

AUDIO ENGINEERING SOCIETY PRESENTS ...



Audio
Engineering
Society
Presents

Immersive Sound

THE ART AND SCIENCE OF
BINAURAL AND
MULTI-CHANNEL
AUDIO

Edited by Agnieszka Roginska and Paul Geluso

A **Focal Press** Book



Immersive Sound

Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio provides a comprehensive guide to multi-channel and binaural sound theory and applications. With contributions from leading recording engineers, researchers, and industry experts, *Immersive Sound* includes an in-depth description of the physics and psychoacoustics of spatial audio as well as practical applications. Chapters include the history of 3D sound, binaural reproduction through headphones and loudspeakers, stereo, surround sound, height channels, object-based audio, sound field (ambisonics), wave field synthesis, and multi-channel mixing techniques. Knowledge of the development, theory, and practice of spatial and multi-channel sound is essential to those advancing the research and applications in the rapidly evolving fields of 3D sound recording, augmented and virtual reality, gaming, film sound, music production, and post-production.

Agnieszka Roginska is Music Associate Professor and Associate Director of the Music Technology program at New York University. Her research focuses on the analysis, simulation, and applications of immersive and 3D audio.

Paul Geluso is Music Assistant Professor at New York University in the Music Technology program. His work focuses on the theoretical, practical, and artistic aspects of sound recording and reproduction.



AUDIO ENGINEERING SOCIETY PRESENTS ...

www.aes.org

Editorial Board

Chair: Francis Rumsey, Logophon Ltd.

Hyun Kook Lee, University of Huddersfield

Natanya Ford, University of West England

Kyle Snyder, Ohio University

Other titles in the Series:

Handbook for Sound Engineers, 5th Edition

Edited by Glen Ballou

Audio Production and Critical Listening, 2nd Edition

Authored by Jason Corey

Recording Orchestra and Other Classical Music Ensembles

Authored by Richard King

Recording Studio Design, 4th Edition

Authored by Philip Newell

Modern Recording Techniques, 9th Edition

Authored by David Miles Huber

*Immersive Sound: The Art and Science of Binaural
and Multi-Channel Audio*

Edited by Agnieszka Roginska and Paul Geluso

Immersive Sound

The Art and Science of Binaural
and Multi-Channel Audio

Edited by Agnieszka Roginska
and Paul Geluso

First published 2018
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2018 Taylor & Francis

The right of Agnieszka Roginska and Paul Geluso to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

Chapter 1 entitled "Perception of Spatial Sound" by Elizabeth M. Wenzel, Durant R. Begault, and Martine Godfroy-Cooper is published with permission and in accordance with US Government Rights.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Names: Roginska, Agnieszka, editor. | Geluso, Paul, editor.

Title: Immersive sound : the art and science of binaural and multi-channel audio / edited by Agnieszka Roginska and Paul Geluso.

Description: New York ; London : Routledge, 2017.

Identifiers: LCCN 2017016918 | ISBN 9781138900011 (hardback) | ISBN 9781138900004 (pbk.)

Subjects: LCSH: Surround-sound systems.

Classification: LCC TK7881.83. I46 2017 | DDC 621.389/334—dc23

LC record available at <https://lccn.loc.gov/2017016918>

ISBN: 978-1-138-90001-1 (hbk)

ISBN: 978-1-138-90000-4 (pbk)

ISBN: 978-1-315-70752-5 (ebk)

Typeset in Sabon
by Apex CoVantage, LLC

Contents

<i>Contributors</i>	ix
<i>Foreword</i>	xii
WIESLAW WOSZCZYK	
<i>Acknowledgements</i>	xiii
Introduction	1
AGNIESZKA ROGINSKA AND PAUL GELUSO	
1 Perception of Spatial Sound	5
ELIZABETH M. WENZEL, DURAND R. BEGAULT, AND MARTINE GODFROY-COOPER	
<i>Auditory Physiology</i>	5
<i>Human Sound Localization</i>	10
<i>Head-Related Transfer Functions (HRTFs) and Virtual Acoustics</i>	19
<i>Neural Plasticity in Sound Localization</i>	22
<i>Distance and Environmental Context Perception</i>	23
<i>Conclusion</i>	33
2 History of 3D Sound	40
BRAXTON BOREN	
<i>Introduction</i>	40
<i>Prehistory</i>	41
<i>Ancient History</i>	41
<i>Space and Polyphony</i>	42
<i>Spatial Separation in the Renaissance</i>	43
<i>Spatial Innovations in Acoustic Music</i>	45
<i>3D Sound Technology</i>	49
<i>Technology and Spatial Music</i>	55
<i>Conclusions and Thoughts for the Future</i>	56

3	Stereo	63
	PAUL GELUSO	
	<i>Stereo Systems</i>	63
	<i>Creating a Stereo Image</i>	72
	<i>Stereo Enhancement</i>	82
	<i>Summary</i>	86
4	Binaural Audio Through Headphones	88
	AGNIESZKA ROGINSKA	
	<i>Headphone Reproduction</i>	90
	<i>Binaural Sound Capture</i>	96
	<i>HRTF Measurement</i>	99
	<i>Binaural Synthesis</i>	101
	<i>Inside-the-Head Locatedness</i>	106
	<i>Advanced HRTF Techniques</i>	108
	<i>Quality Assessment</i>	111
	<i>Binaural Reproduction Methods</i>	113
	<i>Headphone Equalization and Calibration</i>	116
	<i>Conclusions</i>	117
	<i>Appendix: Near Field</i>	122
5	Binaural Audio Through Loudspeakers	124
	EDGAR CHOEIRI	
	<i>Introduction</i>	124
	<i>The Fundamental XTC Problem</i>	128
	<i>Constant-Parameter Regularization</i>	138
	<i>Frequency-Dependent Regularization</i>	147
	<i>The Analytical BACCH Filter</i>	156
	<i>Individualized BACCH Filters</i>	160
	<i>Conclusions</i>	164
	<i>Appendix A: Derivation of the Optimal XTC Filter</i>	169
	<i>Appendix B: Numerical Verification</i>	178
6	Surround Sound	180
	FRANCIS RUMSEY	
	<i>The Evolution of Surround Sound</i>	181
	<i>Surround Sound Formats</i>	184
	<i>Surround Sound Delivery and Coding</i>	191

	<i>Surround Sound Monitoring</i>	198
	<i>Surround Sound Recording Techniques</i>	202
	<i>Perceptual Evaluation</i>	215
	<i>Predictive Models of Surround Sound Quality</i>	217
7	Height Channels	221
	SUNGYOUNG KIM	
	<i>Background</i>	221
	<i>Fundamental Psychoacoustics of Height-Channel Perception</i>	223
	<i>Multichannel Reproduction Systems With Height Channels</i>	225
	<i>Recording With Height Channels</i>	233
	<i>Conclusion</i>	241
8	Object-Based Audio	244
	NICOLAS TSINGOS	
	<i>Introduction</i>	244
	<i>Spatial Representation and Rendering of Audio Objects</i>	245
	<i>Advanced Metadata and Applications of Object-Based Representations</i>	254
	<i>Managing Complexity of Object-Based Content</i>	260
	<i>Audio Object Coding</i>	263
	<i>Capturing Audio Objects</i>	265
	<i>Tradeoffs of Object-Based Representations</i>	267
	<i>Object-Based Loudness Estimation and Control</i>	269
	<i>Object-Based Program Interchange and Delivery</i>	271
	<i>Conclusion</i>	272
9	Sound Field	276
	ROZENN NICOL	
	<i>Introduction</i>	276
	<i>Development of the Sound Field</i>	277
	<i>Higher Order Ambisonics (HOA)</i>	285
	<i>Sound Field Synthesis</i>	295
	<i>Sound Field Formats</i>	299
	<i>Conclusion</i>	300
	<i>Appendix A: Mathematics and Physics of Sound Field</i>	303
	<i>Appendix B: Mathematical Derivation of W, X, Y, Z</i>	308
	<i>Appendix C: The Optimal Number of Loudspeakers</i>	310

10	Wave Field Synthesis	311
	THOMAS SPORER, KARLHEINZ BRANDENBURG, SANDRA BRIX, AND CHRISTOPH SLADECZEK	
	<i>Motivation and History</i>	311
	<i>Separation of Sound Objects and Room</i>	318
	<i>WFS Reproduction: Challenges and Solutions</i>	320
	<i>WFS With Elevation</i>	327
	<i>Audio Metadata and WFS</i>	328
	<i>Applications Based on WFS and Hybrid Schemes</i>	329
	<i>WFS and Object-Based Sound Production</i>	330
11	Applications of Extended Multichannel Techniques	333
	BRETT LEONARD	
	<i>Source Panning and Spreading</i>	333
	<i>An Immersive Overhaul for Preexisting Content</i>	344
	<i>Considerations in Mixing for Film and Games</i>	346
	<i>Envelopment</i>	348
	<i>Musings on Immersive Mixing</i>	354
	<i>Index</i>	357

Contributors

Durand R. Begault, NASA Ames Research Center, works in the area of research and development of 3D audio and multi-modal technologies for aeronautic and space applications, including psychoacoustic research, human factors evaluation, sound quality, acoustical modeling, and communications engineering. He has been associated with the Human Systems Integration Division of NASA Ames Research Center since 1988.

Braxton Boren, American University, is an acoustics and audio researcher specializing in applications of science to music and the humanities. He earned his PhD at the Music and Audio Lab at New York University, and is Assistant Professor in the Audio Technology Program at American University.

Karlheinz Brandenburg, Fraunhofer Institute for Digital Media Technology IDMT, is Full Professor at Technische Universität Ilmenau and Director of the Fraunhofer Institute for Digital Media Technologies in Ilmenau. He is best known for his contributions to audio coding (MP3, AAC). Among others, he is currently supervising research on binaural hearing, immersive audio, and music information retrieval.

Sandra Brix, Fraunhofer Institute for Digital Media Technology IDMT, joined the Fraunhofer Institute for Digital Media Technology IDMT in Germany as the head of the virtual acoustics area in 2000, where she has been leading the Acoustics Department since 2012. Dr. Brix is highly involved in the development of the Wave-Field-Synthesis technology. She has been teaching at the Technical University of Ilmenau since 2007.

