estimation compared with the case when there is no external signal present. Speech signal is used for the virtual signal, while Gaussian white noise signal is used for the simulated external signal. Clearly in the presence of external signal, performance of adaptive equalizer degrades substantially when no adaptive estimation is applied as real signal interferes with the adaptive compensation of HPTF. Similar results were obtained in subsection 4.3.5, when white noise was used as virtual signal for the adaptive headphone compensation. Figure 5.17 shows the results for the proposed adaptive equalizer with adaptive



(a) Adaptive Equalization when no external source present

(b) Adaptive Equalization when external source present

Figure 5.16: Results for the proposed adaptive equalizer without and with presence of external signals: Virtual source is positioned at 0^o azimuth, while external sound is coming from 40^o azimuth and added to virtually reproduced signal at m_{int}
(Top: Ipsilateral ear ; Bottom: Contralateral ear)

estimation of external signals for two delay values: $\triangle_r = 0$ and $\triangle_r = 40$ samples. As shown, when there is no delay applied to input signals, adaptive filter $w_r(n)$ is not able to completely converge for the contralateral ear, resulting in higher MSE. This is due to the fact that for contralateral ear, external signal reaches error microphone position earlier than the external microphone position for 40^o azimuth, while it is opposite for the ipsilateral ear resulting in much higher error reduction. On the other hand when the external signal reaching at m_{int} is delayed by 40 samples, causality issue is resolved and MSE for both the ears are in close agreement with the case

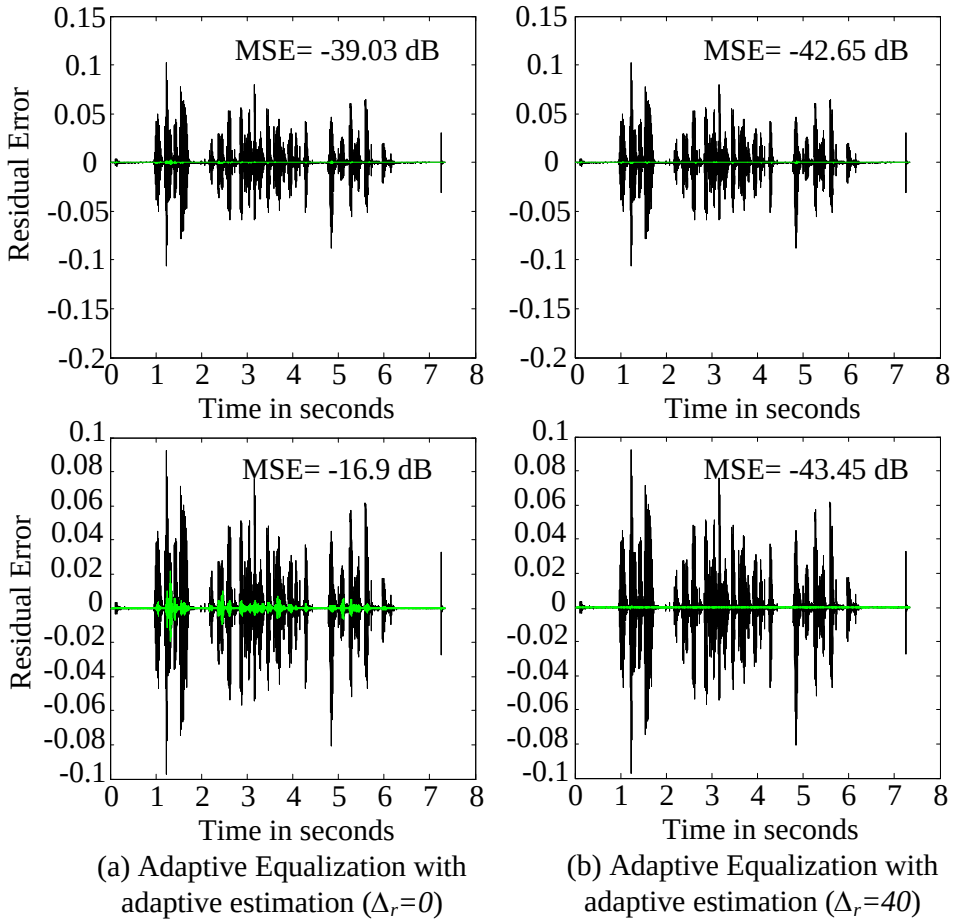


Figure 5.17: Results for the proposed adaptive equalizer with adaptive estimation of external signals. Simulation set up is kept same as of Figure 5.16 and step-size (μ_r) for adaptive estimation is taken as 0.4.
(Top: Ipsilateral ear ; Bottom: Contralateral ear)

when there is no external source present (See Figure 5.16(a) and Figure 5.17(b)). Therefore, it is necessary to avoid the causality issue to ensure stability of the proposed adaptive equalization. However, it is not practically feasible to delay only the real signal $r_{int}(n)$ shown in Figure 5.15, as the error microphone also captures the synthesized virtual signal $v_{int}(n)$.

This issue is resolved by forwarding the delay $\triangle_r$ to the error path as well as weight update path of the adaptive compensation filter $w_v(n)$, as shown in Figure 5.18. Hence, this situation is akin to the delayed LMS (DLMS) algorithm with delayed coefficient update mechanism to account for the inherent computational

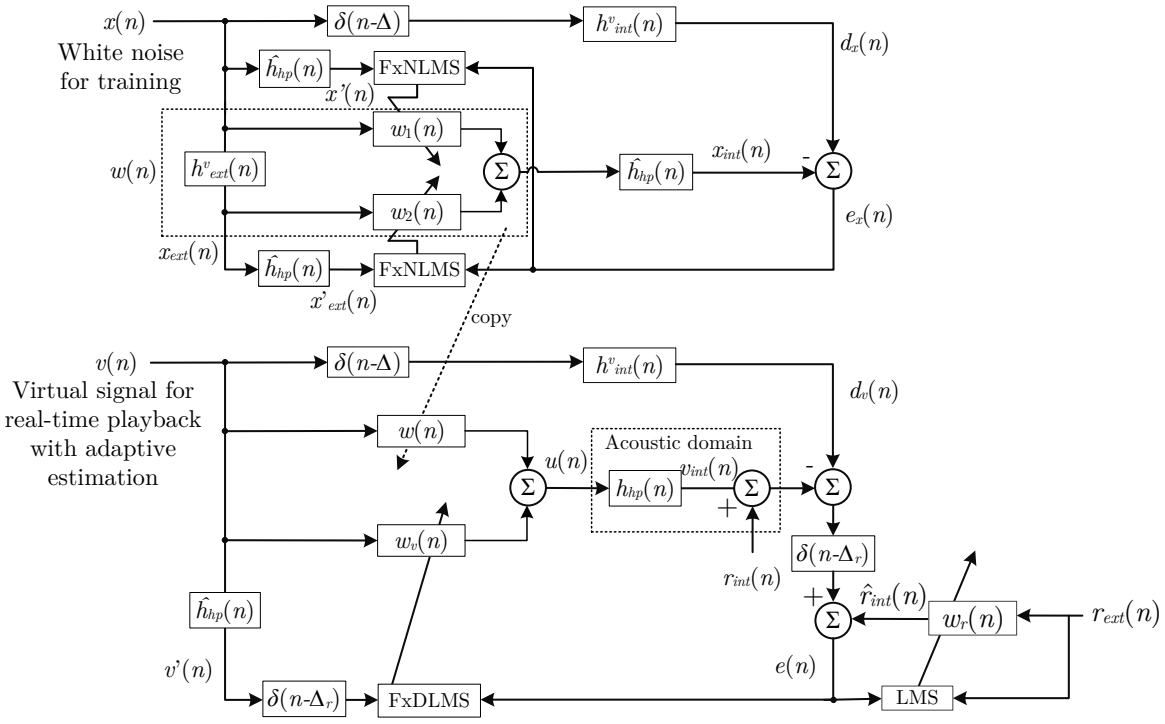


Figure 5.18: Modified adaptive equalizer with adaptive estimation for non-stationary virtual signals resolving causality issue between $r_{ext}(n)$ and $r_{int}(n)$

delay of the system especially in applications involving parallel architecture [175]. It has been shown that DLMS can converge in mean square sense, but there is stringent stability constraint in case of DLMS as compared to the LMS algorithm [175]. The only difference here is that non-stationary virtual signal is used in the adaptive process and thus, stability and convergence is the major issue here but not the steady-state error. In the case of non-stationary signal, trade-off between residual error and convergence speed makes the choice of step-size μ_v more critical to ensure stability [176]. Since the playback uses FxLMS for adaptive compensation, the modified block diagram in Figure 5.18, is now referred as delayed FxLMS (FxDLMS) algorithm and its weight update equation is now expressed as:

$$\mathbf{w}_v(n+1) = \mathbf{w}_v(n) + \mu_v\, \mathbf{v}'(n-\triangle_r)\, \left[e_v(n-\triangle_r) + e_r(n)\right] \ . \tag{5.13}$$

In other words, adaptive filter weights are updated using the delayed versions of

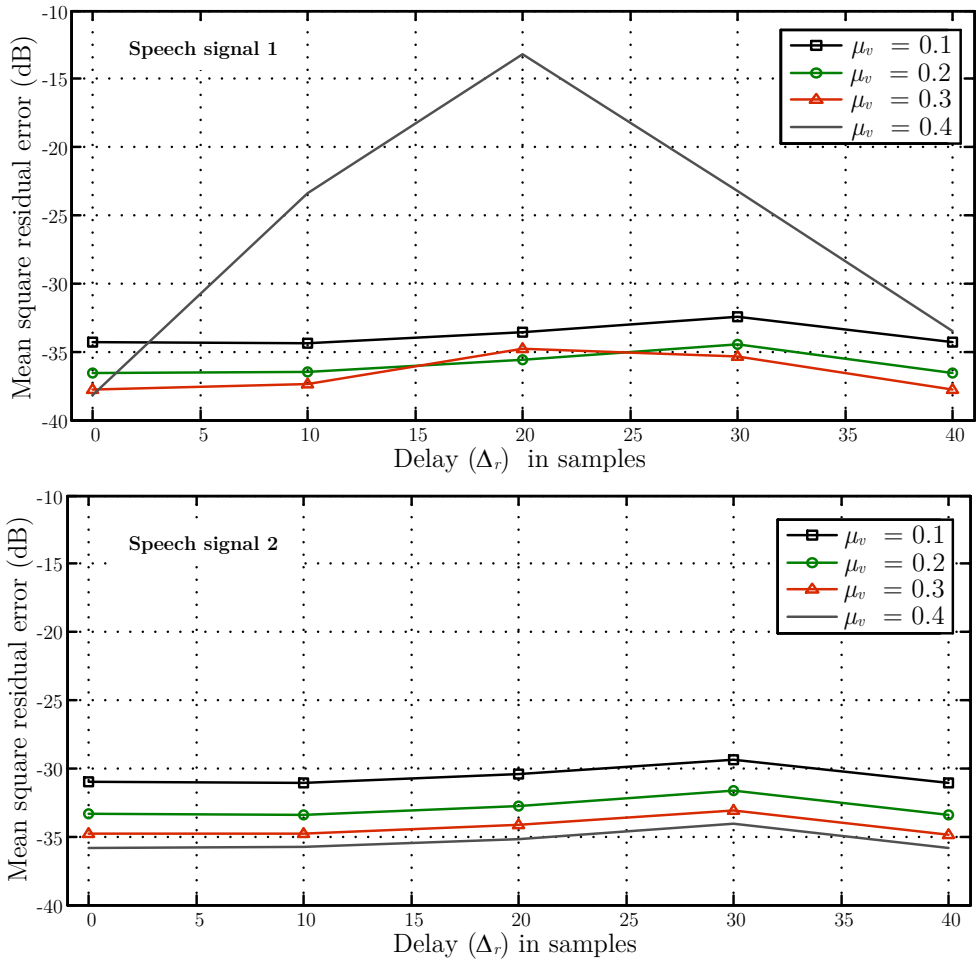


Figure 5.19: Trade-off between MSE and step size for different delay values ($\triangle_r$). Simulation is performed for azimuth: 40^o and contralateral ear.

error signal $e_v(n - \triangle_r)$ and filtered reference signal $\mathbf{v}'(n - \triangle_r)$. Weight update equation for adaptive estimation filter remains same with error signal is modified as:

$$e(n) = e_v(n - \triangle_r) + e_r(n)\,. \tag{5.14}$$

where $e_v(n)$ and $e_r(n)$ are defined in (5.12). Next, we study the effect of different delay values of Δ_r on the performance of residual error versus different step size for FxLMS algorithm of $w_v(n)$. We consider five delay values of 0, 10, 20, 30, 40 samples for 44.1 kHz, which corresponds to delay up to 1 millisecond, and is sufficient for adaptive estimation filter to avoid causality as shown above. Delay of 0 sample

is the reference case, i.e., adaptive compensation works as FxLMS algorithm. In addition, different step sizes from 0.1 to 0.4 are considered to study their effect on the MSE. Two male speech signals were used in the simulations with different transient characteristics. Results are shown in Figure 5.19. As shown, we are easily able to avoid the causality issue in adaptive estimation and achieve the same performance for adaptive compensation filter $w_v(n)$, as indicated by the low values of MSE. Higher value of step-size is favorable to ensure fast convergence and high MSE but it should be chosen slightly lower. This is observed in the top of Figure 5.19 for delay of 10 to 30 samples, where higher step size results in high MSE implying adaptive filter is not able to converge possibly because of sharper transients in the speech signal. Furthermore, delay size does not have much influence on the performance of FxDLMS, although higher delay value will ensure stability and convergence of adaptive estimation process. Interestingly, in our results, we observed that MSE decreases as delay is increased from 30 to 40 samples.

5.6 Conclusions and Further Improvements

5.6.1 Conclusions

In this chapter, we discussed some of the practical limitations of the NAR headset and presented techniques to address them. One of the important assumption for natural listening using NAR headset is that, it must use individualized BRIRs in the adaptive equalization process. Typically, acoustical measurement of individualized BRIRs is a tedious and cumbersome process to carry out, especially in a room environment. Another requirement with NAR headset is that, any one should be able to use it in a real environment and thus, a quick and convenient way measuring desired responses is needed. Using the NLMS technique with perfect sweep signal, individualized BRIRs can be measured readily using the NAR headset microphones. Head tracking can be mounted on the NAR headset to measure the BRIRs for lateral

azimuths.

One important limitation of FxNLMS algorithm for adaptive equalization of headset is that it does not always converge well for non-stationary virtual signals (such as speech) and instead FxLMS is preferred for stability at the cost of slow convergence speed. To ensure both stability and optimum convergence rate for any type of virtual signals, we included a proposed hybrid FxNLMS algorithm using white noise as training phase, while FxLMS is used for the real-time adaption during playback of virtual signal. A practical scenario was emulated in simulation by changing the physical secondary path over time. Online HPTF detection mechanism can be used to detect any large change in the secondary path model using running power estimate and consequently, its accurate estimate can be computed using NLMS technique promptly. It was observed that using HPTF estimation method, adaptive equalization works well similar to the reference case using the exact model of physical secondary path HPTF.

Furthermore, we extended the adaptive equation for non-stationary virtual signals with online adaptive estimation of real signals. However, since in practical scenario real signals can come from any direction and adaptive estimation does not guarantee the causality between reference and error microphone signals. This is resolved by adding a forward delay in error signal path and feedback path of the adaptive compensation of HPTF, which is now referred as delayed FxLMS (FxDLMS). Using simulation results, it was found that an appropriate step-size can be chosen with trade-off between convergence and MSE.

5.6.2 Further Improvements

Below are the two further possible improvements of the NAR headset especially for use in practical scenarios:

1. The main objective of the NAR headset is to reproduce the virtual sources while listener being aware of the surroundings. However, when surroundings

noise are too loud, we must try to control external noise either by amplifying the virtual sounds within permissible limits or generating an anti-noise signal to minimize the external sounds.

2. Head movement in real-life adds to our ability to perceive sound location in a natural manner. Head tracking is a must in AR related applications so that virtual auditory scene can be adapted according to the head movement as well as translation movement. In this context, computation complexity of the proposed algorithm is very crucial and the proposed algorithm should adjust to the dynamically changing head movements in real-time.

Chapter 6

A Hybrid Speaker Array-Headphone System for Immersive Audio Reproduction

In this chapter[1], we present a next generation home entertainment system for home scenario using a simple combination of loudspeaker array based reproduction using WFS and binaural synthesis over headphones. More specifically, physical array is used for the frontal auditory scene playback using WFS, while rear and side auditory scene is reproduced over the NAR headset using virtual WFS. For virtual WFS over NAR headset, we use multichannel version of the adaptive equalization presented earlier. For frontal auditory scene processing, Hybrid WFS methods are introduced in order to further improve the frontal perception. A detailed subjective study is conducted to investigate the performance of the proposed hybrid WFS set up.

This chapter is organized as follows. After a brief introduction of this work in Section 6.1, some of the related works are mentioned in Section 6.2. Section 6.3

[1] Part of this work is published in

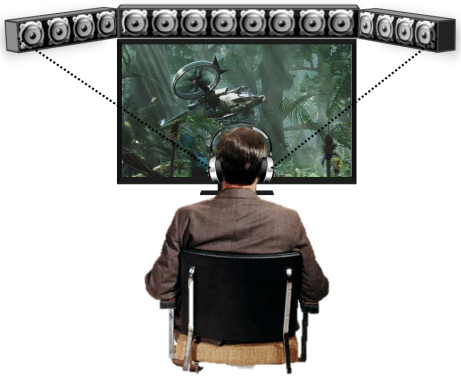
R. Ranjan and W.S. Gan, "A hybrid speaker array-headphone system for immersive 3D audio reproduction," *in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1836-1840.

gives an brief overview of the proposed hybrid WFS system followed by the WFS renderer and sound field analysis of the proposed enclosed structure in Section 6.4. Virtual WFS over headphones is presented in Section 6.5. In Section 6.6, we present the frontal auditory scene processing for physical array and detailed subjective study result is presented in Section 6.7. Section 6.8lists practical limitations of the current proposed system and possible improvements followed by the concluding section Section 6.9 highlighting key results of the objective and subjective experiments.

6.1 Introduction

Audio rendering systems have evolved significantly over the last couple of decades. Different sound technologies have been developed and being commercialized in the consumer market. Two channel stereophony based systems have advanced into multichannel setups creating an immersive experience for the listeners. However, such systems require complex loudspeaker placements, constrained by the room size. Listeners' movements are constrained as the best impression is only achieved in narrow sweet spot of the listening area. Binaural technology, which is also widely used for private listening over headphones, suffers from the problem of in-head localization and front-back confusions if non-individualized HRIRs and headphone equalization is used for the auralization. There are three main multichannel loudspeakers based sound field techniques, namely, vector base amplitude panning (VBAP) [15], higher order ambisonics (HOA) [27–31, 106], and wave field synthesis (WFS) [88, 97]. VBAP is a means for virtual source positioning in any direction using multiple loudspeakers in 3D plane. However, the basic underlying concept is same as in conventional stereo panning extended to a multichannel loudspeaker placement in arbitrary configuration. HOA and WFS are based on sound field synthesis principle to physically reconstruct the true spatial sound field in large sweet area as against narrow sweet spot in conventional multichannel stereo systems. They exhibit homogeneous sound field over extended listening area, while sounds can be perceived to

come from anywhere in the virtual space around you. Both HOA and WFS possess similar physical properties in theory if the listening area is surround by loudspeakers everywhere and is superior to VBAP in terms of stable virtual source position with listener movements and true spatial impression in the entire listening area. In addition, focused sound sources can be reproduced inside the listening area. A recent study by Lopez *et al.* in [177] to compare the sound distance perception between WFS and VBAP reveals that WFS turns out to have better distance perception cues than VBAP especially for frontal sources. Distance perception and capability to reproduce far and near-field sources is very important for immersive 3D audio systems. Spors [42, 43] compared the HOA and WFS with respect to physical properties, spatial sampling artifacts, perceptual attributes and practical aspects. It was highlighted that without spatial sampling there are only minor differences in the reproduced sound field but spatial aliasing results in coloration of sound field differently. HOA results in regular structures in spatial aliasing artifacts as compared to quite irregular spatial artifacts in WFS. However, both of them rely on perceptual insignificance of the aliasing artifacts. In terms of practical aspects, WFS requires convex shape loudspeaker arrays (linear/planar arrays for exact reproduction), while HOA requires strictly circular/spherical shape arrays, which is difficult to realize, especially in home scenarios. In addition, WFS complexity is fairly low and easy to implement as compared to HOA because of the use of spherical harmonics function and are ill-conditioned. Furthermore, we have seen several professional installations of WFS systems in entertainment venues, as well as research and development setups as discussed in 3. In recent years, there have been theoretical advancements as well as corporates increased interests in HOA for its application in practical scenarios, as discussed in 1. However, ambisonics still has not been a commercial success, and mostly used in research centres. With all these practical issues in mind, WFS is chosen in this work for its ease of implementation and can be easily realized in practice.



(a) Hybrid system model with frontal WFS array and headset

(b) WFS system built at DSP Lab, NTU Singapore

Figure 6.1: Proposed hybrid system in a home scenario

In this chapter, a new hybrid system is presented to overcome the practical and physical limitations of the conventional reproduction techniques. The proposed system combines WFS and binaural synthesis over headphones playback to reduce the need of installing many loudspeakers in a home entertainment setups and provide an immersive sound experience close to a full-fledged WFS setup. WFS is used to drive a frontal loudspeaker array, which is positioned on the top of the TV screen, to provide strong frontal localization cues. The rear and side auditory cues are presented over open headphones via a virtual WFS technique to complete the entire 360 degree auditory scene presentation. Figure 6.1 shows the proposed hybrid system model and a prototype of the WFS speaker array system built at the DSP laboratory in NTU, Singapore. One of the main challenges in such a hybrid system is the seamless integration of the two reproduction techniques: WFS and binaural technology over headphones. In other words, headphones should complement the WFS playback from the frontal array so as to provide listener an immersive feeling. In other words, listeners should not be able to distinguish between headphones and loudspeaker array playback. This would require headphones playback to be equalized so as to emulate speakers' playback (for rear and side auditory scenes) and at the same time one technique should not interfere with the other as both of them co-exist. Since, this is a multi-channel array system, headphones need

to be equalized for each of the individual speaker's response to listener position. Multichannel version of the adaptive equalization for the NAR headset is introduced in this Chapter to ensure that virtual WFS rendering over headphones perform as close to the physical speaker (sides and rear) arrays that would be required in a full-fledged WFS with 360^o speaker array system. Measurement results on dummy head showed that the virtual WFS performs very close to an enclosed array setup driven by WFS. We further present hybrid WFS method to enhance sound coloration of the physical frontal array by reproducing aliasing free high frequency components over open headphones, while physical array renders the low frequency components. In this way, strong frontal localization cues are preserved and there is no sound coloration in the synthesized sound signals at listeners' ears. Detailed subjective study was conducted, which confirmed our assumption and it was found that hybrid WFS method have better localization accuracy as well as overall audio quality than pure headphones playback method using virtual WFS. Rear and side virtual sources were also perceived to be highly externalized with good localization accuracy.

6.2 Related Works

There have been some works in recent years to combine WFS and other reproduction techniques. Menzel *et. al.* [118, 178] presented a novel system called "Binaural Sky" to reproduce a virtual headphone using binaural room synthesis. They used a pair of focused sources reproduction, using overhead WFS circular array, in front and close to the listener. These focused sources act as the trans-aural loudspeakers and they can be easily moved around by adjusting the driving signals. In [179], a simulation of wave field synthesis is presented for perceptual quality analysis of different complex setups using virtual WFS with the help of headphone playback. The main objective of the above work was to analyze the WFS through binaural playback. Through subjective tests, it was shown that azimuths were accurately

estimated by the subjects. In [180], Jose et al. proposed a method to combine WFS with HRTF processing for mitigating the absence of elevation cues in conventional horizontal planar loudspeaker array based systems. It is noted that loudspeaker array in azimuthal plane were used to generate the elevation cues. They used the HRTF based elevation cues as pre-processing stage to improve the height perception in median plane without the need of additional loudspeakers. Strauß *et. al.* [181] combined WFS with vector base amplitude panning (VBAP) techniques to create an audio interface for desktop applications. The main aim of this work is to immerse user in the visuals shown on screen and sources can be perceived behind or front of the screen with the help of array of transducers installed on the edges of a desktop. Another hybrid system was introduced by Wittek [101, 104], where phantom source reproduction using stereophony technique is used in conjunction with WFS to improve the sound coloration of the WFS loudspeaker array system. Since WFS reproduce correct sound field below the aliasing frequency and colors the sound spectrum above, WFS is used below aliasing frequency, while selected loudspeakers as phantom sources reproduced high frequencies to improve overall performance of the system.

In this chapter, we introduce a hybrid system combining loudspeaker array using WFS with binaural technology over open headphones. The main purpose of incorporating headphones is to complement the frontal auditory scene reproduction using WFS array with rear and side auditory scene reproduction using virtual WFS without the need of additional loudspeakers to be installed in a home scenario. Furthermore, we aim to improve the spatial aliasing performance of frontal WFS array by reproducing high frequency over headphones. Subjective study is carried out to investigate the performance of physical WFS array playyback, virtual WFS rendering and different combination of hybrid WFS systems.

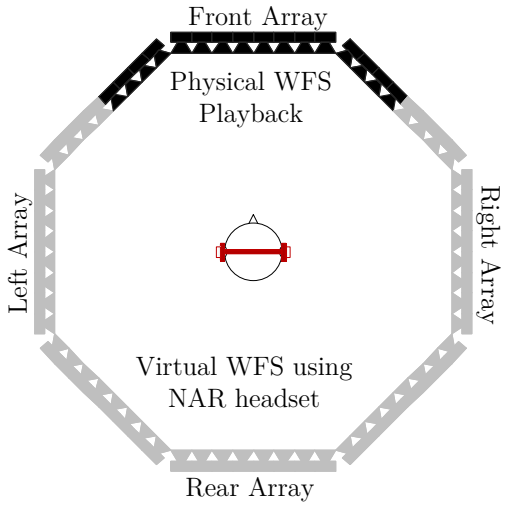


Figure 6.2: Proposed hybrid system structure (: Physical speakers; : Virtual speakers)

6.3 Proposed Hybrid System

The proposed hybrid system consists of a frontal loudspeaker array mounted on a visual display along with an open headphones worn by the listener to perceive both physical frontal sound and virtual side and rear sound image. Figure 6.2 shows the structure of the proposed hybrid system. An open and inverted U-shaped loudspeaker array is considered for the frontal projection, while symmetric virtual WFS array for side and rear is used for binaural reproduction over open headphones. The main objective here is to retain the spatial and temporal characteristics of WFS at the listener position by using a combination of two reproduction techniques. WFS renderer is being used in the back-end to compute all the loudspeaker signals (including physical as well as virtual ones) at the same time. The overall system block diagram is shown in Figure 6.3 and can be divided into three processing stages:

1) WFS renderer

2) Frontal auditory scene processing

3) Rear and side auditory scene processing using virtual WFS over open headphones

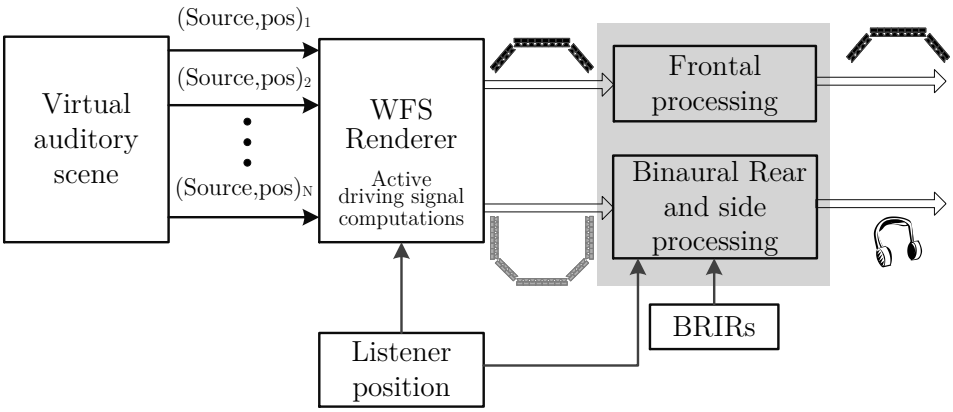


Figure 6.3: Overall hybrid system block diagram

In the next section, we briefly present the WFS renderer and sound field analysis for the proposed structure of the WFS array setup.

6.4 Wave Field Synthesis Renderer

WFS renderer is the processing core of the proposed hybrid system, which computes the driving signals of all the loudspeakers comprising the WFS enclosed setup shown in Figure 6.2. WFS is fundamentally based on the Huygens' principle, which states that secondary sources can be used to synthesize the natural wave fronts of the primary sources [89]. WFS driving signals (loudspeaker signals) are derived by solving the discretized Rayleigh Integral with sound field of a monopole point source and applying stationary phase approximation [48, 88, 97]. WFS driving signal equations are governed by (3.7 and 3.8) as explained in Chapter 3. Driving signal for each loudspeakers can be calculated efficiently by summing the delayed and weighted contribution from filtered source(s) signal(s). An important property of WFS is that it can synthesize virtual source even in front of the loudspeaker array (focused source), creating an illusion of source around the listener. The only restriction is that listener cannot be positioned between array and source. Driving signals for a focused virtual source is defined similarly as for non-focused virtual sources, while

reversing the phase components in (3.7) [48].

One of the drawbacks of using arbitrary shaped bend contours in WFS is that it introduces undesired reflections and leads to the artifacts in reproduced sound field. For bend contours, these prominent artifacts are introduced at those secondary sources, where normal vector ($\mathbf{nx}_n$) of the secondary source does not match with the local wave propagation direction of the virtual source sound field. Spors in his works [97, 98], proposed a secondary source criterion method by introducing a window function in the WFS driving function such that loudspeakers generating undesired reflections are muted. The window function depends on the type of virtual source (plane wave, point source or focused source) as well as on the listener position. Window function for plane wave source with normal vector ($\mathbf{nx}_s$) is defined as:

$$a_{pw}(\mathbf{x}_n) = \begin{cases} 1, & \langle \mathbf{nx}_s, \mathbf{nx}_n \rangle > 0, \\ 0, & else. \end{cases} \tag{6.1}$$

The operator $\langle \mathbf{x}, \mathbf{y} \rangle$ in (6.1) denotes the dot product between vector $\mathbf{x}$ and $\mathbf{y}$. The window function for point source generating spherical wave front with source position $\mathbf{x_s}$is defined as:

$$a_{ps}(\mathbf{x}_n) = \begin{cases} 1, & \langle \mathbf{x}_n - \mathbf{x}_s, \mathbf{nx}_n \rangle > 0, \\ 0, & else. \end{cases} \tag{6.2}$$

It should be noted that plane wave is special case of point source propagation when virtual source is in the far field. However, the window function for focussed source depends also on the reference listener position $\mathbf{x}_L$ in the listener area and is defined as:

$$a_{fs}(\mathbf{x}_n) = \begin{cases} 1, & \langle \mathbf{x}_s - \mathbf{x}_L, \mathbf{x}_n - \mathbf{x}_s \rangle > 0, \\ 0, & else. \end{cases} \tag{6.3}$$

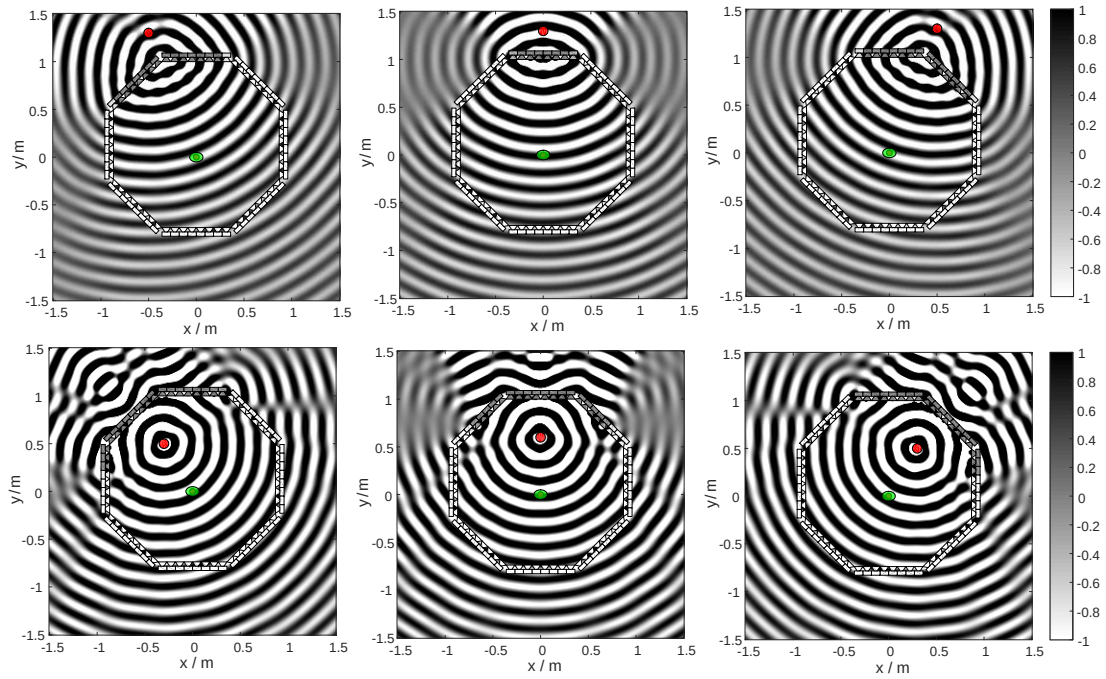


Figure 6.4: WFS Sound field plots for monochromatic source with frequency 2,000 Hz. Green circle indicates the listener position and red circle indicates the virtual source position. (Top: Point source; Bottom: Focussed source)

Using above window functions, WFS renderer computes the driving signals only for active loudspeakers for the enclosed WFS setup based on the source and listener position. If there are more than one source and/or listener present, for each source-listener pair active driving signals are computed and summed together before playback. Figure 6.4 shows the monochromatic sound field plots [2] for the WFS structure shown in Figure 6.2 for active loudspeakers using the above window functions. As shown in Figure 6.4, only desired loudspeakers are activated as per (6.2) and (6.3). For non-focused point source, depending on the source position relative to the loudspeaker array, appropriate loudspeakers are activated. For focused sources, desired loudspeakers are selected based on the position of source and listener relative to the array. It is evident that using an enclosed WFS array setup, sound field

[2] All the sound field plots in this chapter are generated using the Sound Field Sythesis toolbox for MATLAB, which can be found at https://github.com/sfstoolbox/sfs

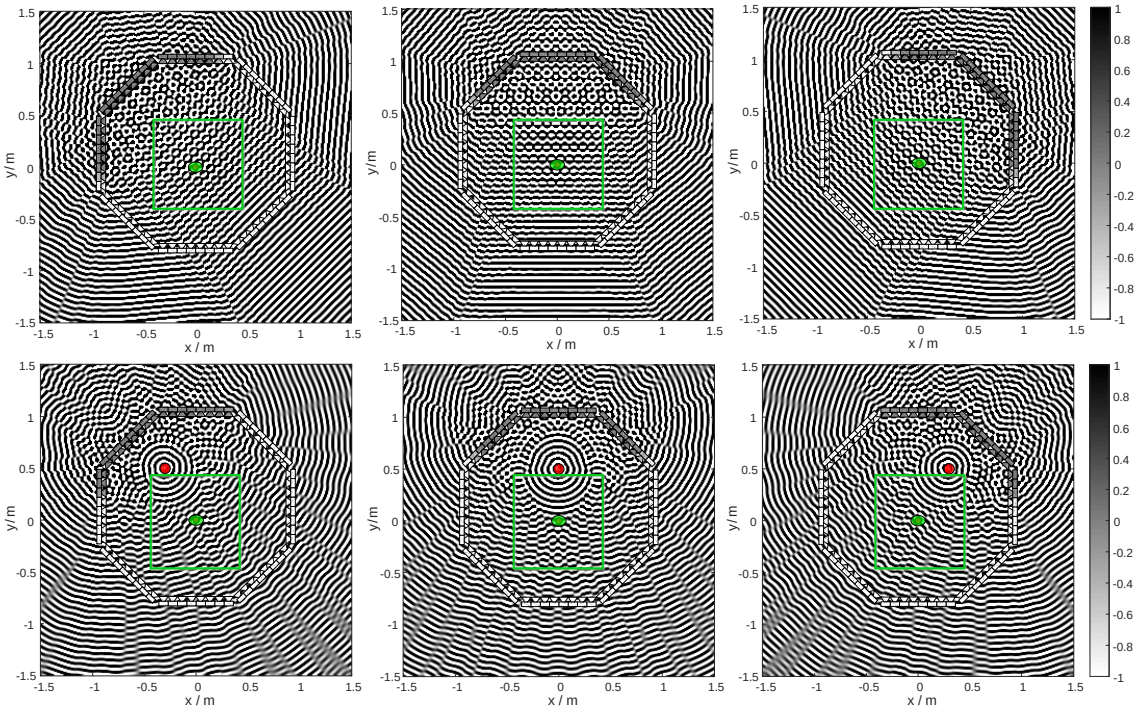


Figure 6.5: WFS Sound field plots for monochromatic source with frequency 6,000 Hz. Listener area is indicated by green square. f_{al} =3,000 Hz (Top: Far-field non-focused source; Bottom: Focussed source)

is correctly reproduced in entire listening area for non-focused point source, while area between the activated loudspeakers and virtual focused source is forbidden zone for listener. Reduced listener area for focused source is due to the fact that wave propagates towards the virtual source in this region, which is not the case in natural propagation of point source. In the next subsection, we study the spatial sampling artifacts in sound field for the WFS structure.

6.4.1 Spatial aliasing in the reproduced sound field

As we discussed in Section 3.4, due to spatial sampling and large loudspeaker inter-spacing in practice, WFS sound field is not correct everywhere in the listening area. Additionally, sound field can be faithfully reproduced only if frequency contents of the source are below the spatial aliasing frequency, or else distortions in the frequency response, physical sound field, and perceptual quality can be observed. Figure 6.5 shows the monochromatic sound field plots for far-field non-focused source and fo-

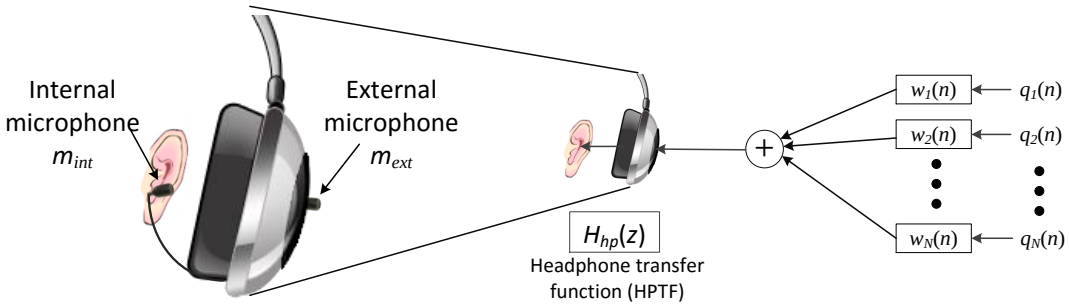


Figure 6.6: Virtual WFS using binaural synthesis over NAR headset

cused source with frequency above than aliasing frequency. Clearly, sound field is now distorted by undesired contributions from other directions. When source is in far field, spatial aliasing can be observed nearer to the loudspeaker array and decreases as one moves away from the array. For the focused source, aliasing is minimum near the source and increases further away from the source. Spatial aliasing in sound field mainly introduces the sound coloration but can also degrade the localization accuracy of the virtual sources. Spatial aliasing artifacts are studied in detail for both focused [106, 107, 182, 183] and non-focused sources [96, 105, 184, 185]. Perceptually, pre-echo artifacts (undesired wave-fronts perceived before the main wave-fronts) are observed in the case of focused sources as a result of the precedence effect in addition to the sound coloration, and are highly undesirable for broadband audio signals with vocals and music [106, 183]. Therefore, it is necessary to minimize its effect at least in the desired listening zone. In this work, we aim to minimize the spatial aliasing artifacts by reproducing the aliasing-free high frequency components of sound field over NAR headset . In the next section, we introduce the virtual WFS reproduction method over the NAR headset.

6.5 Virtual Wave Field Synthesis over Headphones

The main goal of virtual WFS over headphones is to provide listener an enriching listening experience similar to the WFS physical array setup. Therefore, for virtual

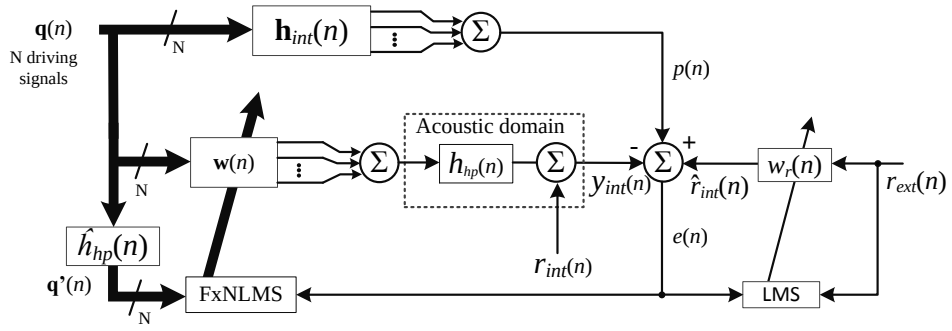


Figure 6.7: Multichannel adaptive equalization for virtual WFS using NAR headset

WFS to emulate real WFS playback in practical scenarios, speakers' room impulses must be used in the WFS synthesis equation (3.5) as:

$$p(r,n) = \sum_{i=1}^{N} q_i(n) * h_i(r,n)\,, \tag{6.4}$$

where $q_i(n)$ is the driving signal computed by the WFS renderer and $h_i(r,n)$ is the impulse response of i^{th}active loudspeaker. $p(r,t)$ is the WFS synthesized signal at any listening position, r. For rear and side auditory scene processing, driving signals corresponding to the virtual loudspeakers are used to synthesize the binaural signals at listeners' ears to be reproduced over open headphones such that they emulate the WFS enclosed array setup. Using the virtual WFS technique, we synthesize the binaural signals by convolving the driving signals computed by WFS renderer with the BRIRs of the virtual speakers to the listeners' ears and summing them together before playing through the open headphones. However, HPTF, which is also unique to every individual modifies the intended sound spectrum at listener's ears and must be compensated. We employ the NAR headset with open structure for binaural synthesis to perform the adaptive equalization of open headphones as shown in Figure 6.6. The individual driving signals ($q_i(n)$, $i = 1, 2, ...N$) are filtered with corresponding headphone equalized BRIR filters ($w_i(n)$, $i = 1, 2, ...N$).

A multichannel version of the adaptive equalization presented for NAR headset in Chapter 4 is implemented in this work to estimate the equalized filters, as shown

in Figure 6.7. WFS driving signals of the virtual loudspeakers are taken as input signals and internal microphones as error sensors for the multichannel version. The main objective is to compensate for individual HPTF, while adapting to the desired WFS response and make the headphones acoustically transparent so as to sound similar to physical WFS playback. Weight update equations for the multichannel adaptive equalization using FxNLMS is defined similarly for single channel FxNLMS as:

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) + \mu \frac{\mathbf{q}_i{}'(n)}{\|\mathbf{q}_i{}'(n)\|^2} e(n), \tag{6.5}$$

where $\mathbf{q}_i{}'(n)$ is the filtered reference vector corresponding to i^{th} driving signal $q_i(n)$ and $e(n)$ is the residual error signal defined as

$$e(n) = p(n) - y_{int}(n). \tag{6.6}$$

$p(n)$ is the desired WFS synthesized signal defined by (6.4) with speaker impulse responses ($h_i(r,n), i = 1, 2, ...N$) replaced by speaker BRIRs as $\mathbf{h}_{int}(n) = [h_1(n)\, h_2(n)\, ...h_N(n)]$ and $y_{int}(n)$ is the signal received at the internal microphone defined as:

$$y_{int}(n) = \left[\sum_{i=1}^{N} q_i(n) * w_i(n)\right] * h_{hp}(n). \tag{6.7}$$

In the next section, we will validate the results for virtual WFS using binaural synthesis with actual measurements of physical WFS array.

6.5.1 Virtual Wave Field Synthesis results

To evaluate the performance of the virtual WFS, virtual sounds reproduced over headphones must be similar to sounds coming from physical WFS array. In other words, both the spectral and temporal features of the actual WFS playback must be retained in the virtual sound field reproduction. The WFS frontal array prototype was built using 16 speakers, with 8 in the middle and 4 each on the either side

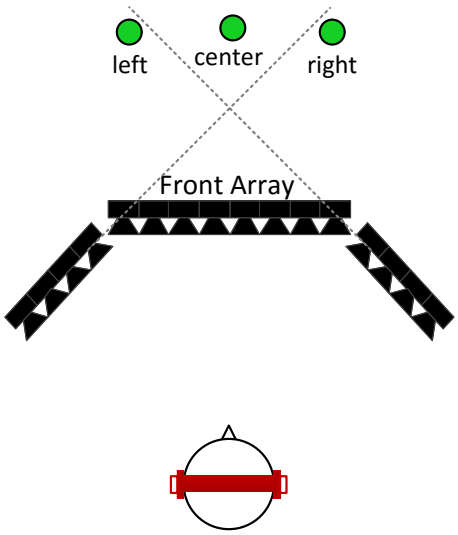


Figure 6.8: Measurement setup for WFS rendering of three virtual WFS non-focused sources

positioned at 45^o, as shown in Figure 6.1 and Figure 6.2. Loudspeaker spacing between any two adjacent speakers was 9 cm. Two 10-channel MOTU Ultralite-mk3 hybrid sound cards were used to drive the WFS frontal array as well as the NAR headset. Speaker responses were measured on the Neumann KU 100 head with and without NAR headset using AKG 417 miniature microphones mounted near ear opening. Open headphones AKG K702 were used for the measurement process. We evaluated the virtual WFS performance by recording response of sine sweep signal played through the loudspeakers. Measurements for virtual WFS via binaural synthesis were carried out using the WFS frontal array (Figure 6.2) by rotating the dummy head in steps of 90^oin clock-wise direction starting from $0^o$to 270^o. Therefore, there were 4 set of measurements corresponding to *front*, *left*, *right* and *rear* array, as shown in Figure 6.2. For all the 4 sets, three non-focused virtual sources, namely, *centre*, *left,* and *right*, were considered as shown in Figure 6.8. Centre virtual source is positioned $1\,m$ behind the array, such that all 16 loudspeakers are active. Left and right virtual sources were positioned respectively to left and right side of the array such that either left or right 4 loudspeakers were active along with the middle array. Individual BRIRs for all the 16 loudspeakers and all the 4 sets were measured

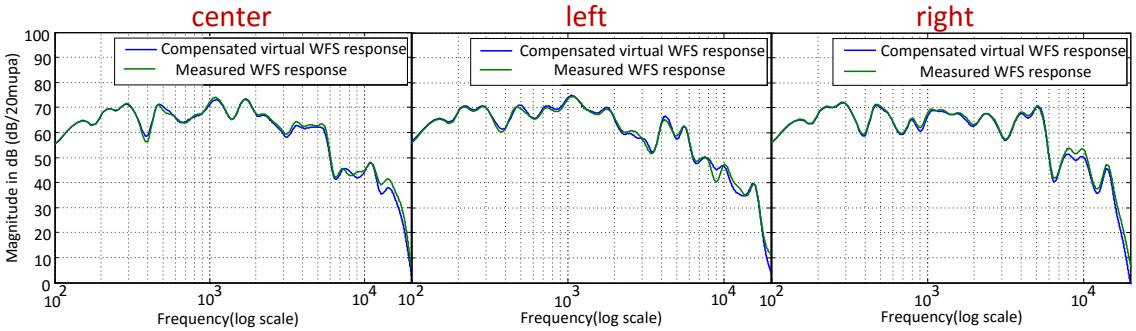


Figure 6.9: Compensated virtual WFS frequency response Vs Measured physical WFS frequency response for frontal WFS array. Three virtual WFS sources (left, center and right) were rendered as shown in Figure 6.8.

Table 6.1: Spectral distortion scores (dB) for the virtual WFS over headphones

Source	Front Array		Left Array		Right Array		Rear Array	
	Simulated	Measured	Simulated	Measured	Simulated	Measured	Simulated	Measured
Far Center	0.18	1.90	0.35	1.79	0.67	2.02	0.40	1.82
Far Left	0.17	1.84	0.10	2.51	0.59	1.80	0.20	1.51
Far Right	0.32	1.71	0.60	1.68	0.45	2.04	0.27	1.73

on the dummy head. Furthermore, BRIRs corresponding to all the 12 WFS virtual source positions (4 sets $\times$ 3 virtual sources) were also measured using dummy head.