Edgar Choueiri, Princeton University, is Professor of Applied Physics at Princeton University. He heads both the Electric Propulsion and Plasma Dynamics Laboratory, where he works on advanced spacecraft propulsion, and the 3D Audio and Applied Acoustics (3D3A) Laboratory, where he works on psychoacoustics and virtual reality 3D audio.

Paul Geluso, New York University, focuses his work on the theoretical, practical, and artistic aspects of sound recording and reproduction. He is an active recording engineer currently serving as an Assistant Music Professor in Music Technology at New York University where he teaches courses in music production, electronics, ear training, and immersive sound technologies.

Martine Godfroy-Cooper, NASA Ames Research Center, has been a research associate on a cooperative agreement between the Human Systems Integration Division at NASA Ames Research Center and the Psychology Department at San Jose State University since 2005. She is currently responsible for the development of 3D audio and multimodal interfaces for the US Army Aviation Development Directorate degraded visual environment mitigation program.

Sungyoung Kim, Rochester Institute of Technology, is a researcher, teacher, recording engineer (Tonmeister), and guitarist. He worked for Korea Broadcasting System (KBS) (1996–2001) and Yamaha Corporation (2007–2012) before joining Rochester Institute of Technology (RIT). His research topics are cross-cultural difference in spatial hearing and rehabilitation of spatial hearing through VR technologies.

Brett Leonard, BLPaudio, is an audio educator, researcher, consultant, and freelance audio engineer with a specialty in acoustic music production. His research includes collaborations with NHK, SwissAudec, and Skywalker Sound, amongst others. His research focuses on understanding the complex interactions of spatial audio, human perception, and small room acoustics.

Rozenn Nicol, Orange Labs, works in the area of research and development of acoustics, 3D audio and sound perception for telecommunication applications at Orange Labs since 2000. She takes part in classes on Spatial Audio for universities (Le Mans, Brest) and engineering schools (ENST, ENSATT) in France.

Agnieszka Roginska, New York University, has been Music Associate Professor and Associate Director of the Music Technology program at New York University since 2006. Her research focuses on the simulation and applications of immersive and 3D audio, including the capture, analysis and synthesis of auditory environments, auditory displays, and their applications in augmented acoustic sensing.

Francis Rumsey, Logophon Ltd., is a technical writer, organist and consultant, Chair of the AES Technical Council and Consultant Editor of the *Journal of the Audio Engineering Society*. Until 2009 he was a Professor in the Music and Sound Recording Department of the University of Surrey, leading a research group concerned with psychoacoustics and sound reproduction.

Christoph Sladeczek, Fraunhofer Institute for Digital Media Technology IDMT, works in the field of spatial audio and acoustics where he (co)-authored numerous papers. He is the head of the Virtual Acoustics research group at Fraunhofer IDMT. Christoph is responsible for the development of object-based audio technology.

Thomas Sporer, Fraunhofer Institute for Digital Media Technology IDMT, works in the area of research and development of 3D audio, audio quality assessment, and applications based on acoustics. He has been associated with Fraunhofer since 1988. He was involved in the development of mp3, aac and MPEG-H.

Nicolas Tsingos, Dolby, leads the virtual and augmented reality exploration group at Dolby Laboratories. Previously, he designed the authoring and rendering tools for the Dolby Atmos cinema sound system and format. Nicolas also holds a tenure research position at INRIA, the French National Institute for Computer Science.

Elizabeth M. Wenzel, NASA Ames Research Center, has been a Research Psychologist with the Human Systems Integration Division at NASA Ames Research Center since 1986, directing development of 3D audio and multimodal display technology and conducting basic and applied research in auditory perception, localization in virtual acoustic displays, and multimodal information presentation for aerospace applications.

Foreword

Immersive sound produced over loudspeakers or headphones has the capacity to deliver a seamless illusion of alternative reality and change the way we relate to and behave in sound. It can revolutionize how we interact with, listen to, and live with music. It can redefine how we entertain, how we communicate, and how we collaborate. It opens creative possibility never imagined before, and it has the potential to improve the quality of life. Immersive sound is strategically aligned with the future of communication and entertainment.

The book you are about to read offers the wealth of information any researcher or practitioner in the audio field would eagerly welcome into his or her personal library. Agnieszka Roginska and Paul Geluso have succeeded in bringing together a capable team of experts who are able to address with eloquence the important aspects of perception, technology, and applications involved in auditory communication through immersive sound.

The goal of this book is to expand the reader's understanding of principles and working methods in immersive sound presented over loudspeakers and headphones using multi-channel and binaural technologies. While spatial audio techniques have a rich historical context, only recently have researchers and developers been able to collectively offer consumers a new paradigm in auditory experience through signal processing and streaming of high-quality digital audio. We are soon to arrive at the forefront of big changes in how people will access and interact with audio information.

The book is aimed at students, practitioners, and researchers who seek a thorough grounding in theory and practice of immersive sound. The information is invaluable, each chapter adding a critical component to the whole. This timely publication will no doubt help readers to acquire an informed and compelling view of the field.

Wieslaw Woszczyk

Acknowledgements

We would like to thank the many people who inspired and contributed to this project. First, we express our gratitude to the amazing team at Routledge, including Lara Zoble and Kristina Ryan, for their constant support and guidance through this process. Our very special thanks to the reviewers and readers for their critiques and comments, especially Wieslaw Woszczyk for his invaluable mentorship, and for contributing his thoughtful review and the foreword of this book. We thank all the chapter authors for sharing their time, expertise, and insight, whose invaluable contributions collectively make up this book. Without their commitment, enthusiasm, and dedication, this book would not have been realized. Working with this incredibly talented and devoted group of experts has been a true honor and inspiration for us. Finally, we would like to give a personal thanks to our very supportive families, colleagues, and friends.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Introduction

Agnieszka Roginska and Paul Geluso

Immersive Sound

Through sound, vision, touch, smell, and taste, multi-sensory integration into a scene can create an immersive experience. Immersive sound can give the listener an experience of *being there* through sound. Compared to vision, sound provides a fully immersive experience and can be perceived from all directions simultaneously. In fact, sound has the ability to ground a listener in a fixed location while other sensory information changes simultaneously. Filmmakers are well aware of this effect, often using sound to establish a fixed location in a scene while having the visual perspective change frequently.

For a moment, let's consider an immersive experience generated using sound alone. The location relationships between the listener, sound source(s) and the boundaries of a room create auditory cues that convey a sense of space. A continuous yet seemingly directionless sea of sound can envelop a listener through the use of environmental elements such as reverberation. Using sound this way, the sense of being immersed can be accomplished through a constructed sound-scape of directional and non-directional sounds surrounding the listener.

For example, New York's Grand Central station creates a natural immersive audio environment through a combination of close directional sounds such as conversation and foot steps. These combine with farther sound sources such as announcements, and interact with the acoustics of the space to create the reverberant and enveloping environment. The combination of sound sources and the enveloping environment create the immersive auditory experience. A sound source can become an environmental and enveloping sound or remain a point source, depending on the focus and perception of the listener.

The Listening Experience

The natural listening environment can be defined as the acoustic space we occupy during our daily lives. A virtual auditory space is an acoustic environment created through the use of loudspeakers or headphones designed to replace or augment the natural listening environment. Using technology, sounds from our natural listening environment can be captured, processed, stored, and/or transmitted to be reproduced in a virtual auditory space. The natural listening environment and the virtual auditory space both contribute to the listener's experience.

Natural listening defines the way we hear normally, without the aid of loudspeakers. Very realistic virtual auditory spaces that approximate the natural listening environment can be created with the use of loudspeakers and headphones through various techniques described in this book. The goal of immersive sound may be to recreate a sound environment that is as close as possible to the real world, or it may be to create an experience that augments the real world and can only exist in the virtual space. To create a listener's virtual experience, natural sounds can be captured or synthesized, processed, and played back using immersive sound reproduction systems.

Listener's Perspective

In a virtual listening environment, we rely on both analytical and psychoacoustic abilities to make sense of the sounds that are being reproduced using loudspeakers or headphones. Consider a mono recording made with a single microphone played back through a single loudspeaker. In

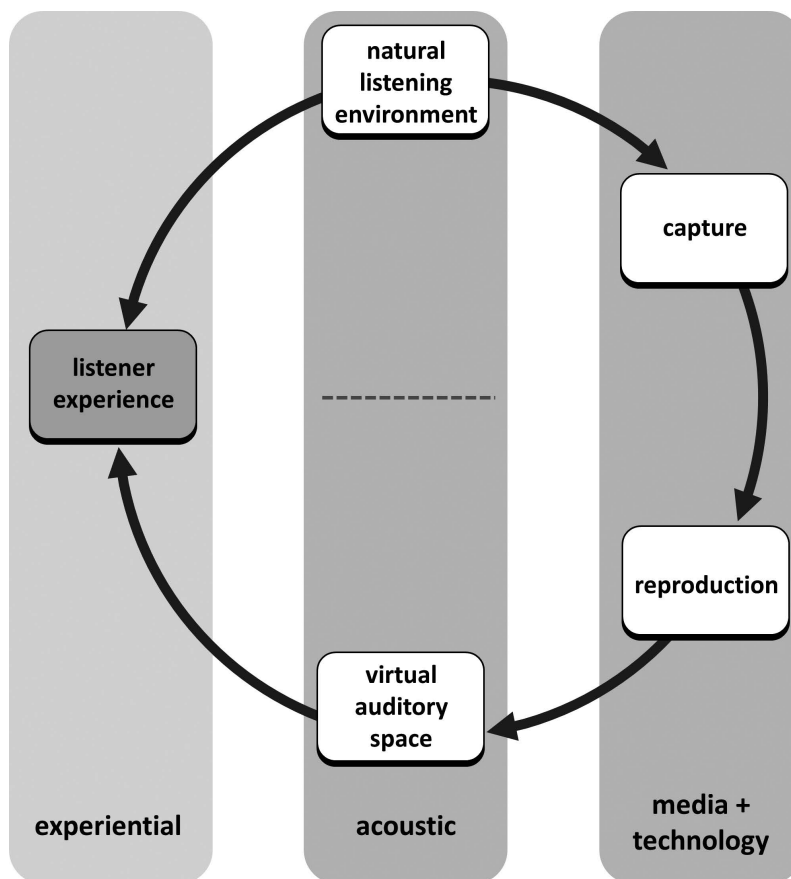


Figure 0.1 The listening experience.

this situation, we can imagine how distant the sound sources are from the microphone and get a general impression of the recording space by analyzing the relative volume, timbre, and room reflections for each sound source captured in the recording. If a second microphone and loudspeaker are added to the system, our auditory system now has two signals to analyze, thus more physical information to process in order to extract spatial information to inform the listener's perspective.