Frequency responses of the compensated virtual WFS over headphones for WFS frontal array was compared with that of the measured physical WFS response. As shown in Figure 6.9, clearly headphone compensated virtual WFS response closely matches with the measured response for all the three virtual sources. Furthermore, SD score was used to objectively quantify the spectral error between frequency response of virtual source for virtual WFS over headphones and physical WFS array playback. Table 6.1 lists all the SD scores for the 3 virtual source positions and 4 sets of measurement of loudspeaker array. Two reference transfer functions were used to compute the SD scores namely, measured reference and simulated reference.

Measured reference is the actual measurement of WFS array response carried out on the dummy head by playing WFS virtual source through the loudspeaker array. Simulated reference represents convolution of WFS driving signals with respective speaker BRIRs and synthesized at the listeners' ears by summing them together. For

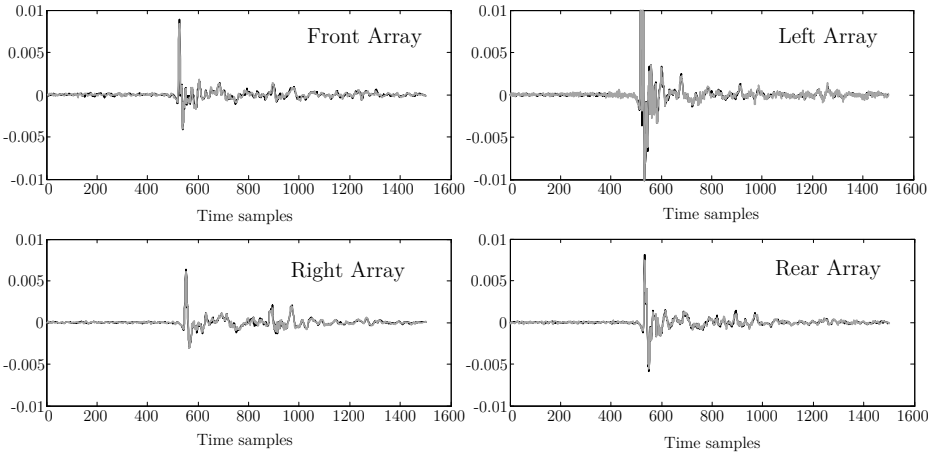


Figure 6.10: Impulse response plots: Real Vs Virtual WFS (Black: Measured IR Grey: Estimated IR)

estimated response, WFS driving signals convolved with the corresponding adaptive headphone equalized filters, summed together, played through the headset and recorded at the dummy head's ears. Clearly, using the multichannel version of hybrid FxNLMS, the spectral distortion is less than 1 dB with the simulated reference, which indicates signals synthesized using virtual WFS are exactly similar to the desired WFS response. However, with measured reference, slightly higher average SD scores were observed of around 2 dB, ensuring no perceptual difference between the physical and virtual WFS performance. The temporal characteristics of virtual WFS were validated by comparing the measured impulse responses (IRs) of physical WFS virtual source with that of virtual WFS.

Figure 6.10 shows the impulse responses of centre WFS virtual sources for all the 4 sets computed for the left ear of dummy head. Evidently, virtual WFS sources are reproduced accurately, with temporal features well matched with the measured IRs as shown. However, very little differences in magnitude were observed similar to the spectral distortion measurement. Hence, it can be concluded that using virtual WFS, virtual auditory scene can be reproduced similar to the physical WFS, giving us an immersive sound experience. Since, there are no physical speakers on rear and side in the proposed setup, we will use the virtual WFS technique to

reproduce the rear and side auditory scene. For the frontal auditory scene, either physical WFS array playback or virtual WFS can be used. In the next section, we present the frontal auditory scene processing and introduce a method to enhance the performance of frontal auditory perception.

6.6 Frontal Auditory Scene Processing

Once WFS render computes the driving signal, either frontal playback using WFS, or rear and side playback over NAR headset, or both is rendered based on which loudspeakers are active in the WFS enclosed array setup (Figure 6.2). In the proposed hybrid setup, we assume that physical WFS array playback will provide the strong frontal sound image quality along with the on-screen visual cues. Nonetheless, virtual WFS over NAR headset can also be used for frontal playback along with the rear and side auditory scene giving listener a personalized immersive sound experience, although in this case frontal perception may not be as strong as in the case of physical WFS array playback. In the next subsection, we will present the headphone isolation compensation for frontal WFS playback.

6.6.1 Frontal array Playback with Headphone Isolation Compensation

NAR headset with open earcup are used in the proposed hybrid system such that direct sounds from the frontal loudspeaker array pass through the headphones ear cup without much attenuation. However, due to the passive structure of the headphone, it acts as a low pass filter and high frequencies are attenuated depending on the transfer function of the headphones ear cup. Headphone isolation effects can be compensated by playing filtered driving signals through the NAR headset as shown in Figure 6.11. Driving signals computed by the WFS renderer are passed through compensation filters, $Wc_i(z)$ corresponding to each physical loudspeakers

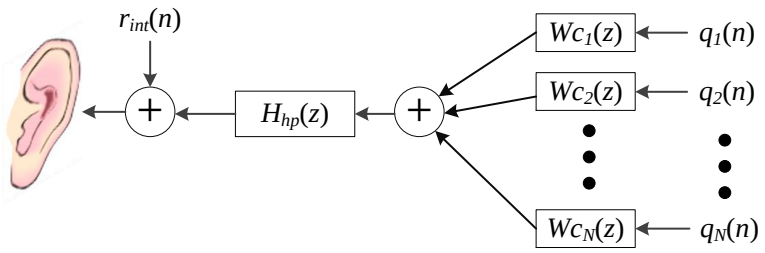


Figure 6.11: Headphone isolation compensation of NAR headset for frontal processing

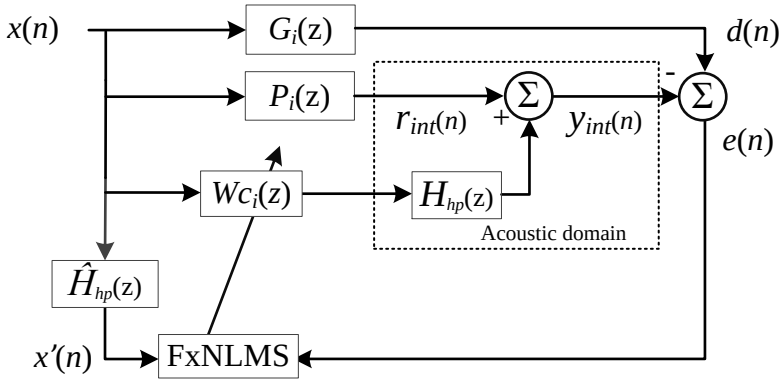


Figure 6.12: Headphone isolation compensation filters estimation

and played back over the NAR headset such that when added to the direct signal $r_{int}(n)$ received at internal microphone position, headphones become acoustically transparent. Compensation filters are estimated individually using the speakers' responses measured with and without headphones ($P_i(z)$ and $G_i(z)$, respectively) and employing a normalized version of the FxNLMS, as shown in Figure 6.12.

The main advantage of this approach is that the set of estimated compensation filters is valid for any virtual source positions rendered by WFS. To validate the headphone compensation approach, we measured the WFS virtual source frequency response at listener position in the center of the listening area by playing the driving signals through the frontal array. Figure 6.13 shows the frequency response of a WFS virtual source measured with and without NAR headset along with the compensated response. Clearly, headphones attenuation of 10-15 dB is observed in the high frequency region above 1.5-2 kHz for the headphone modified frequency response. After applying headphone compensation filters and adding with the direct signal,

the resultant compensated frequency response approaches the original frequency response measured without headset, as shown in Figure 6.13. In the next subsection, we present the frontal WFS playback using a hybrid method to further enhance the performance of physical WFS array by suppressing the spatial aliasing artifacts.

6.6.2 Frontal array Playback with reproduction of high frequency components over Natural Augmented Reality headset

We have seen in subsection 6.4.1 that sound field of the WFS frontal array is only correct up to spatial aliasing frequency. In high frequencies, correct sound field is superimposed by the incorrect high frequency components due to spatial sampling of the discrete loudspeaker array with large spacing. Aliased high frequency components results in timbre change and sound coloration of the desired source spectrum and varies with the source and listener movement. Furthermore, spatial aliasing can also degrade the localization performance, especially in terms of sound image width and locatedness. However, the dominant cue of the localization is provided by ITD in low frequencies below 1.5 kHz. In a perceptual study of linear WFS array with different inter-loudspeaker spacing by Wittek [101, 104], it was found that locatedness of the linear array increases significantly when aliasing frequency increased from 2.5 kHz to 7.5 kHz (i.e., decreased inter-loudspeaker spacing from ~12 cm to ~4 cm). In the same work, significant sound coloration were observed for linear array with $12\,cm$ or more loudspeaker spacing.

In this work, the idea is to substitute the aliased high frequency components of the WFS frontal sound field with correct unaliased sound field using virtual WFS technique over the NAR headset. Therefore, a hybrid WFS method is introduced with low frequency components reproduced over the frontal WFS array, while high frequency above a cross-over frequency is reproduced using virtual WFS over the

Figure 6.13: Headphone isolation compensation results for a WFS virtual source 1 m behind the array

NAR headset. Figure 6.14 illustrates how the reproduction of high frequency over NAR headset can be carried out along with WFS physical array playback for low frequency components with the help of virtual densely spaced speaker array. As shown in the figure, virtual speakers works in tandem with the phsyical speakers such that virtual speaker array with inter-spacing of 1 cm reproduces the high frequency content over NAR headset, while physical speaker array with spacing of 9 cm reproduces only the unaliased low frequency contents of the WFS virtual source. Frontal playback processing method is shown in Figure 6.15. WFS driving signals computed from the WFS renderer for a densely spaced speaker array is passed through a filter-bank comprising of a low-pass and high-pass filter with f_c as cross-over frequency. Low-pass filtered driving signals are played through the physical frontal array, while high-pass filtered driving signals are binaurally synthesized using (6.7) and subsequently played back over the NAR headset after headphone compensation. Figure 6.16 illustrates the hybrid WFS method for frontal auditory scene processing using the frequency responses of WFS array. As we can see in the Figure 6.16, physical WFS array with spacing of 9 cm (referred as WFS_9) suffers from spatial aliasing in high frequency above 3 kHz, while WFS array with 1 cm

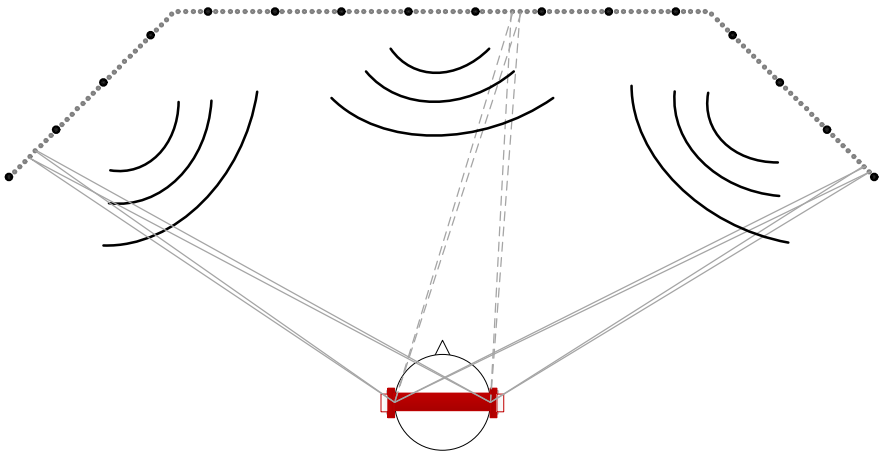


Figure 6.14: Hybrid WFS frontal playback: High frequency reproduction using virtual densely spaced speakers. Black circle represents the physical speakers with inter-spacing of 9 cm, while grey circle denotes the virtual speakers with inter-spacing of 1 cm.

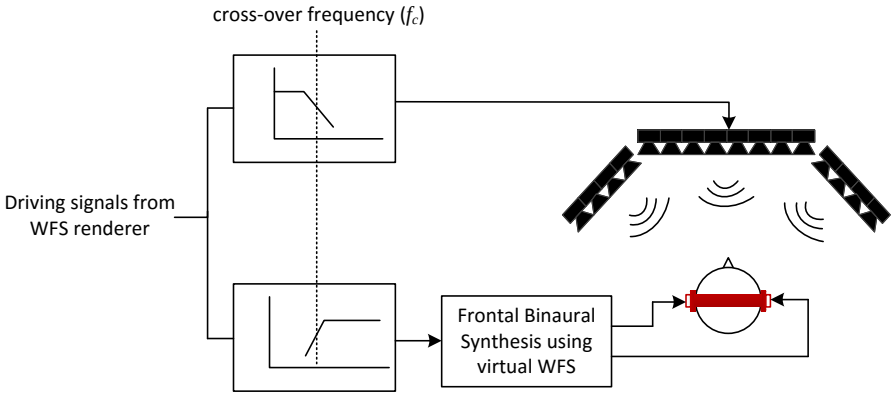


Figure 6.15: Frontal playback processing block diagram using hybrid WFS method

(referred as WFS_1) spacing possess very little aliasing in high frequency. It should also be observed that low frequency responses of WFS_1 matches with that of WFS_9 after level adjustment for larger number of loudspeakers. Therefore, spatial aliasing can be eliminated by combining the aliasing-free high frequency of WFS_1 with low frequency components of WFS_9. As shown in Figure 6.16, hybrid WFS combined spectra matches with that of WFS_1. There are two main advantages of this approach:

1) Almost aliasing free sound field is presented to the listener.

2) There is no need to apply headphone-isolation compensation as high frequency

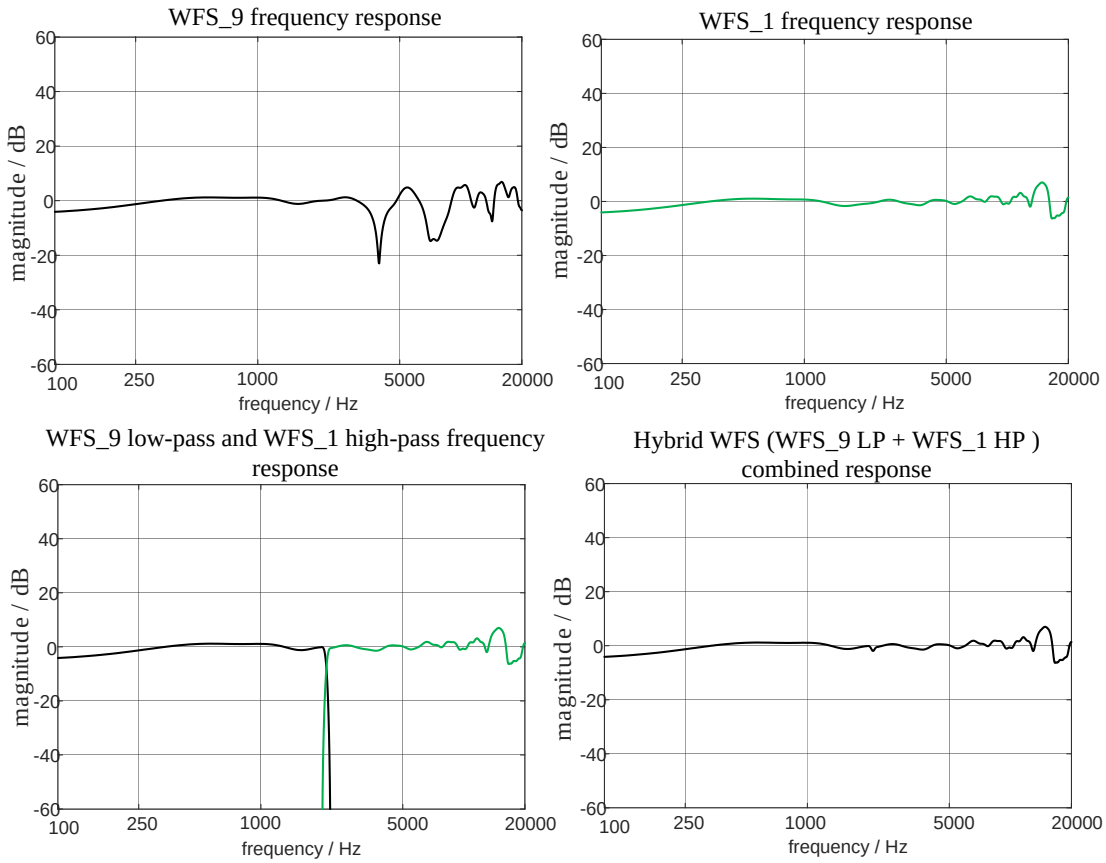


Figure 6.16: Frequency responses illustration for frontal auditory scene processing using hybrid WFS method. WFS_9 in top-left figure represents the speaker array response with inter-spacing of 9 cm emulating phsyical frontal array. WFS_1 in top-right represents speaker array response with spacing of 1 cm emulating virtual WFS. Bottom left figure shows the low-pass spectra of WFS_9 and high-pass spectra of WFS_1. Finally, bottom right figure represents the combined hybrid WFS response. Frequency responses were simulated for a virtual source 0.5 m behind the frontal array with listener position at 0.5 m in front of the array.

contents are reproduced over NAR headset.

By reproducing low frequency over the physical speakers, we aim to provide listeners with strong frontal localization cue dominant by ITD. However, it remains to be seen how localization is affected by reproducing high frequency content over the NAR headset, which also comes at high computational cost of driving many virtual speakers. A subjective test is conducted to study the localization performance of

hybrid WFS method and compared with that of the physical WFS playback as well as the virtual WFS. In the next section, we present the subjective study to investigate the localization and sound coloration performance of the frontal playback as well as rear and side auditory playback.

6.7 Subjective Study

A listening test was conducted using WFS physical array setup shown in Figure 6.1. For practical reasons, non-individualized BRIRs were used in the listening test. Non-individualized BRIRs were measured on the Bruel & Kjaer 4128D head and torso simulator for each of the 16 speakers in frontal array. Furthermore, dummy head was rotated 90^o three times in clockwise direction to measure the BRIRs for *left*, *rear* and *right* array. Listening test was divided into three phases:

- **Listening test 1**: Evaluation of localization accuracy
- **Listening test 2**: Sound coloration in WFS
- **Listening test 3**: Overall sound quality of WFS frontal playback

A total 15 subjects participated in the test with 13 males and 2 females all aged between 20-30 years. Listening test was conducted in a small quiet room with dimension of $3\,m \times 2.5\,m \times 4\,m$. Subjects position were fixed in the center of the room throughout the duration of the test. In the next subsections, we present the results of the three listening tests.

6.7.1 Listening test 1 - *Localization test*

The main objectives of this first phase of listening test is to evaluate the localization accuracy of virtual sound sources in terms of direction (azimuth), externalization, elevation, and locatedness. Subjects were asked to mark the azimuth direction on the scale of -180^o to 180^o with 0^o being the look ahead direction. For externalization

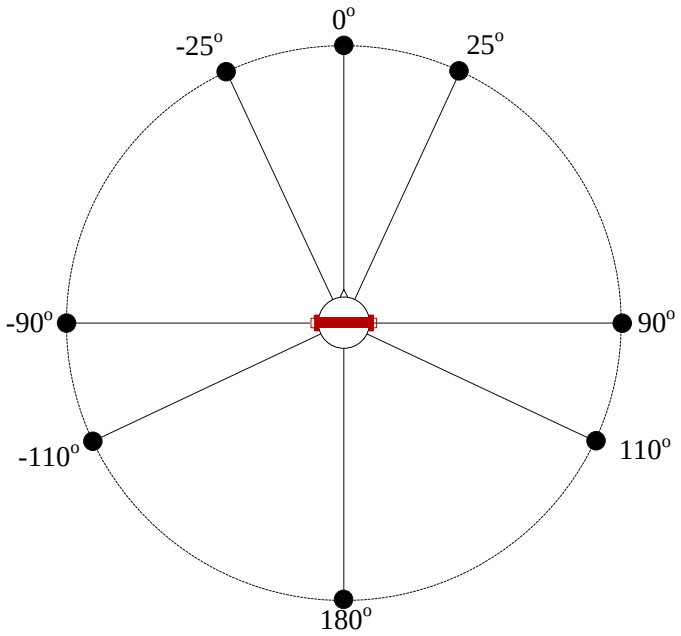


Figure 6.17: Target azimuth directions of virtual sources used in listening test

grading, subjects were asked to grade from inside head to much far. Ratings were defined as, namely, 1) Inside head, 2) Very near, 3) Near, 4) Far, 5) Very Far, and 6) Much Far. Listeners were told to take length of their hand as reference for far sources. Elevation were graded between 6 continuous levels, namely, 1) Much below ear level, 2) Below ear level, 3) Ear level, 4) Just above ear level, 5) Much above ear level, and 6) Near ceiling. For locatedness, subjects were asked to tell how well they can localize the sound and grade between 5 levels from very bad (score of 1) to very good (score of 5). All scores were graded with the help of a listening test GUI developed in MATLAB. There were a total of 8 virtual sources in Figure 6.17 rendered using WFS methods that were presented to the subjects. Three of them were in front direction (-25^o, 0^o, 25^o), two on the each side at $\pm 90^o$, and three on the rear side (-110^o, 180^o, 110^o). For frontal playback, 6 different playback methods were evaluated for localization accuracy:

1) *Pure WFS_9*: Physical WFS frontal array playback with listener always wearing NAR headset.

2) *Virtual WFS_9*: Virtual WFS over NAR headset with speaker array spacing

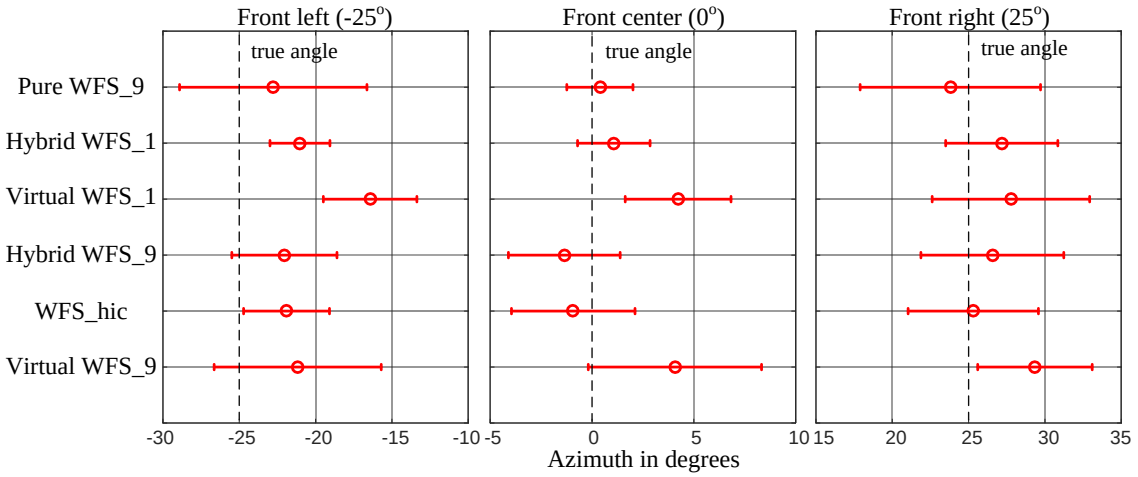


Figure 6.18: Azimuth accuracy for frontal virtual sources: Mean and their 95% confidence intervals

of 9cm.

3) *Virtual WFS_1*: Virtual WFS over NAR headset with speaker array spacing of 1cm.

4) *Hybrid WFS_9*: Hybrid WFS method for frontal playback with high frequency reproduced over NAR headset using virtual WFS with 9cm speaker spacing.

5) *Hybrid WFS_1*: Hybrid WFS method for frontal playback with high frequency reproduced over NAR headset using virtual WFS with 1cm speaker spacing.

6) *WFS_hic:* Physical WFS frontal array playback with headphone isolation compensation.

Since there was no physical speaker were present in the rear and side auditory scenes in the listening room, virtual WFS_9 and virtual WFS_1 were evaluated for localization performance for 5 virtual sources in rear and side in Figure 6.17. Pink-noise burst signal of duration 4 seconds were used in this test to study the localization performance. Pink noise was chosen as its spectrum is close to natural signals, as well as its similarity with how human auditory system perceives sound in logarithmic scale.

Listening test results for azimuth accuracy for three frontal sources are shown in Figure 6.18 with mean subjective scores for each of 6 the methods and their 95% confidence intervals. Below are some of the important observations from Figure 6.18:

- Mean subjective scores for Pure WFS_9 method is closest to the target virtual source directions as compared to others, although larger standard deviations are also observed for left and right sources. This could be due to the fact that NAR headset attenuates high frequencies, which may increase sound image width for some of the sources.

- WFS_hic method results in lesser standard deviation as compared to Pure WFS_9 for non-central positions. This can be explained by the fact that in the WFS_hic method, high frequency components were boosted by headphone isolation compensation over the NAR headset accompanied with physical frontal array playback. However, more results are needed for frontal azimuths to make it a general statement.

- The performance of the Hybrid WFS_9 is similar to the WFS_hic, which indicates that high frequency contents are identically reproduced in both the cases.

- Hybrid WFS_1 has the least standard deviations with mean scores only slightly deviating from target positions for non-central sources. This implies the virtual source is quite stable because of the fact that aliasing free high frequency components are reproduced over the NAR headset.

- Azimuth accuracy for Virtual WFS_1 and Virtual WFS_9 were observed to be worse as compared to the other 4 methods. In addition, standard deviation are also high in most of the cases, as shown in Figure 6.18. This can be attributed to the inherent drawback of binaural technology as non-individualized responses were used. Also, the Virtual WFS_9 method has higher standard deviations than Virtual WFS_1 method, except for frontal right source.

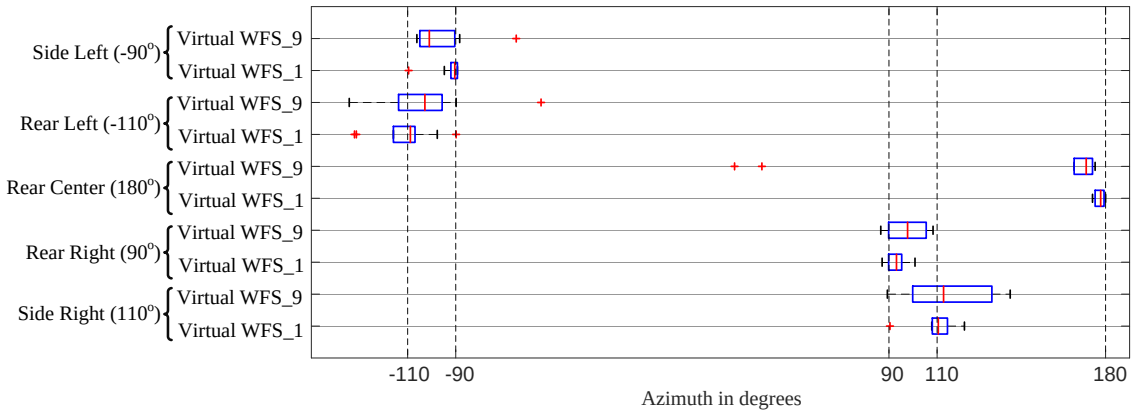


Figure 6.19: Azimuth accuracy for rear and side virtual sources: Center line in the box represents the median value, while edges of the box are 25 and 75 percentiles responses. Top and bottom lines represent the extreme subject responses, while outliers are shown in red.

Using pairwise t-test, it was found that mean scores for the Virtual WFS_9 were significantly different from rest of the methods, except for Virtual WFS_1. Furthermore, couple of subjects also reported front-back confusion for the two virtual WFS methods. Figure 6.19 shows the results for azimuth accuracy of the five rear and side sources with the help of box plots. In general it can be observed that the Virtual WFS_1 has better azimuth accuracy than the WFS_9 method. For the Virtual WFS_9 method, subjects could not distinguish between side left and rear left source positions and their median scores are also similar. For the Virtual WFS_1, azimuth accuracy is good, although few outliers were also observed. For source directly behind, most of the subjects perceived virtual source located slightly on the right side for both the methods, similar to the frontal array case.

Externalization results are shown in Figure 6.20 for all the virtual sources. Since externalization grades were consistent for all the three frontal sources, they were averaged and shown in Figure 6.20. As shown, frontal sources rendered with Pure WFS_9 and Hybrid WFS_1 were perceived significantly farther than the rest of the methods. For the Hybrid WFS_1, this can be due to the fact and there is no spectral peaks in the high frequency components, while for the Pure WFS_9 high frequency attenuation due to headset could lead to increased externalization grades.

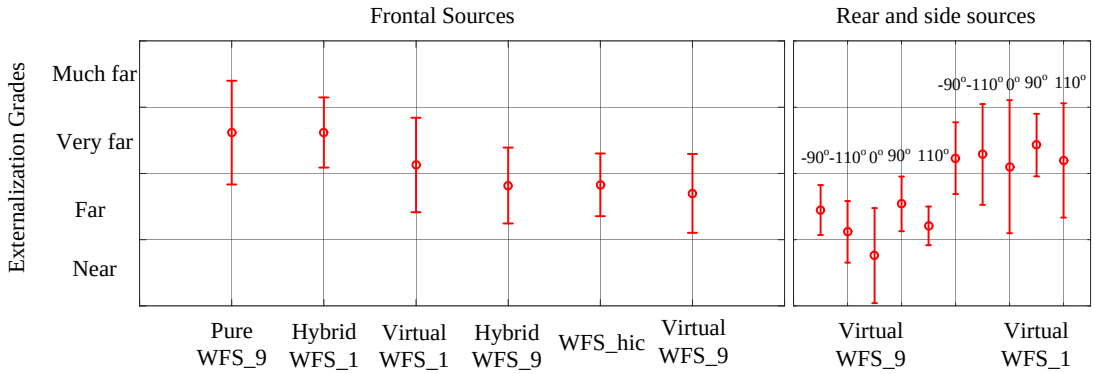


Figure 6.20: Mean externalization grades with 95% confidence interval for both frontal as well as rear and side sources

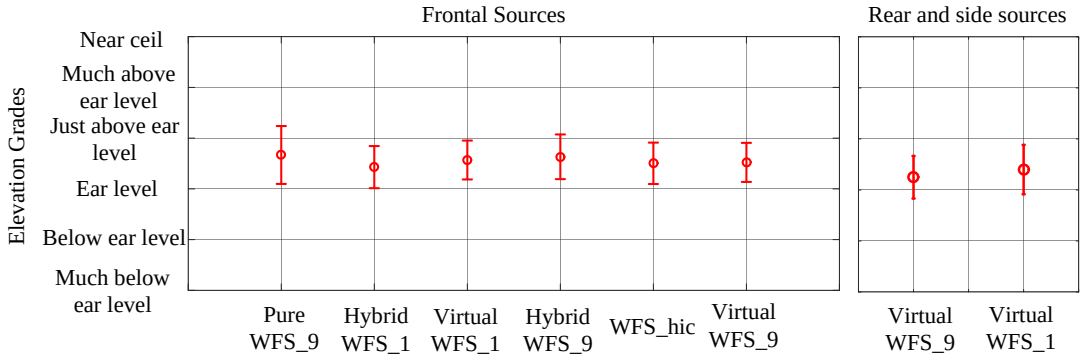


Figure 6.21: Mean elevation grades with 95% confidence interval for both frontal as well as rear and side sources

Additionally, Virtual WFS_1 externalization grade is also significantly higher than the Virtual WFS_9, which is also observed for all the rear and side sources. Furthermore, virtual source directly behind the head were perceived to be near to head using the Virtual WFS_9 method. Although mean externalization grades for Virtual WFS_1 were very high as compared to Virtual WFS_9, their standard deviation were also observed to be high for all the rear sources implying externalization of virtual sources varied across the subjects. Overall, most of the subjects perceived virtual sources to be externalized for all the sources. However, there was visual bias to the listeners as speaker array was placed directly in front of them and could have resulted in higher externalization grades, especially for virtual WFS methods. Elevation results are shown in Figure 6.21 averaged for all the corresponding sources

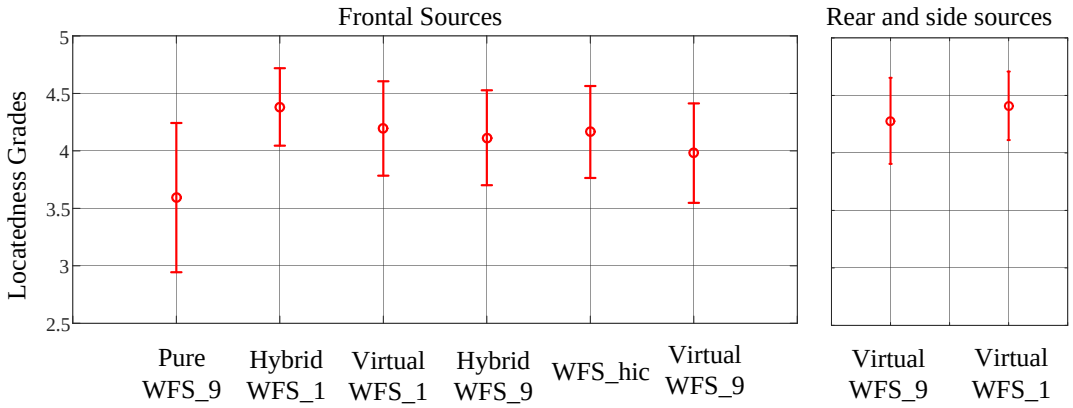


Figure 6.22: Mean locatedness grades with 95% confidence interval for both frontal as well as rear and side sources

in frontal and rear directions. Clearly, the elevation grades lie between *ear level* to *just above ear level* for all the different methods with no significant differences between their means. Locatedness result is shown in Figure 6.22. As shown, Pure WFS_9 have least locatedness, while Hybrid WFS_1 have the highest locatedness grade for frontal sources. Low locatedness of pure WFS_9 can be explained by the NAR headset attenuation, which might disrupt the high frequency spectral cues important for localization. Locatedness grades of other methods are similar and their means are not significantly different. Mean locatedness of the Virtual WFS_1 was found to be slightly better than Virtual WFS_9 method for both frontal as well as rear and side sources as shown in Figure 6.22. This can be attributed to the fact that the Virtual WFS_9 method has incorrect sound field in high frequency due to spatial aliasing, which might affect the localization performance.

6.7.2 Listening test 2 - *Sound coloration test*

The purpose of this listening test is to evaluate the sound coloration of different reproduction methods. More specifically, the goal is to investigate the spectral alterations in the synthesized ear signals due to the spatial aliasing of different WFS systems. Similar to the previous listening set, two virtual WFS methods (WFS_9

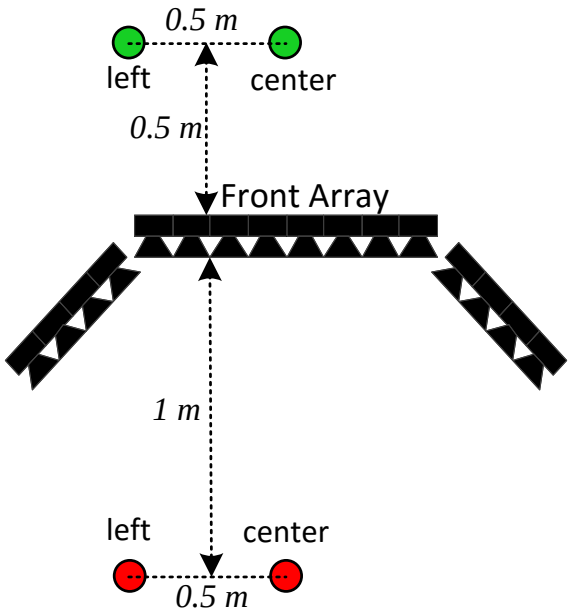


Figure 6.23: Experiment setup for sound coloration test. Source positions are indicated by green circle, while red circle indicates the two listener positions.

and WFS_1) were used. For hybrid method, we also varied the cross-over frequency resulting in three hybrid WFS methods, namely 1) *Hybrid WFS_1500,* 2) *Hybrid WFS_3000,* and 3) *Hybrid WFS_6000* with 1500 Hz, 3000 Hz and 6000 Hz as the cross-over frequencies, respectively. By varying cross-over frequency, we aim to investigate the optimum cross-over frequency for the Hybrid WFS method with the least sound coloration. Sound coloration was tested using virtual acoustics method, where all sounds are reproduced over headphones so as to ensure all other attributes except sound coloration is constant and does not change across the reproduction methods. Any sound coloration introduced by the virtual acoustics system is assumed to be constant throughout all the different methods. Similar method was used in recent works [186, 187], to evaluate the coloration in WFS. Similar to previous test, pink noise burst sequence was chosen as stimuli because of its highest sensitivity to any change in timbre. For evaluation of sound coloration of different WFS methods, standard MUSHRA [188] GUI interface [189] was considered with hidden reference. Two virtual sound source positions and two listener positions were

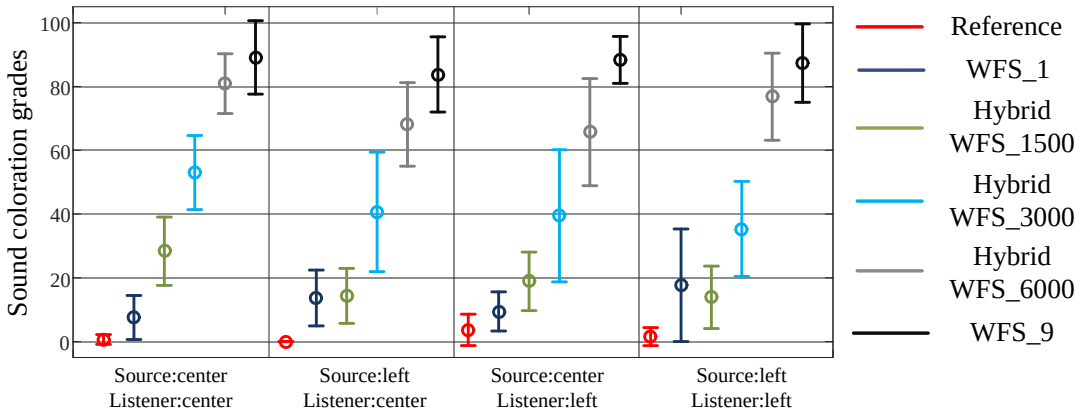


Figure 6.24: Mean coloration grades with 95% confidence interval for 5 reproduction methods

evaluated in 4 sets of experiment, as shown in Figure 6.23. Virtual source was positioned $0.5\,m$ behind the speaker array, while listener was positioned $1\,m$ in front of the array. The two source and listener positions were separated by distance of $0.5\,m$. In each set, one reference sound track along with 6 test tracks including one hidden reference was presented to the listener. For reference, two ear signals were synthesized by convolving the HRIRs of dummy head for virtual source position with pink noise burst stimuli. Test signals were processed using the respective 5 WFS methods and were normalized to match the loudness with reference sound. Subjects were asked to grade the processed tracks on a scale of 0 to 100 as compared to the reference sound. Score of 0 indicates test track is exactly similar to the reference, while score of 100 indicates timbre of test track is very different. Listeners were specially instructed to grade only based on timbre difference and not on localization or any obvious loudness difference.

Result for mean perceived coloration grades along with their 95 % confidence interval for the 5 reproduction methods is shown in Figure 6.24. The main observations from the results are summarized below:

- Reference sound was correctly identified with minimum coloration in all the cases.

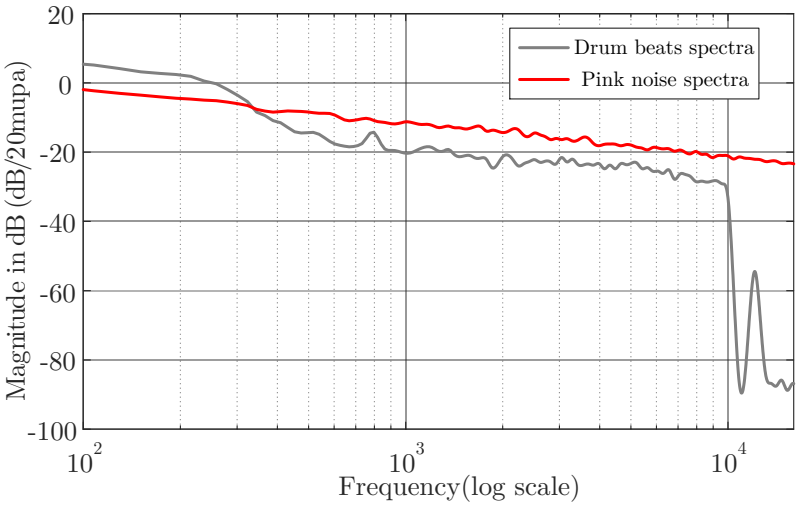


Figure 6.25: Frequency spectra of drum beats stimulus versus pink noise stimulus

- WFS_1 and Hybrid WFS_1500 methods have the least perceived coloration grades among all the reproduction methods. This is expected as both methods have very high spatial aliasing frequency, resulting in the least timbre difference.

- When virtual source is in the left, WFS_1 and Hybrid WFS_1500 have similar coloration grades with no significant difference between their means. However, when virtual source is in center, coloration of Hybrid WFS_1500 is higher than that of WFS_1. This can be explained by the fact that sound coloration in WFS also depends on source and listener position.

- As expected, WFS_9 with spatial aliasing frequency around 2 kHz have highest sound coloration followed by Hybrid WFS_6000 and Hybrid WFS_3000. Clearly, as the cross-over frequency increases, coloration increases with the increase in spatial aliasing in the synthesized sound signal.

- Mean coloration grades of all the methods except for the Hybrid WFS_1500 for two source and two listener positions were found to be insignificant at 95 % significance level.

6.7.3 Listening test 3 - *Overall audio quality test*

The main objective of this listening test is to evaluate the overall audio quality of proposed hybrid WFS method for frontal playback. For this phase, a drum beats stimulus with percussion is chosen for evaluation since its spectrum is similar to a pink noise spectrum (up to 10 kHz), as shown in Figure 6.25. Similar to sound coloration test, MUSHRA interface was used in the listening test for grading. However, reference sound in this phase is the virtual source rendered by physical frontal WFS array with subject wearing the NAR headset. There were 6 other reproduction methods under test in this phase:

1) *Hybrid WFS_9*

2) *Hybrid WFS_1500*: Same as *Hybrid WFS_1* in subsection 6.7.1 for localization test.

3) *Hybrid WFS_3000*

4) *Hybrid WFS_6000*

5) *Virtual WFS_9*

6) *Virtual WFS_1*

Three virtual sources were used in the evaluation for frontal playback, as shown in Figure 6.8 with listener positioned in the center of the listening area. Subjects were asked to rate overall audio quality on a seven-grade (-3 to 3) comparison scale as recommended in ITU-R BS.1284-1 [190]. The definitions of the seven-grade is defined below:

grade	definition
3	Much better
2	Better
1	Slightly better
0	Same as reference
-1	Slightly worse
-2	Worse
-3	Much worse

The overall audio quality as compared to reference sound were judged on the basis of two sound attributes:

1) Frontal image quality, i.e. the sound is predominantly coming from the front direction, and

2) Timbre clarity, i.e. good timbre quality in high frequencies and sound is brighter.

Additionally, any distortions in high frequency may result in poor audio quality as compared to the reference. Result for audio quality grades is shown in Figure 6.26. Reference sound was always correctly identified which imply that reference sound was clearly distinct from the test tracks. The Hybrid WFS_1500 method results in the best audio quality compared to the reference Pure WFS_9 sound, follows by the Virtual WFS_1 and the Hybrid WFS_3000. The Hybrid WFS_1500 method was graded significantly better than the Virtual WFS_1 despite the fact that spectral characteristics of both of them were identical. This result is in agreement with our initial assumption that by using the physical frontal WFS playback, the frontal perception can be more predominant, as compared to pure headphone playback method. Since the Hybrid WFS_1500 method reproduces low frequency components below 1.5 kHz using physical array and un-aliased high frequency components over NAR headset, it enhances the frontal perception of the Virtual WFS_1 method.

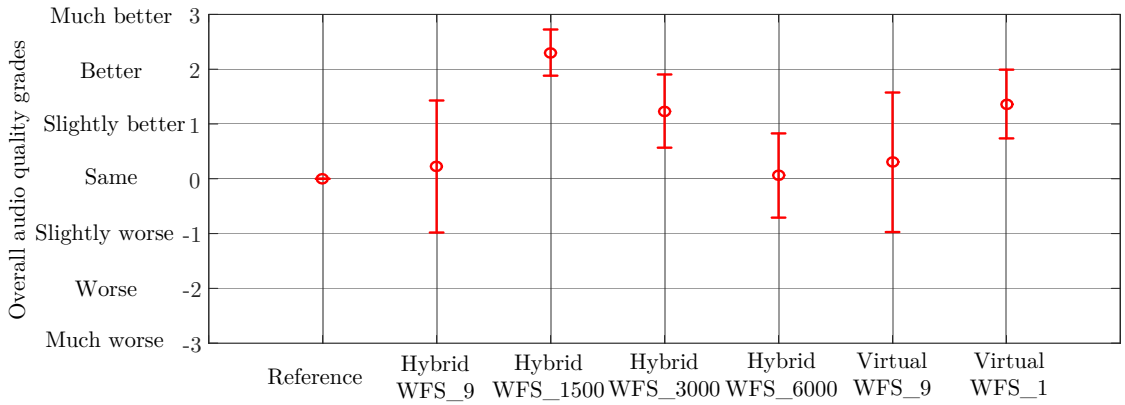


Figure 6.26: Overall audio quality grades: Mean scores with their 95% confidence interval

It should be noted that the reference sound, which is a physical frontal playback rendered sound, suffers from dullness in high frequencies due to headphone attenuation of NAR headset and that is why, most of the test sounds were rated with better audio quality grades. Mean scores of the Hybrid WFS_9 and the Virtual WFS_9 were close to the reference sound but with higher confidence intervals varying from *slightly worse* to *slightly better*. Both Hybrid WFS_9 and Virtual WFS_9 also have similar frequency spectra with significant spatial aliasing in high frequency as observed in sound coloration test. That is why some subjects rated them slightly worse because of the timbre distortions perceived as metallic sound. However, some subjects also rated them slightly better, which can be attributed to the fact that they have better clarity in frequencies as compared to the reference. Furthermore, as cross-over frequency were increased for the three hybrid WFS methods, audio quality of the rendered sounds decrease from much better to slightly worse. On the one hand, the Hybrid WFS_1500 has high timbre clarity as well as aliasing free sound field resulting in highest quality grades. The Hybrid WFS_6000 method on the other hand, has low timbre clarity as well as timbre distortion due to spatial aliasing up to 6 kHz and thus, resulting in mean audio quality score closer to the reference sound. In the next section, we discuss some of the limitations of the proposed system followed by the concluding section summarizing the main findings of

the proposed hybrid WFS setup

6.8 Limitations of the Proposed Hybrid system

Below are some limitations of the current proposed hybrid system with possible improvements:

- Limited to a 2D sound on a horizontal plane reproduction: This is one of the well known practical limitations of WFS as discussed in Section 3.3 and Section 1.2 because of the linear array only in horizontal plane and sources in median plane cannot be reproduced. There are few works in the literature to reproduce elevated sources. Montag in [5] employed multiple line arrays in the vertical plane for elevated sources by applying vertical amplitude panning. This is an interesting approach and can be investigated in home scenarios by employing double layered loudspeaker array mounted on top and bottom of the television screen. Elevation sensation can also be reproduced by combining WFS with HRTF elevation cues using a 2D array as shown by [180]. Similar technique can also be applied in the proposed hybrid setup to perceive height perception which is very important for multimedia application like gaming, movies etc.

- Coloration due to listening room acoustics: The purpose of the hybrid system is present the virtual auditory scene to the user such that one has the illusion of being there in the virtual environment. However, the listening room acoustics add undesired room reflections to the reproduced sound field. There have been several works in the past for compensation of the room acoustics [124–126, 191] which can be applied in our case as well.

- Limited to a single user: Proposed hybrid system can be easily extended to more than one user with the help of multiple NAR headsets and by applying adaptively equalized filters corresponding to each listener position.

- Fixed user position with no head movement: With today's advanced tracking devices, both user head rotation as well as translational movement can be tracked in real-time and accordingly compensation will be applied through NAR headset.