A fixed perspective can be achieved by creating a monophonic or stereophonic sound stage where the virtual sound stage is bounded by the speakers in the frontal plane. Surround sound expands to a panoramic virtual sound stage including the listening environment, or placing the listener on the sound stage surrounded by virtual sound sources. These systems are considered channel based and rely on a priori knowledge about the location of the loudspeakers and their relationship to the listener. Sound field and wave field systems move away from the channel-based model and aim to reproduce the physical wavefront as it would appear in a natural environment by utilizing multiple loudspeakers.

Binaural audio reproduction techniques take advantage of the human natural spatial auditory cues to recreate a virtual auditory environment through headphones or loudspeakers that emulate headphone reproduction. This results in a “you are there”, first-person perspective, in contrast to the loudspeaker “they are here” systems described above.

As our recording and playback systems become more sophisticated, the immersive listening experience is guided by reproduction systems that can better approximate the natural listening environment. Thus, allowing the listener to rely less on their a priori knowledge to make sense of what they hear, and move toward a more natural listening experience. This creates a more compelling and *easier to listen to* virtual auditory environment.

About This Book

The intended audiences for this book are those enrolled in undergraduate, graduate study, and higher learning institutions in the areas of music technology, recording and production, sound design, sound art and film, as well as audiophiles, game designers, simulation and virtual reality professionals, post-production professionals, and entertainment professionals. We assume readers have a fundamental understanding of the physics of sound, digital audio, recording, and reproduction principles.

Chapters are written by experts in their corresponding field of research within immersive audio. Chapter authors are researchers in academic institutions, research laboratories in the industry, US government agencies. The body of the text is grounded in research and empirical work, with each chapter covering the evolution and historical perspective of the development of a reproduction technology.

The organization of the book proceeds with a chapter by Elizabeth Wenzel, Durand Begault, and Martine Godfroy-Cooper about the physiology, psychoacoustics, and acoustics of spatial hearing, including the perception of spatial sound, binaural cues, perceptual plasticity, distance perception, and environmental context for immersive sound. In Chapter Two, Braxton Boren takes the reader on a historical journey across immersive sound starting with the impact of immersive sound on our prehistoric ancestors, through the use of spatial separation of instruments and choirs as a compositional tool in the 15th century, to modern-age techniques and technologies of immersive sound. In Chapter Three, Paul Geluso introduces stereo loudspeaker systems and discusses techniques of sound capture, reproduction, and methods of stereo enhancement.

Chapter Four, written by Agnieszka Roginska, describes the capture, synthesis, and reproduction of binaural sound over headphones, including extended reproduction techniques and applications of binaural sound. In Chapter Five, Edgar Choueiri describes the principles of binaural audio reproduction over loudspeakers using crosstalk cancellation. Continuing in Chapter Six, Francis Rumsey discusses the evolution and principles of surround sound, with speakers located on the horizontal plane. These techniques are extended in Chapter Seven, where Sungyoung Kim describes the methods of surround sound reproduction with height speakers, including the psychoacoustics of height perception, configuration of loudspeakers, and recording techniques.

Nicolas Tsingos discusses the principles of object-based audio in Chapter Eight, where he describes audio objects, their representation, advanced metadata, audio object capture, and rendering. In Chapter Nine, Rozenn Nicol addresses the theory and practice of sound field capture and reproduction, starting from the initial development of the sound field approach to High Order Ambisonics. Wave Field Synthesis is described in Chapter Ten, where Thomas Sporer, Karlheinz Brandenburg, Sandra Brix, and Christoph Sladeczek describe the development of Wave Field Synthesis from one of the earliest examples of immersive sound through the acoustic curtain by Steinberg and Snow in 1934, through the theory and practice of WFS reproduction, and its limitations and applications using modern techniques and signal processing. The book concludes with Brett Leonard describing the applications of extended multi-channel techniques in Chapter Eleven, where he discusses broader concepts and introduces practical mixing techniques for engineers working with immersive sound.

Perception of Spatial Sound

*Elizabeth M. Wenzel, Durand R. Begault,
and Martine Godfroy-Cooper*

Immersion refers acoustically to sounds as coming from all directions around a listener, which normally is an inevitable consequence of natural human listening in an air medium. Audible sound sources are everywhere in real environments where sound waves propagate and reflect from surfaces around a listener. Even in the quietest of environments, such as an anechoic chamber, the sounds of one's own body will be audible. However, the common meaning of immersion in audio and acoustics refers to the psychological sensation of being surrounded by specific sound sources as well as ambient sound. Although acoustically a sound can reach a listener from multiple surrounding directions, its spatial characteristics may be judged as unrealistic, static or constrained. For example, good quality concert hall acoustics has traditionally been correlated with a listener's sensation of being immersed by the sound of the orchestra, as opposed to the sound seeming distant and removed. Spatial audio techniques, particularly 3D audio, can provide an immersive experience because virtual sound sources and sound reflections can be made to appear from anywhere in space around a listener. This chapter introduces a listener to the physiological, psychoacoustic and acoustic bases of these sensations.

Auditory Physiology

Auditory perception is a complex phenomenon determined by the physiology of the auditory system and affected by cognitive processes. The auditory system transforms the fundamental independent aspects of sound stimuli, such as their spectral content, temporal properties and location in space into distinct patterns of neural activity. These patterns will give rise to the qualitative experience of pitch, loudness, timbre and location. They will ultimately be integrated with information from the other sensory systems to form a unified perceptual representation and provide behavior guidance that includes orienting to acoustical stimuli and engaging in intra-species communication.

Auditory Function: Peripheral Processing

The functional auditory system extends from the ears to the brain's frontal lobes with successively more complex functions occurring as one ascends the hierarchy of the nervous system (Figure 1.1). The different functions performed by the auditory system are classically categorized as *peripheral auditory processing* and *central auditory processing*.

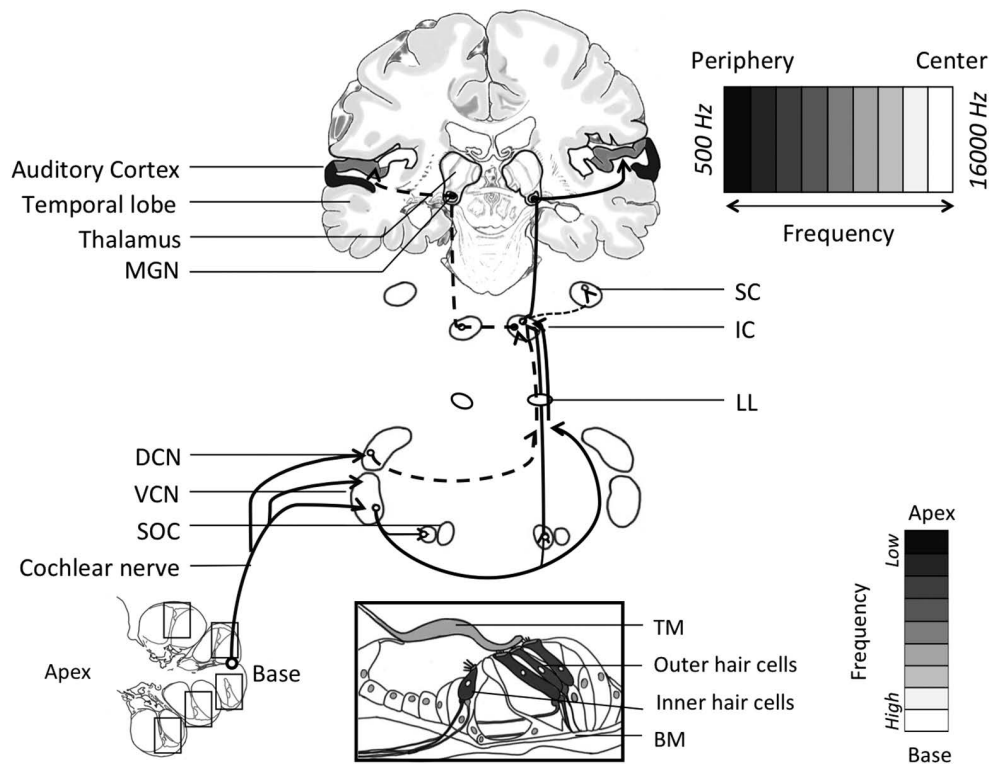


Figure 1.1 Diagram of the major auditory pathways. Note the auditory system entails several parallel pathways and that information from each ear reaches both sides of the system (dashed line: ipsilateral, continuous line: contralateral). Top Right: Tonotopic representation in the auditory cortex. High frequencies are represented at the center of the cortical map while low frequencies are represented at its periphery. Bottom Left: Cochlea. Bottom Center: Organ of Corti. Bottom Right: Cochleotopic coding along the basilar membrane (BM).

The peripheral auditory system includes processing stages from the outer ear to the cochlear nerve. A crucial transformation is performed within these early stages, which is often compared to a Fourier analysis of the incoming sound waves that defines how the sounds are processed at the later stages of the auditory hierarchy. Sound enters the ear as pressure waves. At the periphery of the system, the external and the middle ear respectively collect sound waves and selectively amplify their pressure, so that they can be successfully transmitted to the fluid-filled cochlea in the inner ear.

The *external ear*, which consists of pinna (plural, pinnae) and auditory meatus (or canal), gathers the pressure waves and focuses them on the eardrum (tympanic membrane) at the end of the canal. One consequence of the configuration of the human auditory canal is that it selectively boosts the sound pressure 30 to 100 fold for frequencies around 3 kHz via a passive resonance

effect due to the length of the ear canal. This amplification makes humans especially sensitive to frequencies in the range of 2–5 kHz, which appears to be directly related to speech perception. A second important function of the pinnae is to selectively filter sound frequencies in order to provide cues about the elevation of a sound source: up/down and front/back angles (Shaw, 1974). The vertically asymmetrical convolutions of the pinna are shaped so that the external ear transmits more high frequency components from an elevated source than from the same source at ear level. Similarly, high frequencies tend to be more attenuated for sources in the rear than for sources in the front, as a consequence of the orientation and structure of the pinna (Blauert, 1997).