6.9 Conclusions

In this chapter, we presented a hybrid system combining a physical WFS array reproducing the frontal auditory scene, while rear and side auditory scene is synthesized over NAR headset using virtual WFS. Open headphones, which are embedded with two pairs of microphones, are used to adapt to compensate for the individual HPTF, which is essential for desired sound field reproduction. With emphasis on the strong frontal localization cues, we use the physical WFS frontal array along with visual aid to provide listener an immersive sound experience when used in conjunction with open headset for surround sound. WFS renderer is used as the processing core to compute all the driving signals and therefore, can provide seamless integration of physical WFS with the virtual WFS over headphones. Dummy head measurement results for virtual WFS show that spatial and temporal characteristics are retained in the virtual sound field reproduction as well. Although open headphones were used, high frequencies above 1.5 kHz were still attenuated by the earcup resulting in dullness of the high frequencies sound. Therefore, two hybrid reproduction methods were introduced to further enhance the frontal perception when listening with the NAR headset. In the first method, headphone isolation compensation approach was used so as to make the NAR headset completely transparent. Secondly, a hybrid WFS method was introduced, where aliasing-free high frequency components were reproduced over the NAR headset and only low frequency components were rendered over the frontal physical array. In this way, the entire audio spectra is almost free of sound coloration.

A detailed subjective study was conducted to investigate the performance of proposed hybrid WFS methods in terms of localization, sound coloration and overall audio quality. Since, the aim of the proposed setup is to provide listeners an immersive sound experience, it is important to provide listeners strong frontal perception along with the presence of rear and surround sound without the need of additional surround speakers. In the localization task, subjects were asked to tell the direction, externalization, elevation level and locatedness of the virtual source. For frontal virtual sources, it was found that Hybrid WFS methods had better directional accuracy as compared to only headphones playback method (Virtual WFS_9 and Virtual WFS_1). For rear and side sources, Virtual WFS_1 method performs better than that of frontal directions with good directional accuracy. Externalization of the Hybrid WFS method (Hybrid WFS_1) with minimum spatial aliasing and Pure WFS methods were observed to be significantly better than the others. Virtual WFS_1 method with minimum spatial aliasing clearly outperforms rest of the Hybrid WFS (WFS_3000 and WFS_6000) and Virtual WFS_9 methods. For rear and side sources, same trend was observed with high externalization grades for virtual WFS method with minimum spatial aliasing. Elevation level of all the virtual sources were perceived between ear level and just above ear level. Locatedness of the Hybrid WFS_1 method for frontal sources was observed to be highest, although only slightly better than the rest of the virtual and hybrid methods. For rear and side sources, there were no significant difference between the locatedness of two virtual methods with virtual WFS_1 being slightly better than virtual WFS_9. Overall, using the hybrid WFS method, when there is minimum spatial aliasing, results in high localization accuracy compared to other methods. Sound coloration test between different test methods indicated that Hybrid WFS method cross-over frequency of 1.5 kHz was optimum with minimum sound coloration along with virtual WFS method with similar frequency spectra. As the cross-over frequency spectra increases, spatial aliasing also increases, resulting in severe sound coloration. Fi-

nally, overall audio quality test was conducted using a drum beats stimulus in terms of frontal image quality and timbre clarity. Here also, it was interesting to find that hybrid WFS with optimum cross-over frequency was found to be better than the corresponding virtual WFS method. To summarize, the proposed hybrid WFS setup can be used in home scenarios with complete 360^o sound experience and better frontal playback performance when accompanied with high frequency reproduction over the NAR headset. The hybrid WFS set up proposed in this chapter is able to auralize virtual sound sources all around the listener with the help of virtual WFS rendering hundreds of loudspeakers, which comes at the cost of very high computational complexity. Furthermore, in practical multimedia application with 3D audio-visual immersive experience, multiple virtual sources will need to be rendered adding to the complexity of the proposed hybrid setup. Therefore, a real-time GPU implementation of a three-fold WFS framework is presented in the next chapter. Exploiting the inherent massive parallelism in WFS, hundreds of driving signals can be driven by rendering multiple virtual sources in real-time and at the same-time binaurally synthesized signal can also be computed using virtual WFS for multiple listeners.

Chapter 7

Fast and Efficient Real-Time GPU Based Implementation of Wave Field Synthesis

As concluded in the preceding chapter, hybrid WFS setup is a heavily computational intensive with lot of data parallelism to exploit. For this setup to be realized in practice, system must be able to render all the driving signals along with the binaural signals for NAR headset to provide a seamless listening experience to user. Furthermore, it must be scalable to any number and any size of speaker array setups. In this chapter, we present a fast and efficient real-time GPU implementation of WFS systems.

This chapter[1] is organized as follows. Related works on GPU implementation of WFS systems is mentioned in Section 7.2. In Section 7.3, we give an overview of GPU architecture and CUDA programming model highlighting on ways to optimize the GPU performance. Real-time WFS framework is presented in Section 7.4. In Section 7.5, we present the GPU implementation of WFS with optimization techniques and memory overhead reduction methods. Simulation result is explained in Section 7.6

[1] This work has been published in

R. Ranjan and W.S. Gan, "Fast and efficient real-time GPU based implementation of wave field synthesis," *in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7550-7554.

with key findings reported in the concluding Section 7.7.

7.1 Introduction

Wave Field Synthesis (WFS) is a spatial audio reproduction technique capable of reproducing high fidelity sound in large listening area with the help of loudspeaker arrays [48]. Listeners get to experience realistic sound scene as they are free to move in the listening area and virtual sources are localized as close as possible to their true positions. In practice, such high fidelity systems require many driving units, while rendering multiple virtual sources, making WFS a heavy computationally complex system. Commercially viable solutions from SonicEmotion [116] and IOSONO [115] can render up to 64 real-time sources for 24 and 32 driving units, respectively. Furthermore, before hardware implementation can be realized, the synthesized sound field quality needs to be analyzed across the entire listening area. Thus, two processing blocks, namely, synthesized signal block and sound field synthesis block are added to the system, which can be used for real-time analysis of a WFS setup. Collectively, we call such system: a three-fold WFS setup. Overall, WFS is a highly parallel data intensive application but suffers from limited resource problem and low throughput when implemented on today's multi-core PC platforms.

With the advent of graphics processing units (GPU), maximum resource utilization can be achieved using parallel computing architecture. Recently, modern GPUs like GTX590, C2075, K10 etc. have hundreds to thousands of processing cores, which can handle massively parallel and computationally intensive applications such as WFS. Essentially, algorithms written for small-scale multicore PCs need to be sufficiently parallelized and adapted for multithreading architecture to take full advantage of today's GPUs. Additionally, in real-time spatial audio applications, GPU must process the audio data within a fixed time interval, while also taking account of the GPU-CPU data transfer overheads. This makes parallelization the most critical task for performance improvement.

In this chapter, we present a generic real-time implementation of three-fold WFS setup on GPU using CUDA [192] technology with MATLAB [193]. Low level parallel programming language, CUDA is used to achieve peak performance by giving complete control of GPU architecture to the user. Concretely, the main objective of this work is to develop a fast real-time implementation of WFS, which would ultimately be deployed for the proposed hybrid WFS setup presented in Chapter 6. High system throughput has been achieved by efficiently mapping the massive data parallelism into WFS and thereby, taking advantage of running thousands of threads in parallel. Simulation results for GPU implementation show that peak system throughput of 1,400 Mega samples per second (MSPS) can be achieved with 20 folds improvement over CPU based implementation.

7.2 Related Work

Due to the advent of more powerful GPU, we are seeing new works related to the real-time spatial audio processing applications like, WFS on GPU platforms. Theodoropoulos et al. [194, 195] implemented WFS on different multicore platforms including GPU with the focus on architectural perspectives of these platforms. They reported speed up of around 10-20 times on GPU against Intel core 2 duo PC and estimated up to 64 real-time sources rendering for 96 loudspeakers. In [196], real time implementation of WFS was proposed on GPU and CUDA using NU-Tech framework [197] with peak speed up achieved up to around 4 times, although there is no mention of number of real-time source rendering. In [191], authors implemented WFS and a room compensation block with added computational complexity on three different GPU platforms. Their implementation achieved real-time rendering up to 50, 80 and 300 sources for Tesla, Fermi and Kepler architecture, respectively when room compensation was not applied for 96 loudspeakers. In contrast to above works, our implementation is based on hybrid time-frequency approach, which has lesser

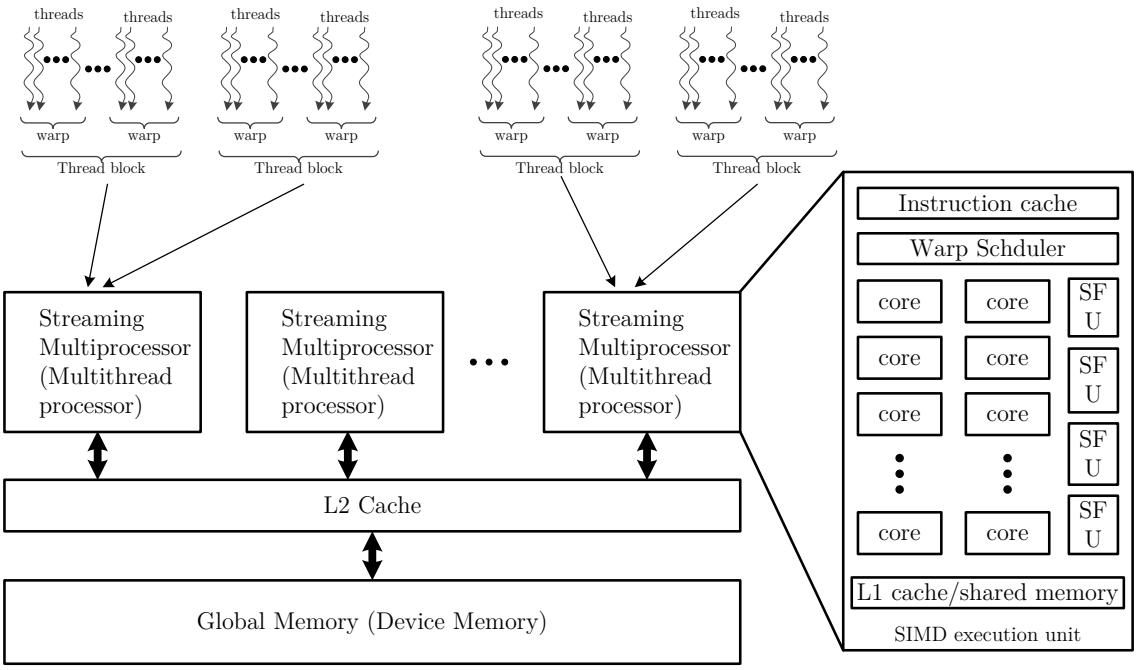


Figure 7.1: NVIDIA GPU architecture overview. Adapted and modified from [7]

computational complexity. In addition, two processing blocks are also implemented in GPU for sound field analysis. System throughput is used as a better measure of system performance instead of speed up. In the end-to-end comparison with [191] and our Fermi GPU, we obtained around 4 fold improvement in number of real-time sources for 96 driving units. In the next section, we present an overview of the NVIDIA GPU and the CUDA programming model.

7.3 GPU and CUDA Overview

With the increasing number of multiple processors and multithreading capabilities, GPU is becoming popular for general purpose parallel computation for wide ranging scientific applications like sparse matrix solvers, molecular simulation, air traffic flow analysis, computational physics etc. Recently, multichannel audio systems are also getting popular at the cost of massive computational complexity to drive all the speakers in real-time. GPUs are being utilized to optimize the signal processing operations like FIR filtering, fast convolution, FFT and inverse transforms [198, 199]

required in multichannel audio systems. GPU is designed such that it is dedicated for high data intensive and parallel computation tasks rather than the control flow and caching mechanism in the case of CPU. Therefore, CPU is optimized to provide low latency task, while GPU supports high throughput by executing tasks in parallel.

A typical NVIDIA Fermi GPU architecture is shown in Figure 7.1 [7]. The GPU consists of scalable number of streaming multiprocessors (SM), which are interconnected through L2 cache and global device memory. Each SM comprises of 32 cores (processors), 4 special function units (SFUs), software programmed shared memory, constant memory and dedicated multithreading scheduler. CUDA is the programming language, which executes a program in parallel as thread using single instruction multiple data (SIMD) software model. CUDA works in conjunction with CPU. While CPU works as host and executes the sequential program, it transfers data to GPU for parallel execution by GPU as CUDA kernel for multiple threads. In the next subsection, we highlight importance of thread blocks, warp and coalesced memory access to optimize GPU performance.

7.3.1 Thread Blocks, Warp and Coalesced Memory access

Threads in a GPU are grouped as thread block (TB), which are executed concurrently on a multiprocessor. Within a TB, threads are executed as warp and all threads in a warp are scheduled together for execution. A warp is a group of 32 threads executed as a batch by SM, which means 32 instructions are dispatched and executed simultaneously on the 32 physical cores. That is why, size of TB is usually taken as multiple of 32 threads to maximize the GPU efficiency i.e.,

$$size\,of\,TB = Q \times 32\,; \qquad Q\,is\,an\,integer. \tag{7.1}$$

But due to the limited size of registers and shared memory in a single SM, maximum number of threads that can be executed at once in a TB is 1024. Since

size of thread blocks are decided by the programmer, the assignment of number of threads per block should be done carefully to maximize the utilization of available resources. Less number of threads per block cause load latency in global memory access. In addition, there should enough be number of thread blocks to be executed simultaneously on all the multiprocessors. Number of TBs to each SM is dependent on the requirement of shared memory and registers by each TB. More memory and registers per block further limits the number of TBs per multiprocessor. For example, if shared memory of size 48 KB is used for data sharing among 1024 threads in a TB, we have only $(48KB/1024 = 48)$ 48 bytes available for storage of data in each thread. Furthermore, we need to have at least 2 TBs per SM so that multiprocessor is not idle during thread synchronization. Therefore, we must choose lesser number of threads per TB so as to maximize the GPU efficiency.

Warp scheduler in each SM uses pipe-lining mechanism to avoid memory latency. For a TB of 256 threads or set of 8 warps, first warp will be executed first. If warp need to read/write data from global memory and while it is waiting for the memory access, SM schedules next set of warp for execution and so on. Since global memory latency is usually around 100 clock cycles, till the time 8^{th} warp is scheduled and 1st warp is still waiting for the data from global memory, then another TB is scheduled for execution to hide the global memory latency. The time for memory access depends on the algorithmic complexity as well as traffic on the GPU bus.

Another important issue in maximizing in the GPU performance, is to ensure coalesced memory access. When a warp executes an instruction that access global memory, it coalesces the memory access of threads into one or more memory transactions. In other words, the data to be read/write must be in contiguous memory locations so that they can be translated into fewer memory accesses by warp because data transfers to and fro from the global memory implicitly affects the instruction throughput and kernel latency. Data re-organizations are needed to ensure maximum coalesced memory access within a warp. In this context, thread block configu-

rations are also crucial to the total throughput of the GPU program. Thread blocks in CUDA can be be chosen as up to 3D matrix and we should choose optimal TB configuration to maximize the coalesced memory access. We will discuss about this further in the Section 7.5 on how to choose optimal TB configuration. Next, we discuss different GPU related overheads limiting overall performance.

7.3.2 GPU Overheads

In the above subsection, we explained how different optimization techniques can be applied to maximize the GPU efficiency once CUDA kernel is launched from the host environment. However, there are other overhead which may affect the overall execution time of an application:

- GPU environment initialization time
- GPU memory setup time
- GPU CUDA kernel launch time
- Data transfer time from host CPU to GPU and vice-versa.

Out of the above, first overhead is due to setup of the GPU environment, which is just one-time. Second overhead is due to time taken in setting up or free some memory on GPU and is usually in the order of few microseconds. GPU CUDA kernel launch time is also usually in the order of 10 microseconds. These three overheads are usually insignificant as there are enough parallelism in the application to be exploited in GPU. Data transfer overhead, which is the time taken to copy data from CPU memory to GPU device memory via the PCIe bus, is a significant factor in deciding how much a workload can be accelerated on GPU. Moreover, it depends on the amount of data to be transferred to-and-fro for GPU processing and limited by the PCIe bus bandwidth. It is possible that data transfer overhead in some applications is more than than the GPU actual computation time and thus, limiting

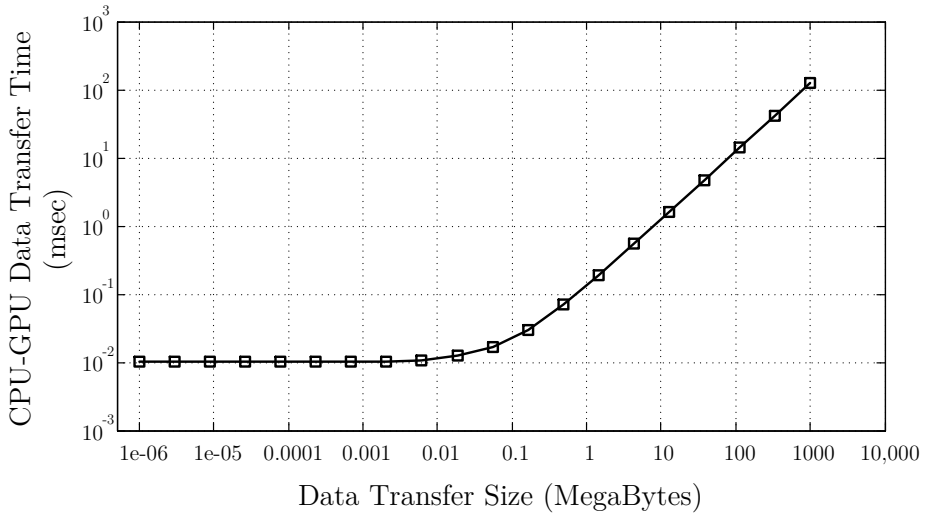


Figure 7.2: Predicted data transfer time using (7.2) for PCIe bandiwdth of 8 GB/s corresponding to Nvidia Tesla C2075 GPU and $\alpha = 10\mu s$

the overall performance [200]. Boyer et al. [201] proposed a model to predict the CPU-GPU data transfer time as function of data size (D):

$$T_{overhead}(D) = \alpha + \frac{D}{Bw} \tag{7.2}$$

where α is the fixed latency representing transfer of first byte and Bw is the PCIe bus bandwidth. Figure 7.2 plots the typical data transfer latency for a Nvidia Tesla C2075 GPU with bandwidth equal to 8 GB/s on PCIe x 16 Gen2. Value of α is taken as $10\mu s$ as found out in other works [201, 202]. Clearly with small data transfer size of less than 10 KB, overhead is dominated by α and constant at $10\mu s$. For larger data transfer size ($>$ 1 MB), overhead increases linearly reaches upto 100 msec for 1 GB of data. With multichannel audio systems requiring a lot of data transfers, this overhead can be a bottleneck in such real-time applications. Below we show how data transfer overhead can affect overall GPU performance using an example for multichannel audio streaming.

Multichannel real-time audio application: Real-time audio systems require multiple source signals (Ns) as input and output multiple loudspeaker signals (L).

In addition, they may also require some static or dynamic filtering. Lets consider a fast block convolution based implementation, which would need $2M$ samples (M previous and M current samples) of each source signals to be transferred to GPU for processing. After computation of all the loudspeaker signals, $2M$ samples for each loudspeaker signals need to be transferred back to CPU. With $M = 512$ samples as audio incoming buffer, overall execution time including data-transfer overhead must be less than 11.2 msec at 44.1 kHz. Lets consider number of sources as Ns and number of loudspeakers L equal to 100. Total amount of data-transfer size with word size of 64 bit (8 Bytes) can be computed as:

$$D = Ns \times 2M \times 8 + L \times 2M \times 8 \approx 1.6\,MB$$

From Figure 7.2, data-transfer size of 1.6 MB will result in approximately 0.2 msec of data transfer overhead which is 1.72 % of the total required execution time (11.6 msec) and thus, limiting the overall speedup of GPU. With the case of dynamic filtering, data transfer overhead becomes more critical as all the FIR filter coefficients need to be transferred every frame as against static case, where coefficients are transferred only once. Therefore, we must try to minimize the data-transfer between CPU and GPU as much as possible even if that mean executing some task on GPU without any speedup. Additionally, pinned (page-locked) CPU memory, batching small data transfers into one large transfer and overlapped execution with other host processing can also be used to reduce data transfer time [203].

7.4 Real-Time WFS Framework

The three-fold WFS spatial reproduction setup is shown in Figure 7.3. Using a linear array of loudspeakers, driving signal (block PB1) for each loudspeaker is governed by (3.7, 3.8) and subsequently, used in synthesis function (3.4, 3.5) altogether for processing blocks PB2 and PB3, as shown in Figure 7.3.

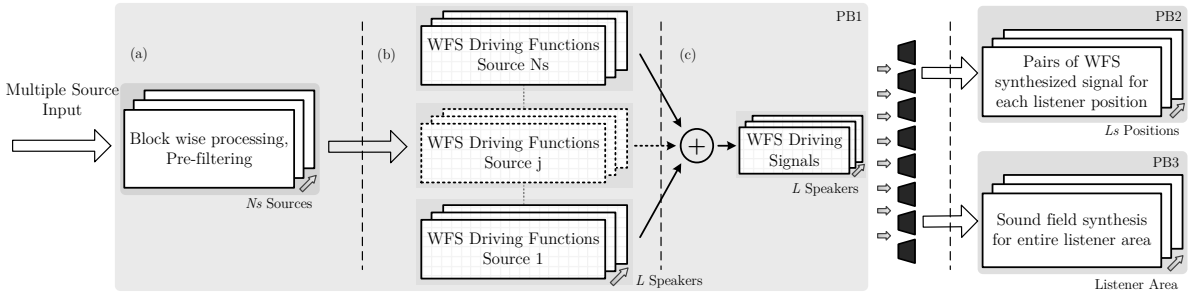


Figure 7.3: Real-time WFS processing framework with processing blocks (PB1, PB2 and PB3)

The processing blocks PB2 and PB3 can be used to assess synthesized sound field quality for different kinds of WFS setup. PB2 computes the virtually synthesized binaural signals at Ls listener positions for real-time playback over headphones. PB3 synthesizes snapshots of the sound field in the entire listening area ($dimx \times dimz$) for different test signals. $dimx \times dimz$ represents the number of sample points in the entire listener area. These snapshots can also be used for analysis of several artifacts, like spatial aliasing, truncation effects and amplitude errors.

For real-time processing, one frame of audio data must be processed within the t_{frame} ($framesize/samplingfrequency$). At the same time, we should aim to maximize the system throughput by processing more data within t_{frame}. The real time implementation of WFS setup is based on overlap-save technique with 50 % overlap using frame size of M current samples and M previous samples. As shown in Figure 7.3, driving signal block (PB1) is divided into three stages, namely, (a) pre-filtering for multiple sources, (b) individual driving signals due to all sources at each loudspeaker, and, (c) compute driving signals using reduction sum of output matrix at stage (b). Real-time filtering using block convolution is generally faster in frequency domain [191]. Therefore, pre-filtering is implemented in frequency domain, while the rest of the stages have been implemented in spatio-temporal domain. Recent contributions [204, 205] have also shown that real-time filtering of multiple data can be processed concurrently on GPU. Pre-filtering is carried out by element-wise complex multiplications of $2M$ FFT transformed multiple source data.

Table 7.1: Computational Complexity of different computation stages in PB1 (MAD: Multiply/Addition; ADD: Addition)

Stage	time-frequency [this work]	time [194, 195]	frequency [191, 196]
FFT (a)	$2M \log 2M \times Ns$	-	$2M \log 2M \times Ns$
MAD (a)	$12M \times Ns$	$8M^2 \times Ns$	$12M \times Ns$
IFFT (a)	$2M \log 2M \times Ns$	-	-
MAD (b)	$L \times M \times Ns$	$L \times M \times Ns$	$8L \times 2M \times 2Ns$
ADD (c)	$L \times M \times Ns$	$L \times M \times Ns$	$2L \times 2M \times Ns$
IFFT (c)	-	-	$2M \log 2M \times L$

7.4.1 Computational complexity of WFS

WFS driving signals can be efficiently computed in time-domain using delayed and weighted version of the pre-filtered source signal. For multiple sources, driving signals for each source is summed together. However, pre-filtering of the source signal can be efficiently computed in frequency domain as against time domain using the FFT. Table 7.1 summarizes the computational complexity of different computations stages in PB1 for the three implementation approaches. Clearly, time-frequency approach seems to have the lowest complexity for larger values of virtual sources and speakers ($say,\ Ns = 100\, and\, L = 100$) but also has high memory usage because previous samples of driving signals need to be stored for delayed value of pre-filtered source signals. Both frequency and time domain approaches have high arithmetic density due to complex arithmetic operations and circular convolutions, respectively, resulting in higher complexity. Blocks PB2 and PB3 are also implemented in time-domain using weighted and delayed contribution from driving signals at listener positions with computational complexity $L \times M \times 2Ls$ and $L \times dimx \times dimz$ respectively.

7.5 GPU Implementation

Most of the audio processing is done in GPU using low level CUDA programming language along with MATLAB as host environment, controlling the GPU execution.

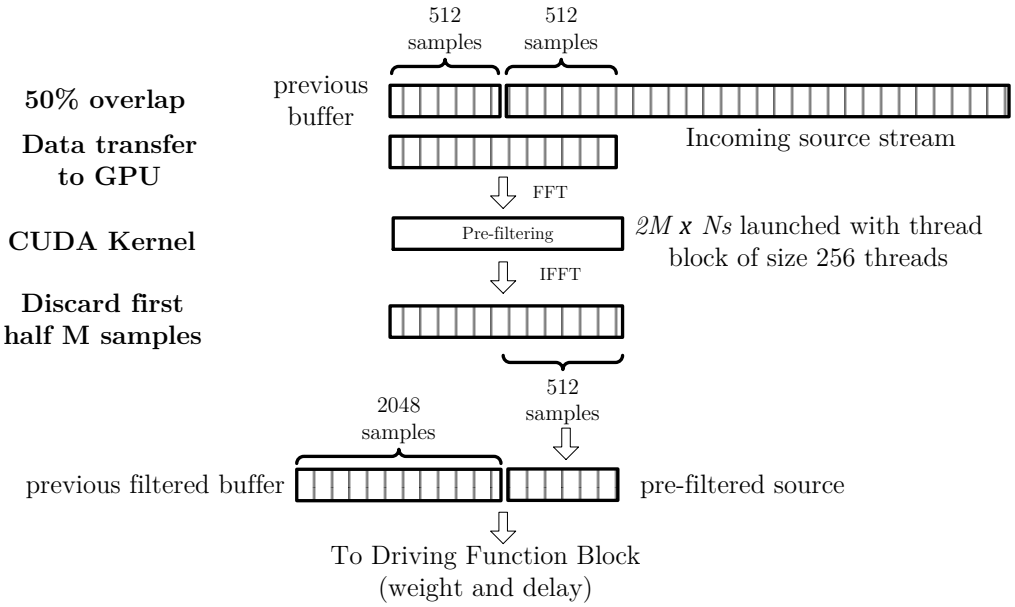


Figure 7.4: Block-wise processing of incoming source data and pre-filtering of one source ($M = 512$) : Stage (a)

Recently, MATLAB has added the support for GPU computing to its parallel computing toolbox (PCT) to take advantage of the parallel computing from MATLAB environment [206]. MATLAB along with the CUDA kernels [207] serves as a useful tool for the fast development of existing MATLAB applications onto GPU using custom CUDA functions, as well as overloaded MATLAB functions for GPU. The datasets to be computed are carefully partitioned into multiple contiguous blocks to take advantage of the data reuse using the on-chip cache and exploit coalesced memory access as much as possible. WFS algorithm is also segregated into parallel functions to exploit maximum data parallelism. Other CPU-GPU optimizations include shared memory, constant memory, data reorganizations, and overlapped executions are also taken into account to further speed up the processing time. We will now describe the implementation of each parallel task on GPU along with the optimization choices made for the best performance.

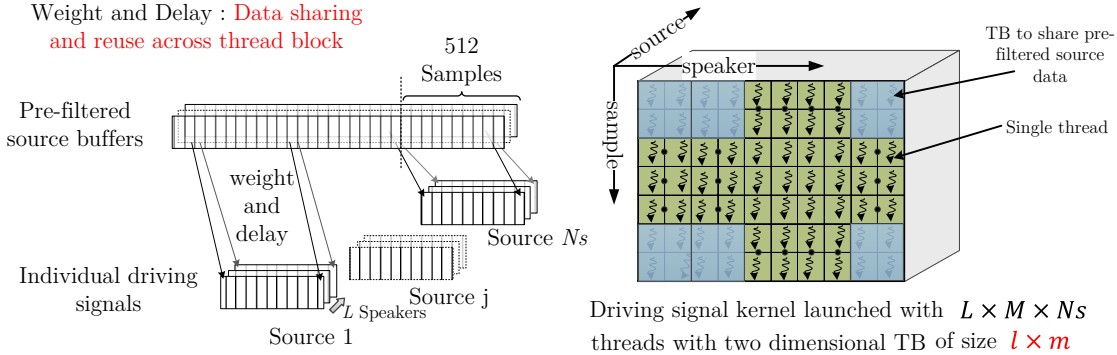


Figure 7.5: Individual driving signals computation and its kernel matrix ($M = 512$) : Stage (b)

7.5.1 Pre-filtering of multiple sources

Once the current M samples of audio source data is available, total $2M$ (M current and M previous samples) samples of the input data is transferred to GPU global memory for pre-filtering of the all the sources in parallel as CUDA kernel as shown in Figure 7.4. Computational complexity of this task is in $O(2M \times Ns)$, where Ns is the number of sources. A total of $2M \times Ns$ threads are launched with thread block size of 256 threads. Each thread computes one complex multiplication for a single source sample with the corresponding filter coefficient. Shared memory is used to synchronize the common filter coefficients within a thread block. Both MATLAB built-in overloaded FFT function for GPU, as well as NVIDIA CUFFT library [208] are considered, since both can perform frequency transformations for multiple sources concurrently.

7.5.2 WFS driving signals computation

Driving signals are computed in time domain after taking inverse Fourier transform and discarding first invalid M samples from output at stage (a). The current M pre-filtered samples are then merged with previous 2,048 samples to form pre-filtered source buffers (to access delayed samples of source signals). As mentioned in Section 7.4, driving signals computation is further divided into two stages to extract the

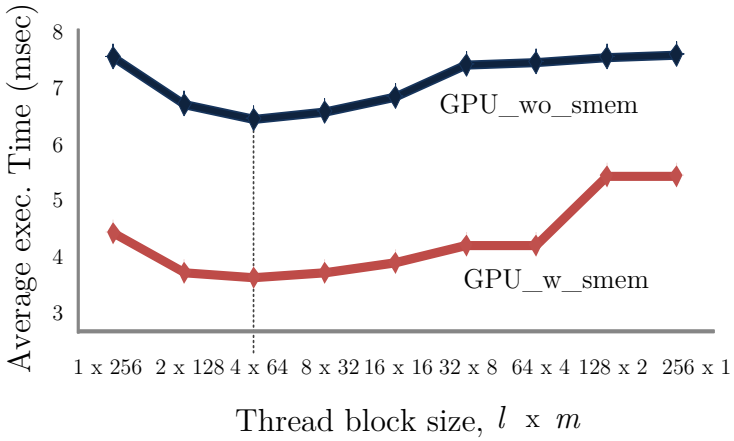


Figure 7.6: CUDA kernel execution times for different TB configurations ($Ns = 100$, $L = 161$, $M = 512$, $Ls = 1$;$l \times m = 256$) : Stage (b)

maximum data parallelism. First, individual driving signals are computed as three-dimensional output matrix of size $(L \times M \times Ns)$. Kernel is launched with $L \times M \times Ns$ threads with two dimensional thread blocks of size $l \times m$ threads, corresponding to m samples of l driving signals. A thread block computes these $l \times m$ samples with each thread computing one sample of the corresponding driving signal. Two dimensional thread block configuration is chosen as driving signal can be computed independently for each virtual source. Weight and delay values are computed once for each speaker position and are reused using shared memory within a TB. Since, each thread need to access previous and current pre-filtered source samples, pre-filtered source buffer is also transferred to shared memory and shared across a thread block to further reduce the global memory latency. The processing stage (b) for individual driving signals for each virtual source is shown in Figure 7.5 along with the 3D driving signal kernel matrix of size $(L \times M \times Ns)$.

Optimum TB configuration $(l \times m)$ **:** Optimum thread block configuration should be chosen such that to maximize the global memory access as coalesced as possible. Let us consider the above processing stage (b) of individual driving signals computation.. The input to this stage is the 2D audio source data

of size $(M + 2048) \times Ns$, where 2048 is the previous samples of pre-filtered source. Output to this stage is a 3D individual driving signal matrix of size $(L \times M \times Ns)$, which is also the size of CUDA kernel as shown in Figure 7.5. For computation of each sample of driving signal, we would need 8 bytes of audio source data to be read and 8 bytes of loudspeaker signal data to be written in addition to other intermediate data like loudspeaker positions, weight and delay values, which approximately equals to 36 bytes per thread. But if we recall, for maximum TB of size 1024 threads, we can only have 48 bytes of data for storage. Thererefore, to be on safer side, we choose TB of size 256 threads so as to allow other TBs to use the shared memory as well during thread synchronization. Since, each l speaker in a TB uses same source data for computation of driving signals, it is obvious that a TB comprising few samples and few speakers can give us higher throughput by employing the sharing of pre-filtered audio data. For our case, $l \times m = 256$ and we wish to determine optimum value of l and m, which will give us the minimum execution time of CUDA kernel for driving signals computation. We consider two extreme cases of $l \times m$:

1) $l \times m = 1 \times 256$ (i.e., 1 speaker $\times$ 256 samples): Since, there is only speaker in the thread block, for all the 256 samples of the driving signals constant delay is to be used. Thus, Memory read can be completed quickly in two clock cycles ($32 \times 8 = 256$ bytes for each warp to be read as 128 byte requests) as filtered source data to be accessed are located in in continuous memory locations. For memory write, since there is only sample or 8 bytes of data to be written, coalesced memory access is not maximized.

2) $l \times m = 256 \times 1$ (i.e., 256 speakers $\times$ 1 sample): Since there are 256 speakers and only one samples to be computed for each of them, there are different delay values needed for all 256 threads (or speakers). In this case for for each 256 speakers a sample is computed. Since the delay values are different,

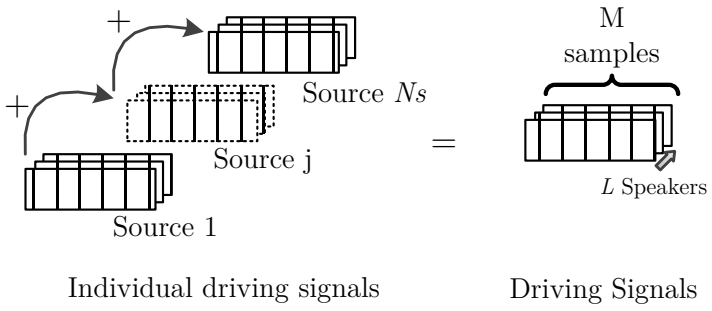


Figure 7.7: Reduction sum for computation of driving signals : Stage (c)

memory read can be sparse and thus, memory read now may take multiple cycles. However, memory write in this case, can be completed in just two cycles as memory address to be written are also contiguous.

Both the extreme cases are not able to maximize the coalesced memory accesses and therefore, optimum TB configuration will be somewhere in between the two extreme cases, which will optimize the memory transactions per warp. Result for execution times of different TB configurations tried is shown in Figure 7.6 along with the impact of shared memory optimization. As shown, there is a trade-off between choices of l and m, with optimum thread block configuration found to be 4×64 for a fixed block size of 256 threads. At the two extremes, execution time of kernel is more than that of the in-between choices of l and m. Similarly, optimum thread block sizes for other kernels have also been found. Furthermore, shared memory usage in stage (b) results in reduction in execution time of CUDA kernel by half as against one without shared memory usage.

Finally, driving signals for each loudspeaker are computed by summing all the individual driving signals for all the sources using reduction sum resulting in 2D output driving signal matrix of size of $L \times M$ as shown in Figure 7.7. Since reduction sum is a sequential operation, kernel with $L \times M$ threads will result in very low throughput with each thread performing Ns serial additions. Furthermore, reduction sum for driving signals for all the source have high data dependency. We

parallelize the reduction sum using binary tree based parallel reduction [209], where partial sums are computed in parallel and synchronized within thread block. Therefore, a separate CUDA kernel is launched with $(L \times M \times Ns)$ threads with one dimensional thread block of size Ns. Each thread block computes one sample of a driving signal using parallel reduction and result is written back to global memory.

7.5.3 WFS synthesized binaural signals computation

Similar to the driving signals computation, processing blocks PB2 and PB3 are implemented in GPU by launching two separate kernels, one for computations of weighted and delayed driving signals and other for the parallel reduction sum (see subsection 7.5.2). First kernel is launched with$L \times M \times 2Ls$ threads and $L \times dimx \times dimz$ threads respectively, for PB2 and PB3. For second kernel, each thread block (of size L threads) computes one sample of synthesized signal at a given listener position using parallel reduction sum for both the processing blocks.

7.6 Simulation Results

Our processing platform consists of the Intel quad core i7 processor as CPU, and Fermi architecture based C2075 as 448-core GPU with 14 streaming multiprocessors (SMs). We analyze the performance of the real-time WFS setup based on implementation aspects, like latency and throughput of the system as well as algorithmic complexity. It should be noted that the CPU implementation inherently takes advantage of the multicore host architecture and multithreading by MATLAB inbuilt functions.

Number of speakers and sound sources are the two main parameters, which control the real-time performance of the WFS driving function block both in terms of efficiency and reproduced sound field quality. In order to create a realistic and practiced WFS system, multiple sources rendering over huge loudspeaker array is

Table 7.2: Average execution times (msec) for WFS processing blocks ($Ns = 1$, $L = 161$, $M = 512$, $Ls = 1$, $dimx = dimz = 256$)

Platform	PB1	PB2	PB3
CPU	1.94	2.08	745.5
CPU+GPU	1.56	0.37	2.7

required. But, a real-time implementation poses constraints on the number of loudspeakers and sources, and often there is a trade-off between performance and behavior of the system. Fewer loudspeakers can result in spatial aliasing and smaller listening area, while limited number of virtual sources may not give an enriching sound experience to the listeners. Since modern GPUs are capable of running thousands of threads in parallel by exploiting massive data parallelism inherent in an application, real-time performance can be improved significantly, while at the same time achieving desired sound field quality.

Table 2 shows the average execution time per frame for the three processing blocks rendering a single source. Execution times reported for GPU is inclusive of the data transfer between host and device. After normalizing the reported execution time by the number of samples processed for each block, PB3 is clearly identified as the slowest block ($PB1 = 0.97$, $PB2 = 0.52$ and $PB3 = 2.91$ msec per sample processed) before GPU optimization. It executes 275 times fast after GPU optimization, which is mainly due to the inherent massive data parallelism involved in the computation of synthesized signals for 256×256 listening points. On the other hand, PB1 is the slowest block after GPU optimization given the lack of much parallelism in single source rendering and overall execution time is dominated by the data transfer latency. It should also be noted that computational complexity of PB1 is directly dependent on number of source signals as against PB2 and PB3, where it is mainly dependent on the number of listener positions and number of loudspeakers. GPU efficiency of the system can be considerably improved when there are many sources to be rendered by extracting more data parallelism. However, increasing the workload on GPU will also incur high global memory overhead. As discussed in Section 7.5,

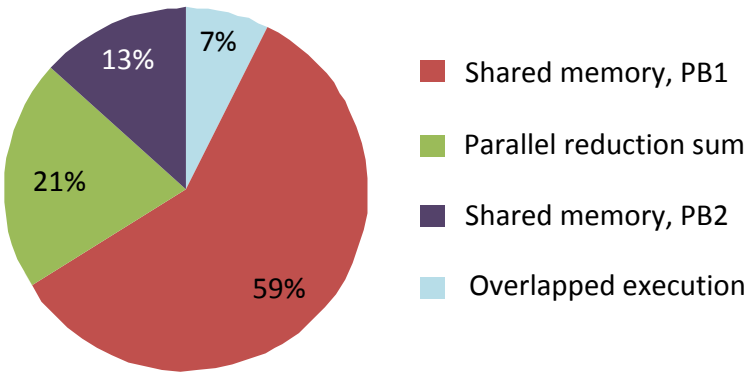
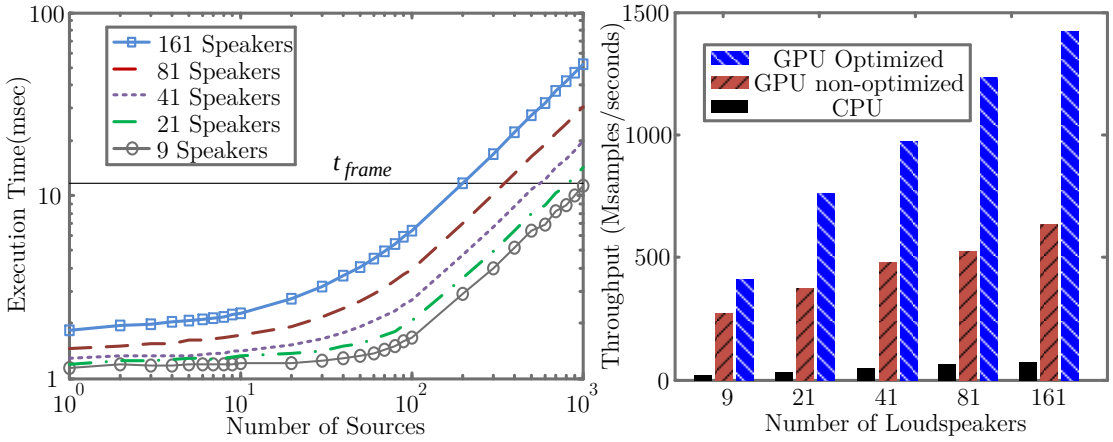


Figure 7.8: Percentage improvement in execution times due to different GPU optimization techniques over GPU non-optimized implementation for PB1+PB2 ($Ns = 100, L = 161, M = 512, Ls = 1$)

several optimization techniques can be applied to speed up the system. Figure 7.8 shows the impact of major optimizations on system performance. As shown in Figure 7.8, shared memory with optimum thread block size has most of the impact in improving the GPU performance with 59% share for the driving function block PB1. However, for block PB2, there is only 13% improvement over non-optimized GPU implementation. This is mainly due to the fewer data parallelism present in PB2 as compared to PB1 when there are multiple sources to rendered. Another significant effect is due to the parallel reduction sum especially, if CUDA kernel is launched with hundreds of threads in case of driving signals computation with many sources. Finally, overlapped execution, which was used to perform some of the data transfers and host processes simultaneously with kernel execution, also resulted in 7% latency savings.

Figure 7.9 shows the average execution times per frame and peak throughput of the overall system for blocks PB1 and PB2. Processing time for the system must be less than t_{frame} for WFS setup to perform in real-time. Thus, GPU can render up to 1,000 real-time sources for 9 speakers or 200 real-time sources for 161 speakers. At the same time one can also listen to the synthesized binaural signals for given listener positions in real-time. System throughput is calculated as



(a) Average execution times per frame

(b) Peak system throughput

Figure 7.9: Average Execution times and Peak throughput of overall system (PB1 + PB2) ($M = 512$, $Ls = 1$)

number of sampled processed per unit time. As shown in Figure 7.9, optimized GPU implementation can have peak system throughput of 1,400 Mega samples per second (MSPS) almost twice of the non-optimized GPU implementation and 20 times of the CPU based implementation. It is also important to note that throughput can be substantially improved by increasing the GPU workload and hence, exploiting the maximum data parallelism in GPU.

7.7 Conclusions

In this chapter, we presented a fast and efficient real-time GPU implementation for a wave field synthesis system. We have successfully accelerated the overall WFS setup with peak throughput of 1,400 MSPS using several GPU optimization techniques. 20-fold improvement was achieved over CPU based implementation, while up to 200 sources can be rendered in real-time using array with hundreds of loudspeakers. In addition, for high-end GPUs running thousands of parallel threads, we are able to synthesize WFS signals at any listener positions in real-time with as many sources. One of the main features of this work is that all the audio processing is done in GPU, while CPU is freed for other independent tasks like I/O buffering or overlapped data

transfers. Among the several optimization techniques, shared memory provides us the most significant performance improvement. The GPU implementation presented in this work can easily be integrated with the hybrid WFS setup for a densely spaced virtual speaker array rendering. As shown in this work, both processing blocks PB1 and PB2 can be carried out in real-time, which implies that virtual WFS rendering for rear and side auditory processing of the hybrid WFS setup can be realized in practice seamlessly with the frontal physical array playback. Furthermore, the proposed GPU implementation can easily be scaled to any number of speakers and sources only to be limited by the number of CUDA cores and real-time constraint of the application.

Chapter 8

Conclusions and Future Works

8.1 Conclusions

In this thesis, we studied the challenges involved in natural listening over both headphones and loudspeakers. Therefore, contributions of this thesis is categorized into two major works:

1) Natural listening over open headphones for augmented reality applications using adaptive filtering techniques and their extensions for practical scenarios.

2) A next-generation entertainment system using a hybrid speaker array-headphones system for use in home scenarios.

In the first part of the thesis, we investigated natural listening over headphones in AR based applications. Natural listening in AR based scenarios requires user to be constantly aware of surroundings with both virtual and physical sound sources present in the real world. However in virtual reality applications, the listener is assumed to be teleported to a virtual auditory environment and is disconnected with the physical surroundings. Therefore, it is essential that virtual objects should sound alike real sources and must give the feeling of *"being there"* in the real auditory environment. Individualized HRTFs and listener environment's transfer functions must be used in order to create an illusion that virtual sources are coming from physical environments. Furthermore, non-flat spectrum of the conventional head-

phones alters the intended sound at listener's ears. HPTF, which represents the headphones' emitter responses as well as reflections with earcup and pinnae, are unique to every individual and individual headphone equalization must be applied to the individualized BRIRs.

We introduced a natural augmented reality (NAR) headset, which employs adaptive filtering techniques for individual headphone equalization for AR applications (Chapter 4). The proposed NAR headset has two pairs of microphones with two microphones (one internal and external) on each side of the ear cup. Internal microphone was attached to the NAR headset such that it was positioned just below ear canal opening, while external microphone was attached just outside the ear cup. With the help of these sensing microphones, individualized responses in the listener environment can be measured readily. Furthermore, open headphones were chosen for the prototype such as to allow external sounds to be heard without much attenuation. For virtual source reproduction via binaural synthesis, individual headphone equalization is applied using adaptive algorithms to compensate for the HPTFs. A hybrid adaptive equalizer (HAE) based on the combination of conventional FxNLMS and modified FxNLMS is introduced to facilitate fast convergence and optimum SS-MSE for the binaurally synthesized signals. Modified FxNLMS is proposed with additional spatial filter introduced in the secondary path to improve the convergence rate as against conventional FxNLMS. However, it was observed that the modified FxNLMS does not adapt entirely to the desired response in high frequencies for some of the source positions, whereas conventional FxNLMS suffers from higher SD in low frequencies. Combining the two approaches helps in converging to the desired sound quickly and optimally. Using simulation with stationary white noise signals, SS-MSE reduction of more than 25 dB were observed with SD values within the perceptually tolerance limit. This implies that virtual sound is reproduced perceptually similar to the direct natural listening. Additionally, the proposed adaptive equalization technique should also work in the presence of environment/physical world

sounds. It was found that HAE method results in lower SS-MSE in the presence of external sounds, implying virtual sounds are not synthesized close to real sounds and might lead to unnaturalness. Therefore, HAE approach is further extended for the augmented reality mode, where both virtual and external sounds are present such that external sounds do not interfere with the convergence process of HAE. Adaptive estimation of external sound signals using the signals received at external microphones can be used to remove their effect from the equalization process. A listening test was conducted using individualized BRIRs to evaluate the naturalness of the virtual sounds and perceptual similarities between the virtual and real sources. Listening test result shows very high source confusion % for the case, where subjects identified both virtual and real sources as real implying virtual sounds were perceived quite realistic and illusioned with the real sounds. Subjects could not differentiate between real and virtual sounds and their positions in 3D space were also in very close vicinity. Moreover, perceptual similarity between real and virtual sound sources further increased the realism in an augmented scenario with both real and virtual sources present.