The *middle ear* is a small cavity, which separates the outer and the inner ear. The cavity contains the three smallest bones (hammer, anvil and stirrup) in the body called ossicles, connected more or less flexibly to each other. Its major function is to match the relatively low impedance (impedance in this context refers to a medium's resistance to movement) airborne sounds to the higher impedance fluid in the inner ear. Without this action, there would be a loss of transmission of 1000:1 corresponding to a loss of sensitivity of 30 dB. The middle ear has two small muscles known as the tensor tympani and the stapedius, which have a protective function. When sound pressure reaches a threshold loudness level (at approximately 85 dB HL¹ in humans with normal hearing), a sensory driven afferent signal is sent to the brainstem via the cochlear nerve, which initiates an efferent reflexive contraction of the stapedius muscle within the middle ears referred to as the stapedius reflex or acoustic middle ear reflex. The excitation of the muscle results in a stimulus level-dependent attenuation of low-frequency (< 1 kHz) ossicular chain vibration reaching the cochlea (Wilson & Margolis, 1999).

Transduction Mechanisms in the Cochlea

The cochlea in the *inner ear* is the most critical structure in the peripheral auditory pathway. The cochlea is a small-coiled snail-like structure that responds to the sound-induced vibrations and converts them into electrical impulses, a process known as mechanoelectrical transduction. Cochlear signal transduction involves amplification and decomposition of complex acoustical waveforms into their component frequencies.

Two membranes, the basilar membrane (BM) and the vestibular membrane (VM), divide the cochlea in three fluid-filled chambers. The organ of Corti sits on the BM and contains an array of sensory hair cells that contact with the tectorial membrane (TM), a structure that plays multiple, critical roles in hearing including coupling elements along the length of the cochlea, supporting a travelling wave and ensuring the gain and timing of cochlear feedback are optimal (Richardson, Lukashkin, & Russell, 2008). The sensory hair cells are responsible for the mechanoelectrical transduction, i.e., the transformation of mechanical stimulus into electrochemical activity. The acoustical stimulus initiates a traveling wave by displacing the hair, thus enabling the encoding of frequency, amplitude and phase of the original sound stimulus by the electrical activity of the auditory nerve fibers. As the BM displaces, it causes deflection in the hair bundles (stereociliae, tiny processes that protrude from the apical ends of the hair cells) of the hair cells of the location-matched inner hair cells, which results in a current flow and ultimately an action potential. Because the stiffness of the BM changes throughout the cochlea, the displacement of the membrane induced by an incoming sound depends on the frequency of that sound. Specifically,

the BM is stiffer near its “base” than near the middle of the spiral (“the apex”). Consequently, high frequency sounds (20 kHz) produce displacement near the base, while low frequency sounds (20 Hz) disturb the membrane near the apex. As a result, each locus on the BM is identified by its characteristic frequency (CF) and the whole BM can be described as a bank of overlapping filters (Patterson et al., 1987; Meddis & Lopez-Poveda, 2010). Because the BM displaces in a frequency-dependent manner, the corresponding hair cells are “tuned” to sound frequency. The resulting spatial proximity of contiguous preferred sound frequency (“place theory of hearing”, von Békésy, 1960; “place code” model, Jeffress, 1948) is referred to as tonotopy or better, cochleotopy. This tonotopic organization is carried up through the auditory hierarchy to the cortex (Moerel et al., 2013, Saenz & Langers, 2014) and defines the functional topography in each of the intermediate relays.

Auditory Function: Central Processing

The central auditory system is composed of a number of nuclei and complex pathways that ascend within the brainstem. The earliest stage of central processing occurs at the cochlear nuclei (dorsal, DCN and ventral, VCN), where the tonotopic organization of the cochlea is maintained. Accordingly the output of the CN has several targets.

One is the superior olivary complex (SOC), the first point at which information from the two ears interacts. The best understood function of the SOC is sound localization. Humans use at least two different strategies, and two different pathways, to localize the horizontal position of sound sources, depending on the frequencies of the stimulus. For frequencies below 3 kHz, which the auditory nerve can follow in a phase-locked manner, interaural time differences (ITDs) are used to localize the source; above these frequencies, interaural intensity differences (IIDs) are used as cues (King & Middlebrooks, 2011; Yin, 2002). ITDs are processed in the medial superior olive (MSO), while IIDs are processed in the lateral superior olive (LSO). These two pathways eventually merge in the midbrain auditory centers. The elevation of sound sources is determined by spectral filtering mediated by the external pinnae. Experimental evidence suggests that the spectral notches created by the shape of the pinnae are detected by neurons in the DCN. See the section on human sound localization for additional discussion of these cues.

The binaural pathways for sound localization are only part of the output of the CN. A second major set of pathways from the CN bypasses the SOC and terminates in the nuclei of the lateral lemniscus (LL) on the contralateral side of the brainstem. These particular pathways respond to sound arriving at one ear only and are thus referred to as monaural. Some cells in the nuclei of the LL signal the onset of sound, regardless of its intensity or frequency. Other cells process other temporal aspects of sound such as duration.

As with the outputs of the SOC, the pathways from the LL project to the midbrain auditory center, also known as the inferior colliculus (IC). This structure is a major integrative center where the convergence of binaural inputs produces a computed topographical representation of the auditory space. At this level, neurons are typically sensitive to multiple localization cues (Chase & Young, 2006) and respond best to sounds originating in a specific region of space, with a preferred elevation and a preferred azimuthal location. As a consequence, it is the first point at which auditory information can interact with the motor system. Another important property of the IC is its ability to process sounds with complex temporal patterns. Many neurons in the

IC respond only to frequency-modulated sounds while others respond only to sounds of specific durations. Such sounds are typical components of biologically relevant sounds, such as those made by predators and in humans, speech.

The IC relays auditory information to the medial geniculate nucleus (MGN) of the thalamus, which is an obligatory relay for all ascending information destined for the cortex. It is the first station in the auditory pathway where pronounced selectivity for combinations of frequencies is found. Cells in the MGN are also selective for specific time intervals between frequencies. The detection of harmonic and temporal combination of sounds is an important feature of the processing of speech.

In addition to the cortical projection, a pathway to the superior colliculus (SC) gives rise to an organized representation of ITDs and IIDs in a point-to-point map of the auditory space (King & Palmer, 1983). Topographic representations of multiple sensory modalities (visual, auditory and somatosensory) are integrated to control the orientation of movements toward specific spatial locations (King, 2005).

Auditory Function: The Auditory Cortex

The auditory cortex (AC, Figure 1.2) is the major target of the ascending fibers from the MGC and plays an essential role in our conscious perception of sound, including speech comprehension, which is arguably the most significant social stimulus for humans.

Although the AC has a number of subdivisions, a broad distinction can be made between a primary area and a secondary area. The primary auditory cortex (BA41, core area) located on the superior temporal gyrus of the temporal lobe receives point-to-point input from the MGC and

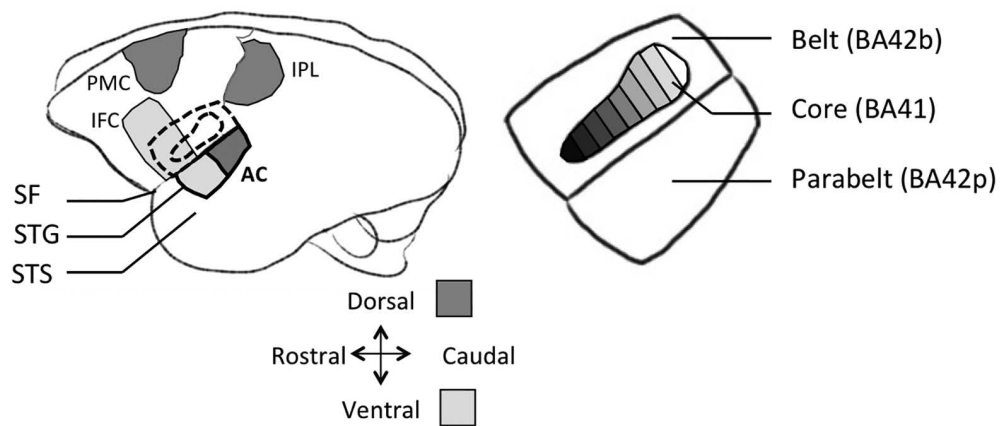


Figure 1.2 Left: Auditory cortex (AC) location in the human brain. The superior temporal gyrus (STG) is bordered superiorly by the Sylvian fissure (SF). The STG is bordered on the inferior side by the superior temporal sulcus (STS). Dorsal stream: intraparietal lobe (IPL), premotor cortex (PMC). Ventral stream: inferior frontal cortex (IFC). Right: The primary auditory cortex is denoted as the auditory Brodmann's area 41 (BA41) in humans. The auditory association cortex is subdivided into lateral belt and parabelt cortices (BA42). The view is from above the STG, after removing the overlying cortex, revealing the superior temporal plane.

comprises three distinct tonotopic fields (Saenz & Langers, 2014). Neurons in BA41 have narrow tuning functions and respond best to tone stimuli, supporting basic auditory functions such as frequency discrimination and sound localization. It also plays a role in processing of within-species communication sounds.

The belt areas of the auditory cortex receive more diffuse inputs from the MGC as well as inputs from BA41, and are less precise in their tonotopic organization. Neurons in the belt areas (BA42b) have broader frequency tuning functions and respond better to complex sounds, such as those that mediate communication. Lateral to the belt is a region of cortex denoted as the parabelt (BA42p) where neurons prefer complex stimuli including band-passed noise, moving stimuli and vocalizations (Rauschecker, Tian & Hauser, 1995).

The projections from the parabelt out of the auditory cortex to higher order cortical structures define the auditory dorsal processing stream and the ventral processing stream (see Figure 1.2). According to the auditory dual-stream model, spatial information (“where”) is primarily processed within the dorsal stream and non-spatial information (object features, i.e., “what”) within the ventral stream (Rauschecker & Tian, 2000; Romanski & Goldman-Rakic, 2002). The auditory ventral stream supports the perception and recognition of auditory objects and is involved in the processing of pitch changes, auditory working memory for words and tones, as well as semantic processing. There is less agreement regarding the functional role of the auditory dorsal stream. The earliest models argued for a role in spatial hearing, but recent research suggests that the auditory dorsal stream supports an interface with the motor system. In fact, this segregation between dorsal/ventral streams appears to be more relative than absolute. It seems that the functions of sound identification and spatial analysis are rather co-localized in the dorsal and ventral auditory streams (Gifford & Cohen, 2005; Lewald et al., 2008) and that human spatial processing is strongly linked with functions of pitch perception (Douglas & Bilkey, 2007).

Connections from the auditory cortex to the frontal lobes mediate a number of functions including language, object recognition and spatial localization. The frontal cortex is a heterogeneous region with multiple functional subdivisions, including the prefrontal cortex (PFC), and is part of the association cortices. The frontal association cortex is importantly involved in guiding complex behavior by planning responses, ongoing stimulation or remembered information. Collectively, the association cortices mediate the cognitive functions of the brain, including speech processing and executive functions that include attention, working memory, planning and decision-making (Fuster, 2009; Plakke et al., 2014).

Human Sound Localization

Primary Localization Cues

Auditory spatial perception refers to the ability to localize individual sound sources in 3D space even when multiple, simultaneous sources are present. Unlike the visual and somatosensory systems, spatial information is not directly represented at the sensory receptor in the auditory system. Instead, spatial locations are estimated by integrating neural binaural properties and frequency-dependent pinna filtering (binaural and monaural cues, Yost & Dye, 1997).

Sound source location is often specified in terms of azimuth, elevation and distance using a coordinate system in which a listener facing directly forward is defined as 0° azimuth and 0°

elevation. Azimuth is defined by the angle (θ) between the source location and the median plane at 0° azimuth (projected onto the horizontal plane) and elevation is the angle (δ) between the source location and the horizontal plane at 0° elevation (projected onto the median plane).

Azimuths to right of the listener are positive, to the left are negative, and the rear is defined as 180° .² Elevations are positive for upper directions and negative for lower directions relative to the listener. Distance is defined as the radius (r) projected along the vector formed by the azimuth and elevation of the source. Another important terminological distinction relevant to interaural cues is between the ipsilateral and contralateral ears. The ipsilateral ear is that closest to the sound source; sound thus arrives first and is greater in intensity at the ipsilateral ear. The contralateral ear is that farthest from the sound source, thus sound arrives later and with less intensity at the contralateral ear.

The localization of a sound source in the horizontal dimension (azimuth) results from the detection of left-right interaural differences in time of arrival and interaural differences in intensity at the two ears (Middlebrooks & Green, 1991). These cues also facilitate speech intelligibility in background noise in human listeners (Culling, Hawley & Litovsky, 2004). To localize a sound in the vertical dimension (elevation) and to resolve front-back confusions, the auditory system relies on the detailed geometry of the pinnae, causing acoustic waves to diffract and undergo direction-dependent reflections (Blauert, 1997; Hofman & Van Opstal, 2003). The two different modes of indirect coding of the position of a sound source in space (as compared to the direct spatial coding of visual stimuli) result in differences in spatial resolution in these two directions.

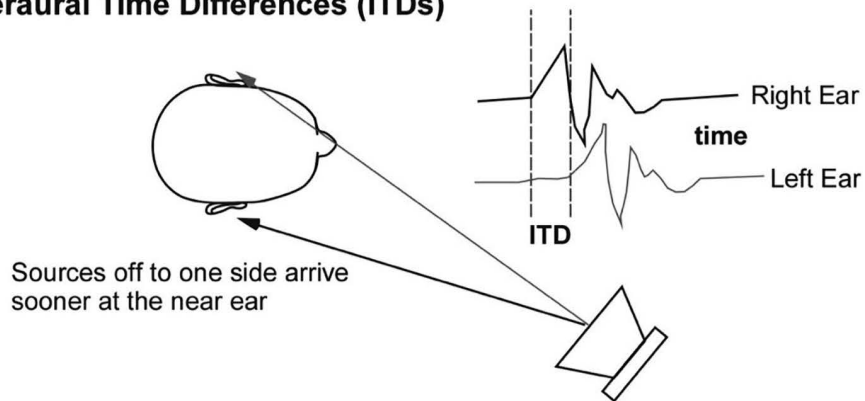
Sound Localization in the Horizontal Dimension

Much of the research on human sound localization in azimuth has derived from Lord Rayleigh's "duplex theory" (1907) which emphasizes the role of two primary cues (top two panels of Figure 1.3): interaural time differences (ITDs) and interaural intensity differences (IIDs, also referred to as Interaural Level Differences, ILDs, particularly when specified in dB SPL). Because the theory had been based primarily on experiments with single-frequency (sine wave) sounds, the original proposal was that IIDs resulting from head-shadowing determine localization at high frequencies (roughly 2000 Hz for larger human heads), while ITDs were thought to be important only for low frequencies because of the phase ambiguities occurring at frequencies greater than ~1000–1500 Hz. Recently, Brughera, Dunai and Hartmann (2013) have demonstrated that humans are sensitive to ITD fine structure in sound (temporal fine structure, TFS; the sound pressure waveform) up to a limit of 1400 Hz. For broadband sounds, the situation is more complex, and ITDs contribute to sound localization at higher frequencies, up to 4000 Hz (Bernstein, 2001).

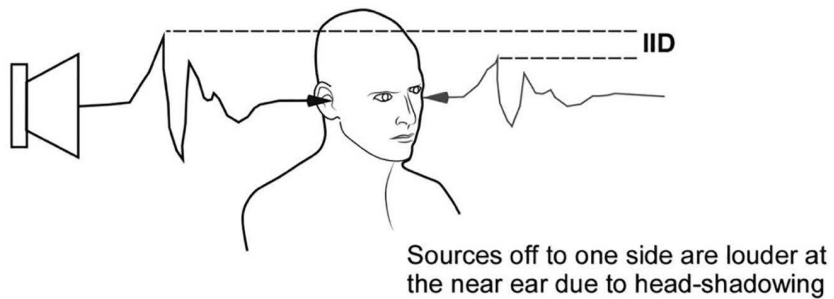
The interaural differences present in natural spatial hearing can be understood by considering a simplified rigid sphere model of a listener with a perfectly round head and no outer ears (spherical head model, Woodworth, 1938) placed at a fixed distance in an anechoic chamber from a broadband sound source at eye level (see Figure 1.4).

Modeling this situation involves calculating two paths representing the sound source wavefront from its center of origin to two points representing the entrance to the ear canals. An additional simplification is the placement of these points exactly at the midline crossing the sphere, at the ends of the interaural axis. With the source at position A at 0° azimuth, the path lengths are equal, causing the wavefront to arrive at the eardrums at the same time and with equal intensity.

Interaural Time Differences (ITDs)



Interaural Intensity Differences (IIDs)



Spectral Shaping by the Pinnae (Outer Ear) Structures

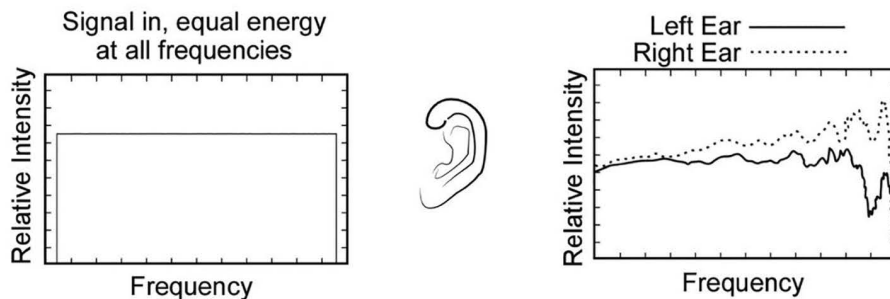


Figure 1.3 Illustration of the primary cues for human sound localization.

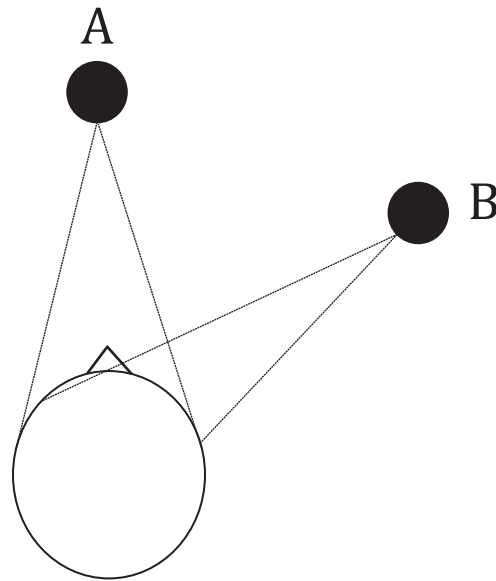


Figure 1.4 A listener in an anechoic chamber, with a sound source oriented directly ahead on the median plane (A, 0°) and displaced to +60° azimuth (B).

At position B, the sound source is at +60° azimuth to the right of the listener, and the paths are now unequal; this will cause the sound source wavefront to arrive later in time at the left ear relative to the right. This sound path length difference is the basis of the ITD cue, and it relates to the hearing system's ability to detect interaural phase differences (IPDs) below approximately 1,000 Hz. If the sounds are pure tones, a simple frequency factor relates the ITDs to the IPDs, for which there are known iso-IPD boundaries (90°, 180°. . .) defining regions of spatial perception. Although dependent upon the nature of the stimulus and the measurement technique, a value of around 650 μ sec (750 μ sec for low-frequency sounds, Kuhn, 1977) is a good approximation of the observed maximum value for an average human head. Figure 1.5 illustrates how the ITD changes as a function of the azimuthal angle of incidence.