We further addressed some of the major practical limitations of the NAR headset in Chapter 5. NAR headset requires individualized impulse responses for virtually synthesized signals to be as close as possible to real sounds. For AR based scenarios, BRIRs must be measured in listener physical environment, it is imperative that NAR headset should be able to acquire listener's transfer function in a convenient, quick and efficient manner. With the help of a head-tracker attached on NAR headset and using continuous acquisition method based on NLMS identification method, it was shown that individualized BRIRs can be measured promptly as user moves his head. However, in this approach one needs a loudspeaker continuously playing a perfect sine sweep signal for the duration of measurement. It needs to be investigated if smartphone can be used to playback and record the perfect sine sweep signal response as well as estimate the BRIRs in real-time. Furthermore, we proposed the

following three extensions of the NAR headset for practical use:

1) Adaptive equalization of NAR headset using non-stationary virtual sounds like speech, music, etc.: In practice, virtual sounds are much different from white noise and are non-stationary in nature. NAR headset should be able to work equally well for non-stationary signals. But, the transient nature of these signals hampers the performance of the proposed HAE method. We extended the HAE approach by introducing a training phase using white noise signal and a real-time playback phase using any virtual signals. The training phase is exactly the same as HAE method, while the playback phase uses a copy of the equalized filter derived from training phase in addition to an adaptive compensation filter for virtual signals. Results from simulations using speech signals show that extension works well with good convergence.

2) Fast detection of any large changes in physical secondary path response and fast estimation of HPTF: Although, the HAE approach is robust to small mismatches between secondary path model ($\hat{h}_{hp}(n)$) and its physical counterpart ($h_{hp}(n)$), any large changes in HPTF can result in lower MSE and thus, must be compensated. HPTF detection method is introduced based on running power estimate to quickly detect any large variation in HPTF due to either headset repositioning or refitting. Once it is detected that physical secondary path response is significantly different from its model, playback of virtual signals is stopped and HPTF estimation begins. HPTF estimation is carried out using a probe signal sent out through headset and its response recorded at internal microphone position. Perfect sine sweep signal is used as probe signal and NLMS algorithm used for secondary path identification. It was found that HPTF can be estimated quickly within 100 millisecond because of the very short length of physical secondary path.

3) Resolving causality issue in adaptive estimation of external signals for HAE

method: We extended the adaptive equation for non-stationary virtual signals with online adaptive estimation of real signals. In practice, environment sounds can come from any direction and consequently, adaptive estimation suffers from causality when external sound reaches internal microphone earlier than the external microphone position. It was observed that performance of adaptive estimation degrades significantly, especially for the contralateral ear, when there is longest difference between arrival time at two ears. This is resolved by adding a forward delay in error signal path and feedback path of the adaptive compensation of HPTF, which is now referred as delayed FxLMS (FxDLMS). Using simulation results, it was found that an appropriate step-size needs to be chosen with trade-off between convergence speed and minimum MSE.

In the second part of this thesis, we presented a hybrid WFS setup combining WFS and binaural synthesis over NAR headset to provide listener an immersive listening experience for home scenarios (Chapter 6). Physical frontal array was used for the frontal playback using WFS, while rear and side auditory scene was virtually synthesized at listener's ears through the NAR headset using virtual WFS method. An enclosed WFS array setup is proposed with rear and side speakers acting as virtual speakers rendered over the NAR headset. WFS renderer is used as the processing core to compute all the driving signals and therefore, can provide seamless integration of physical WFS with the virtual WFS over the NAR headset. For virtual WFS, a multichannel version of the adaptive equalization for NAR headset is presented to compute the equalized speaker BRIR filters for headphones playback. Simulation results using dummy head measurements show that temporal and spectral characteristics of physical WFS sound field were retained in the virtually synthesized WFS sound field. Although the NAR headset is open-back, it attenuates sound signal at high frequency above 1.5 kHz by 10-15 dB. Headphone isolation can result in dullness of the sound and hamper the frontal playback performance especially,

it being crucial for home entertainment applications. Therefore, we presented two methods to compensate for the high frequency attenuation. First, a headphone isolation compensation approach was introduced, where WFS driving signals were convoluted with compensation filters, summed together and played back over the NAR headset along with the physical array rendered sound. The compensation filters account for the high frequency attenuation and thus, compensate for it when played together with the frontal WFS array playback. In the second method, we also aim to minimize the spatial aliasing artifacts of the physical array by reproducing only aliasing free high frequency components of equalized driving signals over NAR headset, while low frequency components were rendered over the physical frontal array. In this way, the entire playback spectra is almost free of sound coloration. It was assumed here that low frequency reproduction over physical speakers ensures better frontal perception as ITD is predominant for localization below 1.5 kHz [210]. However, high frequency spectral cues also help in sound localization and it remains to be seen how localization is affected by reproducing high frequency components over the NAR headset.

A detailed subjective study was conducted to investigate the performance of proposed hybrid and virtual WFS methods for frontal playback and virtual WFS methods for rear and side auditory scene playback in terms of localization, sound coloration and overall audio quality. In the localization task, subjects were asked to tell the direction, externalization, elevation level and locatedness of the virtual source. For frontal virtual sources, it was found that hybrid WFS methods had better directional accuracy as compared to the virtual WFS methods, validating our assumption that the reproduced low frequency over the physical array indeed provides better frontal perception. Physical WFS array playback resulted in best azimuth accuracy but higher standard deviations were observed. This is possibly because of the high frequency headphone isolation, resulting in increased sound image width. For rear and side sources, virtual WFS method with aliasing free

high frequency components performs better than that of frontal directions with good directional accuracy. Virtual sources were found to be well externalized in most of the cases. However, for frontal playback, physical WFS and hybrid WFS playback with minimum spatial aliasing were found to have significantly higher externalization grades than other methods. Among other methods, virtual WFS method with minimum spatial aliasing had slightly higher externalization than rest of the virtual and hybrid playback methods with aliased sound field. For rear and side sources, similar trend was observed for the two virtual WFS methods. All the virtual sources were perceived to be between ear level and just above ear level as reported by most of the subjects. However, few subjects reported that virtual sources, especially in rear directions, were observed to be very near to head and much above ear level. This was more predominant for virtual WFS method with spatial aliasing present in high frequencies. Spectral peaks and notches in the high frequency might disrupt the spectral cues that are important for localization. Locatedness of Hybrid WFS method with very high aliasing frequency was observed to be highest for frontal sources, although only slightly better than rest of the virtual and hybrid methods. For rear and side sources, locatedness for virtual WFS method was only slightly better when aliasing frequency increased from 2 kHz to 18 kHz. However, in the perceptual study conducted by Wittek [101, 104], locatedness of the WFS array degraded significantly when aliasing frequency decreased from 7.5 kHz to 2.5 kHz.

In the second listening test, sound coloration of different hybrid and virtual WFS methods were evaluated as compared to point source as reference sound. It was also investigated how the sound coloration varies as cross-over frequency (f_c) of the hybrid WFS method is increased from 1.5 kHz to 6 kHz. Hybrid WFS method with cross-over frequency of 1.5 kHz was observed with minimum sound coloration grades. As the cross-over frequency increased, increased spatial aliasing resulted in severe sound coloration. Sound coloration of virtual WFS method was found to be better than hybrid WFS with similar frequency spectra for off-center source

position. In the third and final listening test, overall audio quality was evaluated using a drum beats stimuli with spectrum similar to that of pink noise. Subjects were asked to judge the overall audio quality in terms of frontal image quality and timbre clarity. Any timbre distortion in high frequencies could result in degradation of overall audio quality. In this task, physical array playback was chosen as reference. Since NAR headset would attenuate the high frequency sound, reference sound was assumed to have poor timbre clarity. It was interesting to find that hybrid WFS with minimum sound coloration was found to be the best overall audio quality followed by the virtual WFS method with similar spectra. Reproduction method with aliasing characteristics similar to that of physical WFS array were graded between slightly worse to slightly better. To summarize, the proposed hybrid WFS setup can be reliably used in home scenarios with entire 360^o sound experience and better frontal playback performance when accompanied with high frequency aliasing-free reproduction over the NAR headset.

The hybrid WFS setup comes at the cost of high computational complexity as WFS renderer needs to compute driving signals for an enclosed setup comprising of hundreds of loudspeakers. WFS driving signals were binaurally synthesized at listener's ears using the virtual WFS technique and thereby, further increasing computational complexity. With the advent of modern graphics processing units (GPU) like GTX590, C2075, K10 etc. and hundreds to thousands of processing cores, massively parallel and computationally intensive applications, such as WFS can be optimized with significant throughput improvements. In Chapter 7, we presented a fast and efficient real-time GPU implementation for a WFS system comprising of both driving signals computation as well as binaural signals synthesis using virtual WFS. GPU optimization methods like shared memory, overlapped execution, data reorganization, parallel reduction sum, were used to successfully accelerate the WFS system. Peak throughput of 1,400 mega samples per second (MSPS) was achieved using several GPU optimization techniques. 20-fold improvement was achieved over

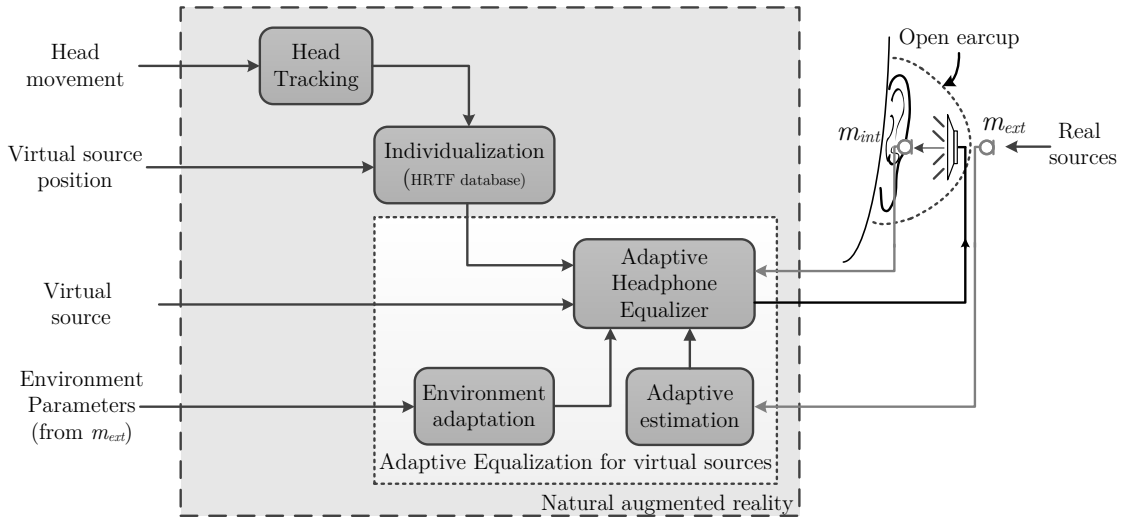


Figure 8.1: Natural listening using NAR headset

CPU based implementation, while up to 200 sources can be rendered in real-time using the array with hundreds of loudspeakers. Shared memory provides the most significant performance improvement among the different optimization techniques applied. Furthermore, we are able to synthesize binaural signals at any listener positions in real-time with as many sources. One of the main features of this work is that all the audio processing is done in GPU using low level CUDA programming, while MATLAB running on CPU as host is free for other independent tasks like I/O buffering or overlapped data transfers.

8.2 Future Works

With the advent of AR wearable devices like Microsoft Hololens [211], Recon Jet [212], Sony SmartEyeGlass, [213] it is becoming more important to provide natural listening experience to the user so that one can connect with what they are seeing through the devices. In this thesis, a conscious effort has been made to provide listener with a realistic experience via headphones in augmented reality scenarios. With today's embedded processors operating at hundredth of MHz, its processing power is more than sufficient to handle the computational complexity

of the proposed algorithms. The proposed NAR headset can be further refined so as to integrate it into conventional wearable devices as shown in Figure 8.1. NAR headset currently works best when individualized BRTFs of the user are available. Individualization techniques can be added to the NAR headset so as to enhance the listening experience when using generic/non-individualized HRTFs. Out of the existing individualization techniques, one simple method is subjective tuning from an extensive set of per-measured HRTFs database. Sunder et al. [142] proposed an individualization technique using frontal projection headphones as the frontal emitter inherently contains the frontal pinna cues. With the NAR headset, similar methods can be applied for individualization. We can take advantage of the fact that transfer functions measured using NAR headset at external microphones does not contain any pinnae cues and can be used for virtual sound synthesis in conjunction with inverse of free-field headphone ear-cup response. In this context, the position of the microphones are critical and need to be investigated firther. Furthermore, constant adaption of environment characteristics are required when listener moves from one place to another. The two external microphones can be used to capture surroundings sound and extract important environment parameters. Head tracking should also be added into the NAR headset so that virtual sound objects can be adjusted in physical space according to the head orientation as well listener translation movements. NAR headset can also be extended for noise canceling mode when surroundings are too loud and external sounds must be suppressed by generating an anti-noise signal through the NAR headset.

With the arrival of HDTV, 3D TV and curved TV, users at home are now more engrossed in watching movies, sports or playing games. However, spatial 3D audio system is still not fully viable in the home entertainment systems, which can be integrated seamlessly with the 3D video content to fully immerse viewers in a 3D audio-visual experience. There is a great research potential in this direction to make the user experience more immersive by combining 3D audio with visuals. The

hybrid WFS setup proposed in this work aims to fulfill this objective by providing an immersive audio experience. However, for such setups to be practically realizable in TV, audio contents must be recorded and spatially encoded so that it can be rendered over any flexible playback system. Recently, we are seeing standardization of 3D audio content via MPEG-H with support added for audio objects based encoding and transmission, which are apt for speaker array playback systems like WFS. It needs to be further investigated if the proposed hybrid WFS setup can be readily applied in different multimedia applications. The current setup can also be further extended to multiple listeners especially, for rear and side auditory scenes playback with no physical speakers present.

Author's Publications

[a] R. Ranjan and W. S. Gan, "Natural Listening over Headphones in Augmented Reality Using Adaptive Filtering Techniques," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1988-2002, Nov. 2015.

[b] R. Ranjan and W.S. Gan, "A hybrid speaker array-headphone system for immersive 3D audio reproduction," *in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr 2015, pp. 1836-1840.

[c] R. Ranjan, W.S. Gan, and C. Yong-Kim, "Applying Active Noise Control Technique for Augmented Reality Headphones," *in Proceedings of Internoise*, Melbourne, Nov 2014.

[d] R. Ranjan and W.S. Gan, "Fast and efficient real-time GPU based implementation of wave field synthesis," *in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 7550-7554.

[e] R. Ranjan and W.S. Gan, "Wave Field Synthesis: The Future of Spatial Audio,"*IEEE Potentials*, vol. 32, pp. 17-23, Mar 2013.

[f] R. Ranjan and W. S. Gan, "On the use of Dynamically Varied Loudspeaker Spacing in Wave Field Synthesis," *in 133rd Audio Engineering Society Convention*, San Francisco, USA, Oct 2012.

Bibliography

[1] R. Ranjan and W. Gan, "Wave Field Synthesis: The Future of Spatial Audio," *IEEE Potentials*, vol. 32, no. 2, pp. 17–23, Mar 2013.

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

[3] D. Hammershøi and H. Møller, "Binaural technique—Basic methods for recording, synthesis, and reproduction," in *Communication Acoustics*. Springer, 2005, pp. 223–254.

[4] Merchel, S., Franco, A. F., Pesqueux, L., Rouaud, M., and Soerensen, M. O., "Sound Reproduction by Wave Field Synthesis," Aalborg University, Project Report, 2004.

[5] M. N. Montag, "Wave Field Synthesis In Three Dimensions by Multiple Line Arrays," Master's thesis, University of Miami, Florida, 2011.

[6] C. Antweiler, "Multi-Channel System Identification with Perfect Sequences," *Advances in Digital Speech Transmission*, p. 171, 2008.

[7] S. Hong and H. Kim, "An analytical model for a GPU architecture with memory-level and thread-level parallelism awareness," in *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3, New York, NY, USA, June 2009, pp. 152–163.

[8] F. Rumsey, *Spatial Audio*. CRC Press, Sep 2012.

[9] E. Berdahl, D. Harris, G. Niemeyer, and J. Smith III, "An electroacoustic sound transmission system that is stable in any (dissipative) acoustic environment: An application of sound portholes," *Proceedings of the NOISE-CON*, Apr 2010.

[10] W. Orfali, "Room Acoustic and Modern Electro-Acoustic Sound System Design during Constructing and Reconstructing Mosques," Ph.D. dissertation, Technischen Universitat Berlin, 2007.

[11] P. Fellgett, "Ambisonics. Part One: General system description," *Studio Sound*, vol. 17, pp. 20–22, Aug 1975.

[12] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *Journal of Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, Feb 1973.

[13] A. J. Berkhout, "A Holographic Approach to Acoustic Control," *Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, Dec 1988.

[14] A. D. Blumlein, "British Patent Specification 394,325 (Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems)," *Journal of the Audio Engineering Society*, vol. 6, no. 2, pp. 91–130, Apr 1958.

[15] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, June 1997.

[16] R. Nicol, *Binaural Technology*. Audio Engineering Society MonoGraph, 2010.

[17] T. Cashion and S. Williams, "Apparatus for creating 3D audio imaging over headphones using binaural synthesis," U.S. Patent US5 809 149 A, Sep, 1998. [Online]. Available: http://www.google.com/patents/US5809149

[18] P. H. Myers, "Three-dimensional auditory display apparatus and method utilizing enhanced bionic emulation of human binaural sound localization," United States Patent US4 817 149 A, Mar., 1989.

[19] C. B. Jensen and others, "Binaural synthesis, head-related transfer functions, and uses thereof," Patent US Patent 6,118,875, Sep, 2000.

[20] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Proceedings of the 107th Audio Engineering Society Convention*, New York, NY, USA, Sep 1999.

[21] L. Rayleigh, "On our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[22] W. G. Gardner, *3-D Audio Using Loudspeakers*. Springer Science & Business Media, Apr. 1998.

[23] M. J. Evans, J. A. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2400–2411, Oct 1998.

[24] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights into head-related transfer function: Spatial dimensionality and continuous representation," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2347–2357, Apr 2010.

[25] W. L. Martens, *Principal components analysis and resynthesis of spectral cues to perceived direction*. Ann Arbor, MI: MPublishing, University of Michigan Library, 1987.

[26] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Modal expansion of HRTFs: Continuous representation in frequency-range-angle," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr 2009.

[27] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Proceedings of the 114th Audio Engineering Society Convention*, Amsterdam, The Netherlands, Mar 2003.

[28] J. Daniel and S. Moreau, "Further study of sound field coding with higher order ambisonics," in *Proceedings of the 116th Audio Engineering Society Convention*, Berllin, Germany, Mar 2004.

[29] J. Ahrens and S. Spors, "Analytical driving functions for higher order ambisonics," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Las Vegas, USA, Mar 2008, pp. 373–376.

[30] Y. J. Wu and T. D. Abhayapala, "Theory and design of soundfield reproduction using continuous loudspeaker concept," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 107–116, Jan 2009.

[31] P. N. Samarasinghe, M. A. Poletti, S. A. Salehin, T. D. Abhayapala, and F. M. Fazi, "3d soundfield reproduction using higher order loudspeakers," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Vancouver, BC, Canada, May 2013.

[32] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 697–707, Sep 2001.

[33] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field re-

production in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, Apr 2005.

[34] M. Poletti, "Robust two-dimensional surround sound reproduction for nonuniform loudspeaker layouts," *Journal of the Audio Engineering Society*, vol. 55, no. 7/8, pp. 598–610, July 2007.

[35] F. Fazi, P. Nelson, J. E. Christensen, and J. Seo, "Surround system based on three-dimensional sound field reconstruction," in *Proceedings of the 125th Audio Engineering Society Convention*, San Francisco, CA, USA, Oct 2008.

[36] Y. J. Wu and T. D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1711–1720, Aug 2011.

[37] S. Bertet, J. Daniel, and S. Moreau, "3d sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone," in *Proceedings of the 120th Audio Engineering Society Convention*, Budapest, Hungary, Aug 2006.

[38] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, Nov 2005.

[39] M. Poletti, "A spherical harmonic approach to 3d surround sound systems," in *Proceedings of the Forum Acousticum*, 2005, pp. 311–317.

[40] A. Gupta and T. D. Abhayapala, "Three-dimensional sound field reproduction using multiple circular loudspeaker arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1149–1159, July 2011.

[41] W. Zhang and T. D. Abhayapala, "Three dimensional sound field reproduction using multiple circular loudspeaker arrays: functional analysis guided

approach," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1184–1194, July 2014.

[42] S. Spors and J. Ahrens, "A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling," in *Proceedings of the 125th Audio Engineering Society Convention*, San Francisco, CA, USA, Oct 2008.

[43] S. Spors, "Comparison of wave field synthesis and higher-order ambisonics," in *Ambisonics Symposium*, Graz, Austria, June 2009.

[44] J. Ahrens, H. Wierstorf, and S. Spors, "Comparison of Higher Order Ambisonics and Wave Field Synthesis with respect to spatial discretization artifacts in time domain," in *Proceedings of the 40th Audio Engineering Society Conference*, Tokyo, Japan, Oct 2010.

[45] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix-a suggested ambisonics format," in *Proceedings of the 3rd Ambisonics Symposium*, Lexington, KY, June 2011.

[46] H. Pomberger, F. Zotter, and A. Sontacchi, "An ambisonics format for flexible playback layouts," in *Proceedings of the Ambisonics Symposium*, Graz, Austria, June 2009.

[47] M. Kronlachner, "Ambisonics plug-in suite for production and performance usage," in *Linux Audio Conference*, Graz, Austria, May 2013.

[48] D. De Vries, "Wave Field Synthesis," *AES Monograph. New York: Audio Engineering Society*, 2009.

[49] G. Enzner, C. Antweiler, and S. Spors, "Trends in Acquisition of Individual Head-Related Transfer Functions," in *The Technology of Binaural Listening*, 2013, pp. 57–92.

[50] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA*, New York, US, Oct. 2009, pp. 325–328.

[51] J. Blauert, "An introduction to binaural technology," in *Proceedings of the Binaural and spatial hearing in real and virtual environments*. American Psychological Association, 1997, pp. 593–609.

[52] J. C. Middlebrooks, J. C. Makous, and D. M. Green, "Directional sensitivity of sound-pressure levels in the human ear canal," *The Journal of the Acoustical Society of America*, vol. 86, no. 1, pp. 89–108, July 1989.

[53] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 921–930, Aug 2015.

[54] M. Zhang, R. A. Kennedy, and T. D. Abhayapala, "Empirical determination of frequency representation in spherical harmonics-based hrtf functional modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 351–360, Feb 2015.

[55] W. Zhang *et al.*, "Measurement and modelling of head-related transfer function for spatial audio synthesis," Ph.D. dissertation, 2013.

[56] A. W. Mills, "Auditory localization(Binaural acoustic field sampling, head movement and echo effect in auditory localization of sound sources position, distance and orientation)," *Foundations of modern auditory theory.*, vol. 2, pp. 303–348, 1972.

[57] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S.

Colburn, and E. M. Wenzel, "On the Externalization of Auditory Images," *Presence: Teleoper. Virtual Environ.*, vol. 1, no. 2, pp. 251–257, May 1992.

[58] V. Larcher, J.-M. Jot, P. I. Stravinsky, and F. Paris, "Techniques d'interpolation de filtres audio-numériques, Application à la reproduction spatiale des sons sur écouteurs," in *Proceedings of the French Society of Acoustics*, Apr. 1997.

[59] Véronique Larcher, "Techniques de spatialisation des sons pour la ralit virtuelle," PhD Thesis, University of paris, 2001.

[60] D. Satongar, C. Pike, Y. W. Lam, and T. Tew, "On the Influence of Headphones on Localization of Loudspeaker Sources," in *Proceedings of the 135th Audio Engineering Society Convention*, New York, NY, USA, Oct 2013.

[61] P. L. Søndergaard, J. F. Culling, T. Dau, N. Le Goff, M. L. Jepsen, P. Majdak, and H. Wierstorf, "Towards a binaural modelling toolbox," in *Proceedings of Forum Acusticum*, Aalborg, Denmark, June 2011.

[62] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *the Journal of The Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, Sep. 1999.

[63] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, May 1999.

[64] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *the Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, July 1993.

[65] J. C. Middlebrooks, "Narrow-band sound localization related to external ear

acoustics," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2607–2624, Nov 1992.

[66] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: effect of pinnae cavity occlusion," *Journal of the Acoustical Society of America*, vol. 53, no. 2, pp. 400–408, Feb 1973.

[67] S. R. Oldfield and S. P. Parker, "Acuity of sound localisation: a topography of auditory space. II. Pinna cues absent," *Perception*, vol. 13, no. 5, pp. 601–617, Oct. 1984.