The auditory system can phase-lock, or change the rate of neural firing corresponding to the peaks in a stimulus waveform, as long as the inter-peak interval is above about 1 msec. Such phase-locking can occur for the peaks in either sine waves or the envelopes of complex signals. One theory is that ITDs are estimated by the auditory system by comparing peaks in neural firing rates between the two ears in a manner known as coincidence detection (place code or topographic model, Jeffress, 1948), a neural mechanism that enables interaural cross-correlation. More recent data support the opponent channel coding of auditory space in humans (van Bergeijk, 1962; Magezi & Krumbholz, 2010; Salminen et al., 2010) where ITD is represented by a non-topographic population rate code, which involves only two opponent (left and right) channels,

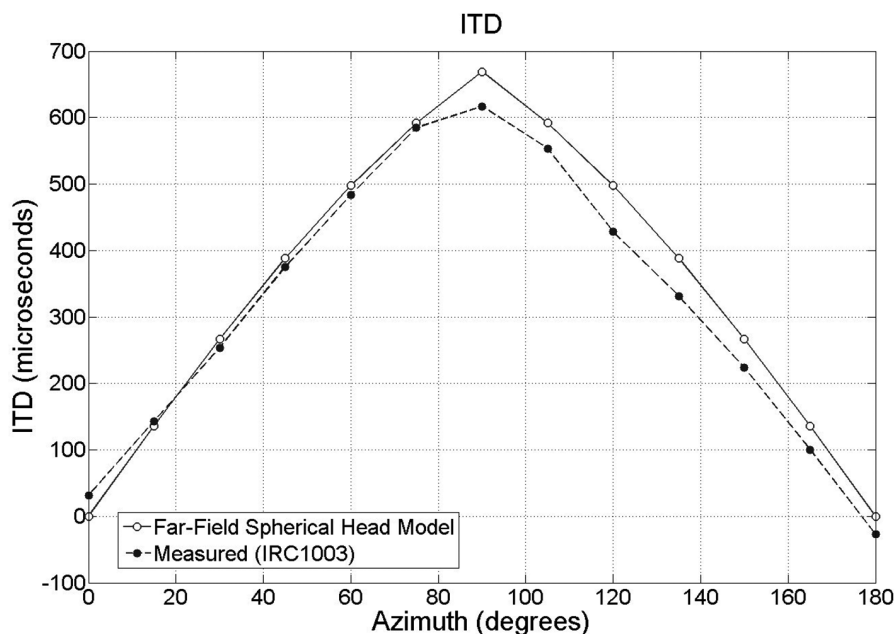


Figure 1.5 Interaural time differences (ITDs) plotted as a function of sound source azimuth (similar to data published by Feddersen et al., 1957). The open dots represent ITDs computed from the spherical head model and the closed dots are measured data at 0° elevation for subject IRC_1003 of the IRCAM Listen database (<http://recherche.ircam.fr/equipes/salles/listen/>). ITDs were computed using the method described in Miller, Godfroy-Cooper and Wenzel (2014).

broadly tuned to ITDs from the two auditory hemifields. The data suggest that the majority of ITD-sensitive neurons in each hemisphere are tuned to ITDs from the contralateral hemifield.

The sound source at position B in Figure 1.4 will also yield a significant interaural intensity difference cue, but only for those waveform components that are smaller than the diameter of the head, i.e., for frequencies greater than about 2,000 Hz. Higher frequencies will be attenuated at the left ear because the head acts as an obstacle, creating a “head shadow” effect at the opposite side. The relation of a wavefront to an obstacle of fixed size is such that the shadow effect increases with increasing frequency (i.e., decreasing size of the wavelength). However, below 2,000 Hz, the IID is no longer effective as a natural spatial hearing cue because longer wavelengths will diffract (“bend”) around the obstructing surface of the head, thereby minimizing the intensity differences. Figure 1.6 illustrates this head-shadow effect by plotting the IID as a function of azimuth location and stimulus frequency. Measured data from the literature for IIDs shows that a 3,000 Hz sine wave at 90° azimuth will be attenuated by about 10 dB, a 6,000 Hz sine wave will be attenuated by about 20 dB, and a 10,000 Hz wave by about 35 dB (Feddersen et al., 1957; Middlebrooks & Green, 1991). Measured IIDs derived from an individual listener are also shown in Figure 1.6 (bottom). Note that the pattern of IIDs can be quite complex across both frequency and azimuth.

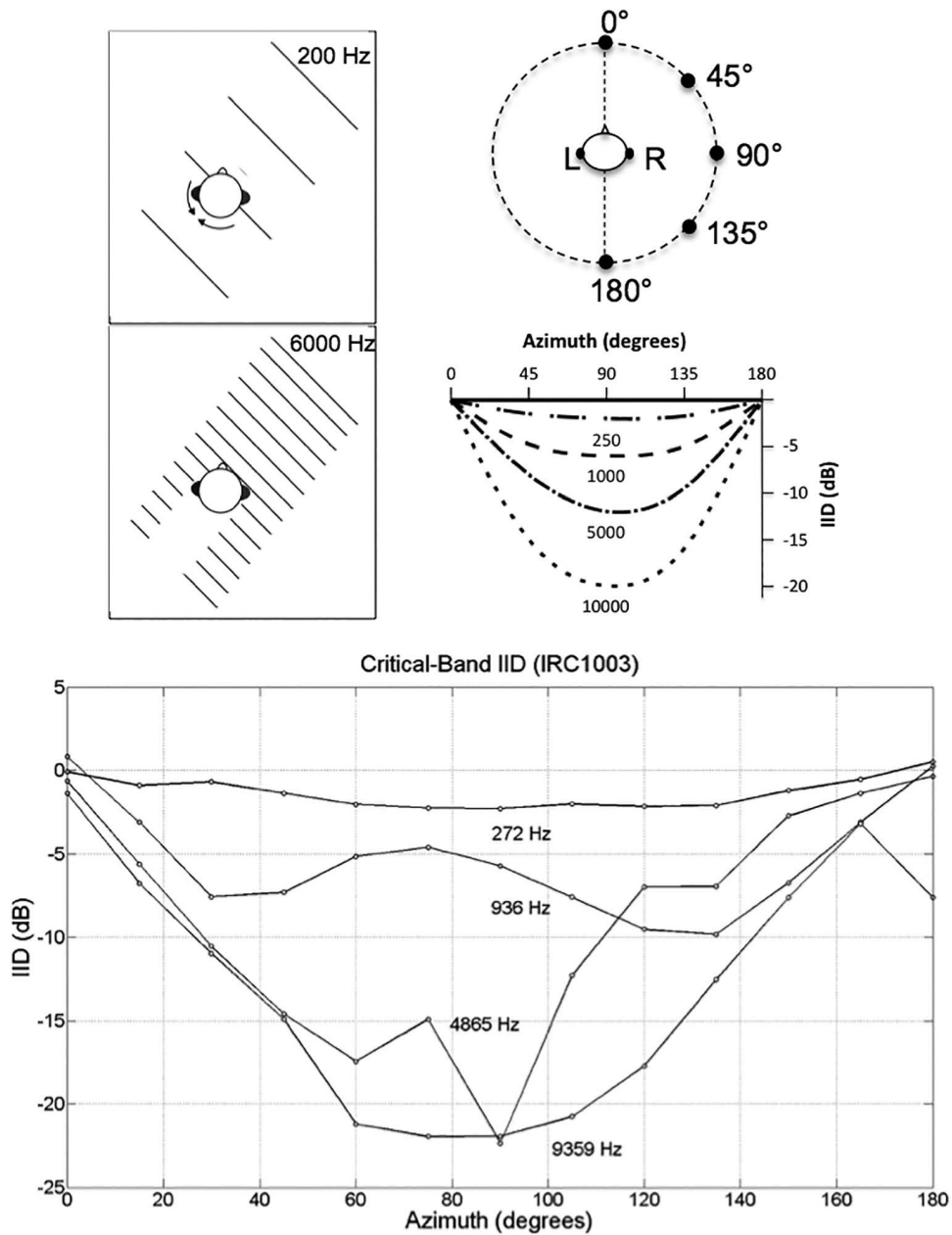


Figure 1.6 Top Panels: Illustration of the head-shadow effect and the resulting interaural intensity differences (IIDs) for 70 dB SPL tones of different frequencies plotted as a function of the azimuth of the source location (adapted from Gulick, Gescheider & Frisina, 1989, p. 324). Bottom Panel: IIDs as a function of critical-band center-frequency and azimuth are plotted for measured data at 0° elevation for subject IRC_1003 of the IRCAM Listen database. Critical bands were chosen to parallel the illustration in the top panel and were computed using the method described in Miller, Godfroy-Cooper and Wenzel (2014).

Independent of frequency content, variations in the overall difference between left and right intensity levels at the eardrum are interpreted as changes in the sound source position from the perspective of the listener. Consider the primary spatial auditory cueing device built into stereo recording consoles, the panpot (short for “panoramic potentiometer”). Over headphones, the panpot creates IIDs without regard to frequency, yet it works for separating sound sources in most applications. This is because the frequency content of typical sounds includes frequencies above and below the hypothetical “cut-off” points for IID and ITD, and listeners are sensitive to IID cues for localization across most of the audible frequency range, down to at least 200 Hz (Blauert, 1997).

Binaural research over the last few decades, however, points to serious limitations of the duplex theory. For example, for nearby sources, the IID is available even at low frequencies (Shinn-Cunningham, 2000). Similarly, it is known that ITD cues based on the relative timing of the amplitude envelopes (“envelope ITD”) of high-frequency sounds can be used by a mechanism such as interaural coincidence detection (Henning, 1974, 1980; van de Par & Kohlrausch, 1997; Bernstein & Trahiotis, 2010). Finally, in theory, the azimuth of a sound source can also be determined monaurally because the high-frequency components of the sound are more attenuated compared to low-frequency components as the sound source moves contra-laterally (Shub, Durlach & Colburn, 2008).

Sound Localization in the Vertical Dimension

The duplex theory cannot account for the ability of subjects to localize sounds on the vertical median plane (directly in front of the listener), where interaural cues are minimal. Similarly, when subjects listen to stimuli over headphones, the sounds are perceived as being lateralized inside the head even though interaural temporal and intensity differences appropriate to an external source location are present.

The results of many studies now suggest that these deficiencies of the duplex theory reflect the important contribution to localization of the direction-dependent filtering that occurs when incoming sound waves interact with the outer ears or pinnae and other body structures such as the shoulders and torso.

The main cue the human auditory system uses to determine the elevation of a sound source is the monaural spectrum determined by the interaction of the sound with the pinnae (Wightman & Kistler, 1997). However, small head asymmetries may provide a weak binaural elevation cue. Specifically, there is a spectral notch that moves in frequency from approximately 5 kHz to 10 kHz as the source moves from 0° (directly ahead of listener) to 90° (above the listener’s head) that is considered to be the main elevation cue (Musicant & Butler, 1985; Moore, Oldfield & Dooley, 1989). As sound propagates from a source to a listener’s ears, reflection and refraction effects tend to alter the sound in subtle ways and the effect depends on frequency. For example, for a particular location, a group of high-frequency components centered at 8 kHz may be attenuated more than a different band of components centered at 6 kHz. Such frequency dependent effects or filtering also vary greatly with the direction of the sound source. Thus, for a different source location, the band at 6 kHz may be more attenuated than the higher frequency band at 8 kHz. It is clear that listeners use these kinds of frequency-dependent effects to discriminate one location from another. Experiments have shown that spectral shaping by the pinnae is

highly direction-dependent, that the absence of pinna cues degrades localization accuracy, and that pinna cues are partially responsible for externalization or the “outside-the head” sensation (Gardner & Gardner, 1973; Oldfield & Parker, 1984a, b; Plenge, 1974; Shaw, 1974).

Other monaural cues are provided by the ratio of direct-to-reverberant energy that expresses the amount of sound energy that reaches our ears directly from the source versus the amount that is reflected off the walls in enclosed spaces (Larsen et al., 2008). In general, monaural cues are more ambiguous spatial cues than binaural cues because the auditory system must make *a priori* assumptions about the acoustic features of the original sound in order to estimate the filtering effects corresponding to the monaural spatial cues. Environmental cues will be discussed in more detail in a later section.

Factors Affecting Localization Performance

Localization performance generally refers to the degree of accuracy with which listeners can identify and/or discriminate the location of a sound source. It may be measured using a variety of experimental paradigms with different types of response measures. For example, in a direct localization task a sound source (real, recorded or virtual) is presented to a listener who is asked to report its location, perhaps in terms of estimates of azimuth, elevation and distance.

Several kinds of error are usually observed in perceptual studies of localization when listeners are asked to judge the position of a static sound source in the free field. One, which Blauert (1997) refers to as localization blur, is a relatively small error in resolution on the order of about 5° to 20°. A related measure of localization accuracy is the minimum audible angle (MAA), the minimum detectable angular difference between two successive sound sources. MAAs increase from about 1° for a sound directly ahead, to 20° or more for a sound directly to the right or left (Mills, 1958; Perrott & Saberi, 1990). The minimum audible movement angle (MAMA) is the minimum detectable angular difference of a continuously moving sound source (Perrott & Tucker, 1988). The MAMA depends on the speed of the moving sound source and ranges from about 8° for a velocity of 90°/s to about 21° for a velocity of 360°/s.

Another class of error observed in nearly all localization studies is the occurrence of front-back “reversals” (Figure 1.7, right). These are judgments indicating that a source in the front hemisphere was perceived by the listener as if it were in the rear hemisphere. Occasionally, back-to-front confusions are also found (e.g., Oldfield & Parker, 1984a, b). Confusions in elevation, with up locations heard as down, and vice versa, have also been observed (Wenzel, 1991).

Although the reason for such reversals is not completely understood, they are probably due in large part to the static nature of the stimulus and the ambiguities resulting from the so-called cone of confusion (Woodworth, 1938; Woodworth & Schlosberg, 1954; Mills, 1972). Assuming a stationary, spherical model of the head and symmetrically located ear canals (without pinnae), a given interaural time or intensity difference will correlate ambiguously with the direction of a sound source, with a conical shell describing the locus of all possible sources (Figure 1.7, left). Intersection of these conical surfaces with the surface of a sphere results in circular projections corresponding to contours of constant ITD or IID (i.e., considering sources at an arbitrary fixed distance). While the rigid sphere model is not the whole story, the observed pattern of such iso-IDT and iso-IID contours indicates that the interaural characteristics of the stimulus

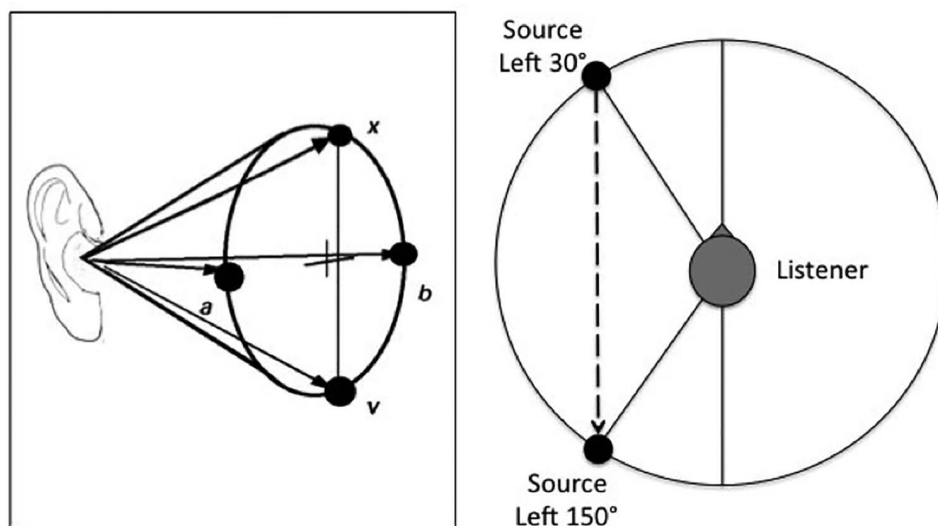


Figure 1.7 Cone of confusion effects (left panel) cause perceptual reversals in location for sound sources with identical or near-identical ITDs or IIDs (right panel).

are inherently ambiguous. In the absence of other cues, both front-back and up-down reversals would seem to be quite likely.

Several cues are thought to help in disambiguating the cones of confusion. One is the complex spectral shaping provided by the pinnae as a function of location that was described above. For example, because of the orientation and shell-like structure of the pinnae, high frequencies tend to be more attenuated for sources in the rear than for sources in the front [e.g., see Blauert's (1997) discussion of “boosted bands”, pp. 111–116]. For the case of static sounds, such cues would essentially be the only clue to disambiguating source location. With dynamic stimuli, however, the situation improves greatly. A variety of studies have shown that allowing listeners to move their heads substantially improves localization ability and can almost completely eliminate reversals (e.g., Wallach, 1939, 1940; Thurlow & Runge, 1967; Fisher & Freedman, 1968; Wightman & Kistler, 1999; Begault, Wenzel & Anderson, 2001). With head-motion, the listener can apparently disambiguate front-back locations by tracking changes in the magnitude of the interaural cues over time; for a given lateral head movement, ITDs and IIDs for sources in the front will change in the opposite direction compared to sources in the rear (Wallach, 1939, 1940). Time-varying cues provided by moving sources may also aid in disambiguation, particularly if there is *a priori* knowledge about the direction of motion (Wightman & Kistler, 1999), although relatively little research has been done on the topic of source motion in this context.

In addition to the primary localization cues, the localizability of a sound also depends on other factors such as its spectral content: narrowband (pure) tones are generally difficult to localize while broadband, impulsive sounds are the easiest to locate. A closely related issue in the localizability of sound sources is their degree of familiarity. Logically, localization based on spatial

cues other than the interaural cues, e.g., cues related to spectral shaping by the pinnae, is largely determined by a listener's *a priori* knowledge of the spectrum of the sound source. The listener must “know” what the spectrum of a sound is to begin with to determine that the same sound at different positions has been differentially shaped by the effects of his or her ear structures. Thus both the perception of elevation and relative distance, which depend heavily on the detection of spectral differences, tend to be superior for familiar signals like speech (e.g., Plenge & Brunschen, 1971, in Blauert, 1997, p. 104; Coleman, 1963). Similarly, spectral familiarity can be established through training (Batteau, 1967).

In an acoustical environment multiple acoustic objects can be present and room reverberation can also distort spatial cues. When the listener is in a room or other reverberant environment the direct sound received at the ears is combined with multiple copies of the sound reflected off the walls before arriving at the ears. Reverberation alters the monaural spectrum of the sound as well as the IIDs and IPDs of the signals reaching the listener (Shinn-Cunningham, Kopco & Martin, 2005). These effects depend on the source position relative to the listener as well as on the listener position in the room. On the other hand, the ratio of direct to reverberant energy itself can provide a spatial cue.

Such phenomena suggest that the auditory system exhibits neural plasticity, adapting over time in response to perceptual experience and changes in the sensory environment. Neural plasticity will be discussed in a later section.

Head-Related Transfer Functions (HRTFs) and Virtual Acoustics

The data on the primary localization cues suggest that perceptually veridical localization over headphones is possible if the spectral shaping by the pinnae and other body structures as well as the interaural difference cues can be adequately reproduced in a 3D sound system or virtual acoustic display. There may be many cumulative effects on the sound as it makes its way to the eardrum, but it turns out that all of these effects can be expressed as a single filtering operation much like the effects of a graphic equalizer in a stereo system. The exact nature of this filter can be measured by a simple experiment in which an impulse (a single, very short sound pulse or click) or other broadband probe stimulus is produced by a loudspeaker at a particular location. The acoustic shaping by the two ears is then measured by recording the outputs of small probe microphones placed inside an individual's ear canals (Figure 1.8). If the measurement of the two ears occurs simultaneously, the responses, when taken together as a pair of filters, include an estimate of the interaural differences as well. Thus, this technique allows one to measure all of the relevant spatial cues together for a given source location, a given listener and in a given room or environment.