[68] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *the Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, Dec 1974.

[69] K. H. Shin and Y. Park, "Customization of head-related transfer functions using principal components analysis in the time domain," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3284–3284, Nov 2006.

[70] S. Hwang and Y. Park, "HRIR customization in the median plane via principal components analysis," in *Proceedings of the Audio Engineering Society Conference*, London, UK, June 2007.

[71] 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," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2053–2064, Nov 2002.

[72] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. Davis, "Hrtf personalization using anthropometric measurements," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, NY, USA, Oct 2003, pp. 157–160.

[73] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan, "Psychophysical customization of directional transfer functions for virtual sound localization," *The Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 3088–3091, Dec 2000.

[74] K. McMullen, A. Roginska, and G. H. Wakefield, "Subjective selection of head-related transfer functions (HRTF) based on spectral coloration and interaural time differences (ITD) cues," in *Proceedings of the 133rd Audio Engineering Society Convention*, San Francisco, CA, USA, Oct 2012.

[75] Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoustical science and technology*, vol. 27, no. 6, pp. 340–343, 2006.

[76] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1997.

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

[78] 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.

[79] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 55, no. 7/8, pp. 623–637, July 2007.

[80] M. Pollow, B. Masiero, P. Dietrich, J. Fels, and M. Vorländer, "Fast measurement system for spatially continuous individual hrtfs," York, UK, Mar 2012.

[81] D. Brungart, "Near-Field Virtual Audio Displays," *Presence*, vol. 11, no. 1, pp. 93–106, Feb. 2002.

[82] G. Plenge, "On the differences between localization and lateralization," *The Journal of the Acoustical Society of America*, vol. 56, no. 3, pp. 944–951, Sep 1974.

[83] D. R. Begault, "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems," *Journal of the Audio Engineering Society*, vol. 40, no. 11, pp. 895–904, Nov. 1992.

[84] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?" *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, June 1996.

[85] K. I. McAnally and R. L. Martin, "Sound localization with head movement: implications for 3-d audio displays," *Front Neurosci.*, vol. 8, Aug. 2014.

[86] H. Wallach, "The Role of Head Movements and Vestibular and Visual Cues in Sound Localization," *Journal of Experimental Psychology*, vol. 27, no. 4, p. 339, Oct 1940.

[87] H. Møller, *Fundamentals of Binaural Technology.* Aalborg Universitetscenter, Institut for Elektroniske Systemer, Afdeling for Telekommunikation, 1992.

[88] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[89] P. Vogel, *Application of wave field synthesis in room acoustics.* TU Delft, Delft University of Technology, 1993.

[90] D. de Vries and M. M. Boone, "Wave field synthesis and analysis using array

technology," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, NY, USA, Oct 1999, pp. 15–18.

[91] M. M. Boone, E. N. Verheijen, and P. F. Van Tol, "Spatial sound-field reproduction by wave-field synthesis," *Journal of the Audio Engineering Society*, vol. 43, no. 12, pp. 1003–1012, Dec. 1995.

[92] A. Berkhout, "Wave-front synthesis: A new direction in electroacoustics," *The Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 2396–2396, Oct 1992.

[93] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[94] E. N. G. Verheijen, "Sound reproduction by wave field synthesis," Ph.D. dissertation, TU Delft, Delft University of Technology, 1998.

[95] A. Berkhout and M. Boone, "Application of wave field synthesis in enclosed spaces: new developments," *ACUSTICA*, vol. 82, pp. S218–S218, Jan. 1996.

[96] S. Spors and R. Rabenstein, "Spatial Aliasing Artifacts Produced by Linear and Circular Loudspeaker Arrays used for Wave Field Synthesis," in *Proceedings of the 120th Audio Engineering Society Convention*, Paris, France, May 2006.

[97] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of Wave Field Synthesis Revisited," in *Proceedings of the 124th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 2008.

[98] S. Spors, "Extension of an analytic secondary source selection criterion for wave field synthesis," in *Proceedings of the 123rd Audio Engineering Society Convention*, New York,NY,USA, Oct 2007.

[99] E. W. Start, "Direct sound enhancement by wave field synthesis," PhD Thesis, TU Delft, Delft University of Technology, 1997.

[100] J.-J. Sonke, J. Labeeuw, and D. de Vries, "Variable acoustics by wavefield synthesis: A closer look at amplitude effects," in *Proceedings of the 104th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 1998.

[101] H. Wittek, F. Rumsey, and G. Theile, "Perceptual enhancement of wavefield synthesis by stereophonic means," *Journal of the Audio Engineering Society*, vol. 55, no. 9, pp. 723–751, Sep 2007.

[102] W. P. J. De Bruijn, "Application of wave field synthesis in videoconferencing," Ph.D. dissertation, TU Delft, Delft University of Technology, 2004.

[103] H. Wierstorf, "Perceptual assessment of sound field synthesis," Ph.D. dissertation, Technische Universität Berlin, 2014.

[104] H. Wittek, "Perceptual differences between wavefield synthesis and stereophony," PhD Thesis, University of Surrey, 2007.

[105] E. Corteel, C. Kuhn-Rahloff, and R. Pellegrini, "Wave field synthesis rendering with increased aliasing frequency," in *Proceedings of the 124th Audio Engineering Society Convention*, 124th, Amsterdam, The Netherlands, May 2008.

[106] S. Spors, H. Wierstorf, M. Geier, and J. Ahrens, "Physical and Perceptual Properties of Focused Virtual Sources in Wave Field Synthesis," in *Proceedings of the 124th Audio Engineering Society Convention*, New York, NY, USA, Oct 2009.

[107] R. Oldfield, I. Drumm, and J. Hirst, "The perception of focused sources in

wave field synthesis as a function of listener angle," in *Proceedings of the 128th Audio Engineering Society Convention*, London, UK, May 2010.

[108] M. Geier, H. Wierstorf, J. Ahrens, I. Wechsung, A. Raake, and S. Spors, "Perceptual evaluation of focused sources in wave field synthesis," in *Proceedings of the 128th Audio Engineering Society Convention*, London, UK, May 2010.

[109] D. De Vries, "Sound reinforcement by wave field synthesis: Adaptation of the synthesis operator to the loudspeaker directivity characteristics," *Journal of the Audio Engineering Society*, vol. 44, no. 12, pp. 1120–1131, 1996.

[110] D. de Vries, E. W. Start, and V. G. Valstar, "The Wave-Field Synthesis Concept Applied to Sound Reinforcement Restriction and Solutions," in *Proceedings of the 96th Audio Engineering Society Convention*, Amsterdam, The Netherlands, Feb. 1994.

[111] D. de Vries and P. Vogel, "Experience with a sound enhancement system based on wavefront synthesis," in *Proceedings of the 95th Audio Engineering Society Convention*, New York, NY, USA, Oct. 1993.

[112] D. de Vries and J. Baan, "Auralization of sound fields by wave field synthesis," in *Proceedings of the 106th Audio Engineering Society Convention*, Munich, Germany.

[113] E. Start, D. de Vries, and A. Berkhout, "Wave field synthesis operators for bent line arrays in a 3d space," *Acta Acustica united with Acustica*, vol. 85, no. 6, Nov.

[114] S. Brix, T. Sporer, and J. Plogsties, "CARROUSO-An European approach to 3d-audio," *PREPRINTS-AUDIO ENGINEERING SOCIETY*, 2001.

[115] "IOSONO." [Online]. Available: http://www.iosono-sound.com/

[116] "SonicEmotion." [Online]. Available: http://www2.sonicemotion.com/

[117] J. Ahrens, M. Geier, and S. Spors, "The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Proceedings of the 124th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 2008.

[118] D. Menzel, H. Wittek, G. Theile, H. Fastl, and others, "The binaural sky: A virtual headphone for binaural room synthesis," in *Proceedings of the International Tonmeister Symposium*, Schloss Hohenkammer, Oct. 2005.

[119] de Vries, D. and colleagues, "Special event: Wave Field Synthesis Demonstration," Amsterdam, The Netherlands., 2011.

[120] M. M. Boone, E. N. Verheijen, and G. Jansen, "Virtual reality by sound reproduction based on Wave Field Synthesis," in *Proceedings of the 100th Audio Engineering Society Convention*, Copenhagen, Denmark, May 1996.

[121] M. M. Boone, "Multi-actuator panels (maps) as loudspeaker arrays for wave field synthesis," *Journal of the Audio Engineering Society*, vol. 52, no. 7/8, pp. 712–723, July 2004.

[122] M. M. Boone and W. P. de Bruijn, "On the applicability of distributed mode loudspeaker panels for wave field synthesis-based sound reproduction," in *Proceedings of the 108th Audio Engineering Society Convention*, Paris, France, Feb 2000.

[123] E. Corteel, "Equalization in an extended area using multichannel inversion and wave field synthesis," *Journal of Audio Engineering Society*, vol. 54, no. 12, pp. 1140–1161, Dec 2006.

[124] S. Spors, A. Kuntz, and R. Rabenstein, "An approach to listening room compensation with wave field synthesis," in *Proceedings of the 24th Audio Engineering Society Conference: Multichannel Audio, The New Reality*, Banff, Canada, Jun 2003.

[125] J. J. López, A. González, and L. Fuster, "Room compensation in wave field synthesis by means of multichannel inversion," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, NY, USA, Oct 2005, pp. 146–149.

[126] S. Spors, H. Buchner, and R. Rabenstien, "A novel approach to active listening room compensation for wave field synthesis using wave-domain adaptive filtering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Quebec, Canada, May 2004.

[127] R. T. Azuma and others, "A survey of augmented reality," *Presence*, vol. 6, no. 4, pp. 355–385, Aug 1997.

[128] J. R. Blum, M. Bouchard, and J. R. Cooperstock, "What's around me? Spatialized audio augmented reality for blind users with a smartphone," in *Proceedings of the 8th Mobile and Ubiquitous Systems: Computing, Networking, and Services*, Copenhagen, Denmark, Dec 2012, pp. 49–62.

[129] T. Nilsen, S. Linton, and J. Looser, "Motivations for augmented reality gaming," vol. 4, Dunedin, New Zealand, June 2004, pp. 86–93.

[130] T. Sielhorst, M. Feuerstein, and N. Navab, "Advanced medical displays: A literature review of augmented reality," *Journal of Display Technology*, vol. 4, no. 4, pp. 451–467, Dec 2008.

[131] T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen, "Application scenarios of wearable and mobile augmented reality audio," in *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, Germany, May 2004.

[132] M. Billinghurst and H. Kato, "Collaborative augmented reality," *Communications of the ACM*, vol. 45, no. 7, pp. 64–70, July 2002.

[133] T. Miyashita, P. Meier, T. Tachikawa, S. Orlic, T. Eble, V. Scholz, A. Gapel, O. Gerl, S. Arnaudov, and S. Lieberknecht, "An augmented reality museum guide," in *Proceedings of the IEEE/ACM International Symposium on Mixed and Augmented Reality*, Sep 2008, pp. 103–106.

[134] H. Ishii and B. Ullmer, "Tangible bits: towards seamless interfaces between people, bits and atoms," in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, New York, NY, USA, Mar 1997, pp. 234–241.

[135] R. W. Lindeman, H. Noma, and P. G. De Barros, "Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio," in *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, Darmstadt, Germany, Nov 2007, pp. 1–4.

[136] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, June 2004.

[137] M. Tikander, M. Karjalainen, and V. Riikonen, "An augmented reality audio headset," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Espoo, Finland, Sep 2008.

[138] J. Rämö and V. Välimäki, "Digital Augmented Reality Audio Headset," *Journal of Electrical and Computer Engineering*, vol. 2012, Oct 2012.

[139] D. Schobben and R. Aarts, "Three-dimensional headphone sound reproduction based on active noise cancellation," in *Proceedings of the 113th Audio Engineering Society Convention*, Los Angeles, CA, USA, Oct.

[140] 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, May 2005.

[141] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747–749, Dec 1998.

[142] K. Sunder, E.-L. Tan, and W.-S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *Journal of the Audio Engineering Society*, vol. 61, no. 12, pp. 989–1000, Dec 2013.

[143] H. Han, "Measuring a dummy head in search of pinna cues," *Journal of the Audio Engineering Society*, vol. 42, no. 1/2, pp. 15–37, Feb 1994.

[144] S. Brunner, H.-J. Maempel, and S. Weinzierl, "On the audibility of comb filter distortions," in *Proceedings of the 122nd Audio Engineering Society Convention*, Vienna, Austria, May 2007.

[145] J. Rämö, "Equalization techniques for headphones listening," Ph.D. dissertation, Aalto University, Department of Signal Processing and Acoustics, 2014.

[146] F. Brinkmann and A. Lindau, "On the effect of individual headphone compensation in binaural synthesis," *Fortschritte der Akustik: Tagungsband d. 36. DAGA*, pp. 1055–1056, 2010.

[147] M. Bouchard, S. G. Norcross, and G. A. Soulodre, "Inverse filtering design using a minimal-phase target function from regularization," in *Proceedings of the 121st Audio Engineering Society Convention*, SanFrancisco, CA, USA, Oct 2006.

[148] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, July 1979.

[149] S. M. Kuo and D. Morgan, *Active noise control systems: algorithms and DSP implementations*. John Wiley & Sons, Inc., 1995.

[150] 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, Aug 2007.

[151] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance dependent head-related transfer function database of KEMAR," in *Proceedings of the International Conference on Audio, Language and Image Processing*, Shanghai, China, July 2008, pp. 466–470.

[152] R. Ranjan, G. Woon-Seng, and C. Yong-Kim, "Applying Active Noise Control Technique for Augmented Reality Headphones," in *Proceedings of the Internoise*, Melbourne, Australia, Nov 2014.

[153] B. De Man and J. D. Reiss, "A pairwise and multiple stimuli approach to perceptual evaluation of microphone types," in *Proceedings of the 134th Audio Engineering Society Convention*, Rome, Italy, May 2013.

[154] E. Parizet, N. Hamzaoui, and G. Sabatie, "Comparison of some listening test methods: a case study," *Acta Acustica united with Acustica*, vol. 91, no. 2, pp. 356–364, Mar 2005.

[155] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Proceedings of the Independent Component Analysis and Signal Separation*. Springer, Sep 2007, pp. 454–461.

[156] A. H. Sayed, *Adaptive filters*. John Wiley & Sons, 2011.

[157] D. Duttweiler, "Proportionate normalized least-mean-squares adaptation in

echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 508–518, Sep 2000.

[158] H.-C. Huang and J. Lee, "A New Variable Step-Size NLMS Algorithm and Its Performance Analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 2055–2060, April 2012.

[159] R. H. Kwong and E. W. Johnston, "A variable step size lms algorithm," *IEEE Transactions on Signal Processing*, vol. 40, no. 7, pp. 1633–1642, July 1992.

[160] H.-C. Shin, A. H. Sayed, and W.-J. Song, "Variable step-size nlms and affine projection algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132–135, Feb 2004.

[161] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electronics and Communications in Japan (Part I: Communications)*, vol. 67, no. 5, pp. 19–27, 1984.

[162] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, June 2001.

[163] G. Enzner, "Analysis and optimal control of LMS-type adaptive filtering for continuous-azimuth acquisition of head related impulse responses," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,(ICASSP)*, Las, Vegas, USA, Mar 2008, pp. 393–396.

[164] T. Ajdler, L. Sbaiz, and M. Vetterli, "Dynamic measurement of room impulse responses using a moving microphone," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1636–1645, Sep 2007.

[165] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, and others, *Advances in network and acoustic echo cancellation*. Springer, 2001.

[166] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters–an overview," *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, 2000.

[167] C. Antweiler and M. Dörbecker, "Perfect sequence excitation of the NLMS algorithm and its application to acoustic echo control," vol. 49, pp. 386–397, July 1994.

[168] A. Telle, C. Antweiler, and P. Vary, "Der perfekte Sweep-Ein neues Anregungssignal zur adaptiven Systemidentifikation zeitvarianter akustischer Systeme," Berlin, Germany, Mar 2010, pp. 341–342.

[169] L. Hakansson, "The filtered-X LMS algorithm," *Lecture Notes, University of Karlskrona, Ronneby*, 2004.

[170] S. J. Elliott and P. A. Nelson, "Multiple-point equalization in a room using adaptive digital filters," *Journal of the Audio Engineering Society*, vol. 37, no. 11, pp. 899–907, Nov 1989.

[171] M. Zhang, H. Lan, and W. Ser, "On comparison of online secondary path modeling methods with auxiliary noise," *IEEE Transactions of Speech and Audio Processing*, vol. 13, no. 4, pp. 618–628, July 2005.

[172] S.-C. Chan and Y. Chu, "Performance analysis and design of FxLMS algorithm in broadband ANC system with online secondary-path modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 982–993, Mar 2012.

[173] M. Zhang, H. Lan, and W. Ser, "Cross-updated active noise control system

with online secondary path modeling," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 598–602, July 2001.

[174] M. Guldenschuh, "Secondary-path models in adaptive-noise-control headphones," in *Proceedings of the International Conference on Systems and Control (ICSC)*, Algiers, Algeria, Oct 2013, pp. 653–658.

[175] G. Long, F. Ling, and J. G. Proakis, "The LMS algorithm with delayed coefficient adaptation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 9, pp. 1397–1405, Sep 1989.

[176] H. Huynh, P. Fortier, and J. Martinet, "Generalized DLMS algorithm," in *Proceedings of the IEEE Conference on Communications, Computers and Signal Processing*, vol. 1, Victoria, BC, Canada, Aug 1997, pp. 448–452.

[177] J. Lopez, P. Gutierrez, M. Cobos, and E. Aguilera, "Sound distance perception comparison between wave field synthesis and vector base amplitude panning," in *Proceedings of the 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Athens, Greece, May 2014, pp. 165–168.

[178] K. Laumann, G. Theile, and H. Fastl, "A virtual headphone based on wave field synthesis," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3515, May 2008.

[179] F. Völk, J. Konradl, and H. Fastl, "Simulation of wave field synthesis," in *Proceedings of the Acoustics*, vol. 8, Paris, June 2008, pp. 1165–1170.

[180] J. J. Lopez, M. Cobos, and B. Pueo, "Elevation in Wave-Field Synthesis using HRTF Cues," *Acta Acustica united with Acustica*, vol. 96, no. 2, pp. 340–350, Mar 2010.

[181] M. Strauß, A. Sontacchi, M. Noisternig, and R. Holdrich, "A spatial audio in-

terface for desktop applications," in *Proceedings of the 24th Audio Engineering Society Conference*, Banff, Canada, June 2003.

[182] H. Wierstorf, A. Raake, M. Geier, and S. Spors, "Perception of Focused Sources in Wave Field Synthesis," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 5–16, Jan. 2013.

[183] H. Wittek, S. Kerber, F. Rumsey, and G. Theile, "Spatial perception in wave field synthesis rendered sound fields: Distance of real and virtual nearby sources," in *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, Germany, May 2004.

[184] J. Ahrens and S. Spors, "Sound field reproduction using planar and linear arrays of loudspeakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2038–2050, Nov 2010.

[185] S. Spors, "Investigation of spatial aliasing artifacts of wave field synthesis in the temporal domain," *Fortschritte der Akustik, DAGA*, Mar 2008.

[186] H. Wierstorf, C. Hohnerlein, S. Spors, and A. Raake, "Coloration in Wave Field Synthesis," in *Proceedings of the 55th Audio Engineering Society Conference*, Helsinki, Finland, Aug. 2014.

[187] C. Hohnerlein, "Coloration of virtual sources in Wave Field Synthesis for different loudspeaker spacings," Ph.D. dissertation, Technical University of Berlin, 2013.

[188] I. Recommendation, "BS. 1534-1. Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," *Proceedings of the International Telecommunications Union, Geneva*, 2001.

[189] E. Vincent, "MUSHRAM: A MATLAB interface for MUSHRA listening tests," *Online] http://www. elec. qmul. ac. uk/people/emmanuelv/mushram*, 2005.

[190] T. Assembly, "ITU-R BS. 1284-1: EN-General methods for the subjective assessment of sound quality," Technical Report. ITU, Tech. Rep., 2003.

[191] J. A. Belloch, M. Ferrer, A. Gonzalez, J. Lorente, and A. M. Vidal, "GPU-based WFS Systems with Mobile Virtual Sound Sources and Room Compensation," in *Proceedings of the 52nd Audio Engineering Society Conference*, Guildford, UK, Sep. 2013.

[192] C. Nvidia, "C programming guide version 4.2," *NVIDIA Corporation, Santa Clara, CA*, Apr. 2012.

[193] Mathworks T., "MATLAB: The Language of Technical Computing." [Online]. Available: http://www.mathworks.com/products/matlab/index.html

[194] D. Theodoropoulos, G. Kuzmanov, and G. Gaydadjiev, "Multi-core platforms for beamforming and wave field synthesis," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 235–245, Apr 2011.

[195] D. Theodoropoulos, C. B. Ciobanu, and G. Kuzmanov, "Wave field synthesis for 3D audio: architectural prospectives," in *Proceedings of the 6th ACM conference on Computing frontiers*, New York, NY, USA, May 2009, pp. 127–136.

[196] A. Lattanzi, E. Ciavattini, S. Cecchi, L. Romoli, and F. Ferrandi, "Real-Time Implementation of Wave Field Synthesis on NU-Tech Framework Using CUDA Technology," in *Proceedings of the 128th Audio Engineering Society Convention*, London, UK, May 2010.

[197] A. Lattanzi, F. Bettarelli, and S. Cecchi, "NU-Tech: the entry tool of the hArtes toolchain for algorithms design," in *Proceedings of the 124th Audio Engineering Society Convention*, New York, NY, USA, May 2008.

[198] F. Wefers and J. Berg, "High-performance real-time fir-filtering using fast

convolution on graphics hardware," in *Proceedings of the 13th Conference on Digital Audio Effects*, Graz Austria, Sep 2010.

[199] L. Savioja, V. Välimäki, and J. O. Smith, "Audio signal processing using graphics processing units," *Journal of the Audio Engineering Society*, vol. 59, no. 1/2, pp. 3–19, Mar 2011.

[200] C. Gregg and K. Hazelwood, "Where is the data? why you cannot debate cpu vs. gpu performance without the answer," in *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software, (ISPASS)*, Austin, TX, USA, Apr 2011, pp. 134–144.

[201] M. Boyer, J. Meng, and K. Kumaran, "Improving gpu performance prediction with data transfer modeling," in *Proceedings of the IEEE 27th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW)*, Cambridge, Massachusetts, USA, May 2013, pp. 1097–1106.

[202] "Memory transfer overhead," 2015. [Online]. Available: https://www.cs.virginia.edu/~mwb7w/cuda_support/memory_transfer_overhead.html

[203] "How to optimize data transfers in cuda c/c++," 2012. [Online]. Available: https://devblogs.nvidia.com/parallelforall/how-optimize-data-transfers-cuda-cc/

[204] J. A. Belloch, M. Ferrer, A. Gonzalez, F.-J. Martínez-Zaldívar, and A. M. Vidal, "Headphone-based spatial sound with a GPU accelerator," *Procedia Computer Science*, vol. 9, pp. 116–125, Dec 2012.

[205] J. A. Belloch, A. Gonzalez, F.-J. Martínez-Zaldívar, and A. M. Vidal, "Real-time massive convolution for audio applications on GPU," *The Journal of Supercomputing*, vol. 58, no. 3, pp. 449–457, Dec 2011.

[206] Mathworks T., "MATLAB GPU Computing Support for NVIDIA CUDA-Enabled GPUs." [Online]. Available: http://www.mathworks.com/discovery/matlab-gpu.html

[207] ——, "MATLAB-CUDA. CUDA kernel integration in MATLAB applications." [Online]. Available: http://www.mathworks.com/help/distcomp/executing-cuda-or-ptx-code-on-the-gpu.html

[208] C. Nvidia, *CUFFT library*. Version, 2010.

[209] M. Harris, S. Sengupta, and J. D. Owens, "Parallel prefix sum (scan) with CUDA," *GPU gems*, vol. 3, no. 39, pp. 851–876, Apr 2007.

[210] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, Mar 1992.

[211] Microsoft, "Microsoft hololens," 2015. [Online]. Available: https://www.microsoft.com/microsoft-hololens/en-us

[212] R. Jet, "Recon jet smart eyewear," 2015. [Online]. Available: http://store.reconinstruments.com/Recon-Jet

[213] Sony, "Sony smarteyeglass," 2015. [Online]. Available: http://developer.sonymobile.com/products/smarteyeglass