The bottom panel of Figure 1.3 illustrates these effects for the transfer functions of the ears. The illustration on the left shows the frequency domain representation, derived from a mathematical operation known as the Fourier Transform, of an acoustic impulse in the time domain before interaction with the outer ear (and other body) structures. The illustration on the right shows what happens to the frequency response of an impulse delivered from a loudspeaker located directly to the right of a listener after interaction with the outer ear structures, as measured in the left (solid line) and right (dashed line) ear canals of a listener. The differences between the left and right intensity curves are the IIDs at each frequency. Spectral phase effects (frequency-dependent

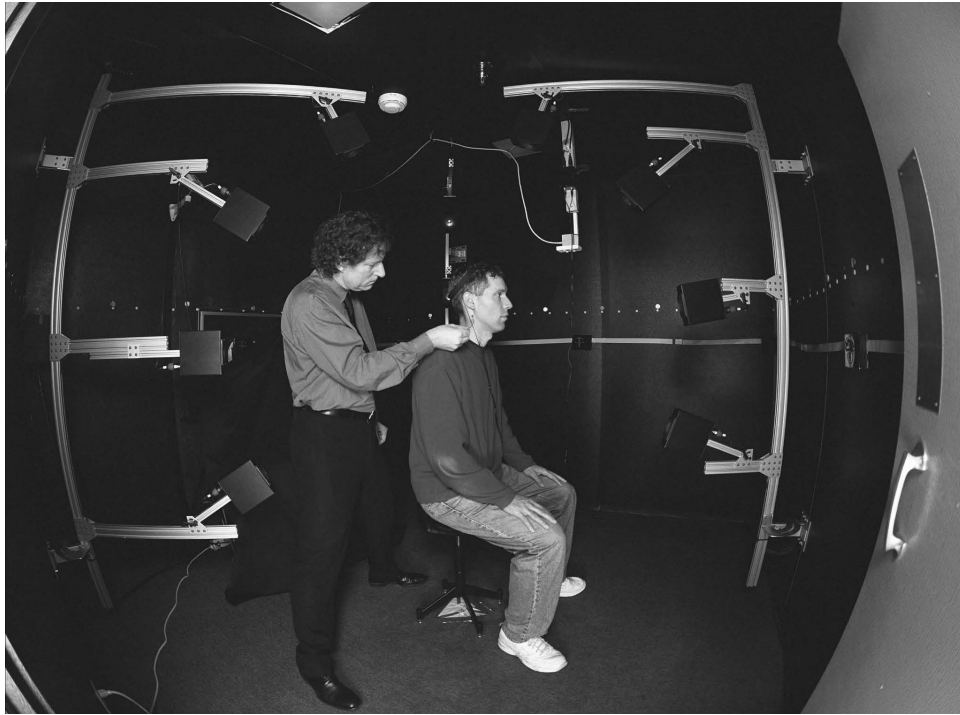


Figure 1.8 Facility at the NASA Ames Spatial Auditory Displays Lab for measuring HRIRs. The 12-speaker system in a double-walled soundproof booth can measure 432 locations at 10° intervals. The measurement signal is a Golay code (Foster, 1986) and responses are “quasi-anechoic”, i.e., responses are windowed to remove possible reflections. The impulse response (down-sampled from 96 kHz to 44.1 kHz sample rate) is high-pass filtered at 700 Hz to avoid the influence of reflections. The amplitudes of frequencies below 700 Hz are not significantly affected by the head due to diffraction. The system calculates a frequency-independent ITD that is applied to the measured (minimum phase) HRIR that includes a flat extension of the frequency response below 700 Hz.

phase, or time delays) are also present in the measurements, but are not shown here for clarity. The filters constructed from these ear-dependent characteristics are examples of Finite Impulse Response (FIR) filters and are often referred to as Head-Related Impulse Responses (HRIRs) in the time domain, and Head-Related Transfer Functions (HRTFs) in the frequency domain. Filtering in the frequency domain is a point-by-point multiplication operation while filtering in the time domain occurs via a somewhat more complex operation known as convolution [see Brigham (1974) for a useful pictorial discussion of filtering and convolution]. By filtering an arbitrary sound with these HRTF-based filters, it is possible to impose spatial characteristics on the signal such that it apparently emanates from the originally measured location. If the filtering occurs in real-time, the effects of source motion and the listener’s head motion can also be simulated.

Figure 1.9 provides examples of the magnitude responses for the left and right ears derived from measured HRTFs. The top panels show the frequency responses for azimuths of -90° , 0° and $+90^\circ$ at 0° elevation. Note that one can see how the magnitude responses are similar at 0° azimuth and are larger in the ipsilateral ear for the -90° (left ear) and $+90^\circ$ (right ear) source locations. The difference between the left and right ear responses represents the IIDs as a function of frequency. The overall IIDs averaged across frequency, as well as the ITDs, tend to be similar between individual subjects. Consequently, accuracy in azimuth perception is generally observed to be reasonably comparable when listening to stimuli generated from either individualized (one's own) or non-individualized HRTFs (Wightman & Kistler, 1989; Wenzel et al., 1993). The bottom panels show the frequency responses for elevations of -45° , 0° and $+45^\circ$ at $+45^\circ$ azimuth. Note that one can see how the center frequency of the notches in the magnitude spectra shift toward different frequencies as the elevation moves from -45 to $+45^\circ$ in the ipsilateral (right ear). Such frequency notches are thought to be a primary cue for elevation (Hebrank & Wright, 1974;

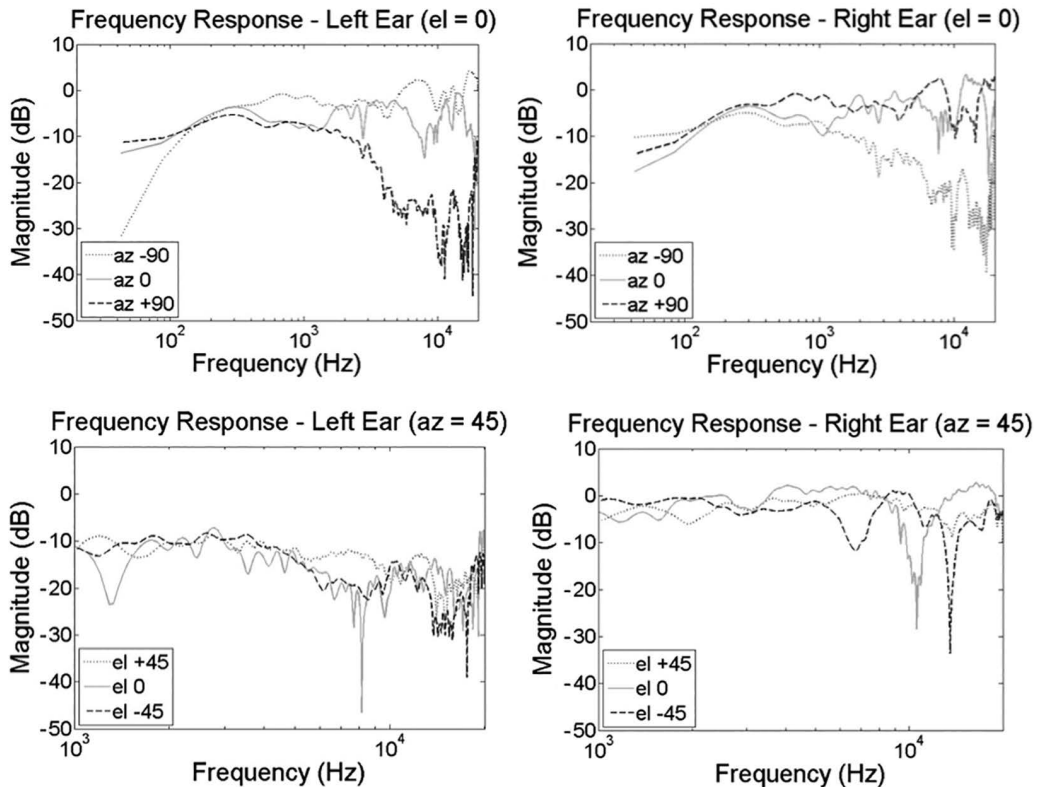


Figure 1.9 Examples of the magnitude responses for the left and right ears derived from measured HRTFs (subject IRC_1003 of the IRCAM Listen database). The top panels show the frequency responses for azimuths of -90° , 0° and $+90^\circ$ at 0° elevation. The bottom panels show the frequency responses for elevations of -45° , 0° and $+45^\circ$ at $+45^\circ$ azimuth.

Middlebrooks, 1992). Since these notch locations are highly dependent on specific pinna structures their particular frequency location may vary greatly between individuals. Thus, elevation perception is generally observed to be more accurate when listening to stimuli generated from individualized HRTFs.

It should be noted that the spatial cues provided by HRTFs, especially those derived from simple anechoic (free-field or echoless) environments, are not the only cues likely to be necessary to achieve veridical localization in a virtual display. Anechoic simulation is merely a first step, allowing a systematic study of the technological requirements and perceptual consequences of synthesizing spatial cues by using a less complex, and therefore more tractable, stimulus. The section on Distance and Environmental Context Perception will discuss the impact of environmental cues on localization and the perception of distance and immersion in more detail.

Neural Plasticity in Sound Localization

In natural environments, neural systems must be continuously updated to reflect changes in sensory inputs and behavioral goals. Recent studies of sound localization have shown that adaptation and learning involve multiple mechanisms that operate at different timescales and stages of processing, with other sensory and motor-related inputs playing a key role. In recent years, studies on sound localization have provided evidence to support the view that neural processing is rapidly updated to reflect changes in sensory conditions.

Spatial Cue Remapping

A popular approach to studying the plasticity of auditory spatial processing has been to reversibly alter the relation between stimulus location and the binaural cues available. This can be easily achieved by occluding one ear so that the acoustical input is attenuated and delayed, thereby changing the IID and ITD values corresponding to each direction in space. In the barn owl, this procedure leads to adaptation to the abnormal binaural cues, with neurons shifting their sensitivity to these cues in a way that compensates for the effects of monaural occlusion (Gold & E. I. Knudsen, 2000; E. I. Knudsen, P. F. Knudsen & Esterly, 1984). Recent work has shown that mammals also possess the ability to developmentally remap spatial position onto abnormal IIDs, which can be observed both behaviorally and in the responses of neurons in the primary auditory cortex (Keating, Dahmen & King, 2015). This capacity to accommodate altered spectral cues is not restricted to development and adult humans can learn to localize accurately using altered binaural (Bauer et al., 1966; Mendonça et al., 2013) or spectral localization cues (Hofman, Van Riswick & Van Opstal, 1998; Hofman & Van Opstal, 1998; Carlile & Blackman, 2014; Majdak, Walder & Laback, 2013). Interestingly, no aftereffect is seen in adult humans following adaptation to altered spatial cues, implying that different sets of spatial cues can be mapped onto the same location (Hofman et al., 1998; Hofman & Van Opstal, 1998; Carlile & Blackman, 2014; Majdak et al., 2013).

Spatial Cue Reweighting

In situations where some, but not all, of the spatial cues are altered, an alternative form of plasticity to cue remapping is to down-weight the spatial information provided by the altered cues and

instead, to rely more on the cues that remain intact. In the specific case of monaural hearing loss, a number of studies have shown that sound localization behavior in mammals adapts both during development and adulthood by giving greater weight to the unchanged monaural spatial cues provided by the normal hearing ear (Agterberg et al., 2014; Kumpik, Kacelnik & King, 2010; Keating et al., 2013).

Importance of Behavioral Context

In addition to adapting to changes in the localization cues available, auditory spatial processing can be refined in situations where its behavioral importance is increased even if the acoustical input remains the same. Studies in humans have shown that training-induced improvement in spatial processing are specific to individual binaural cues, though cue specificity may be asymmetric, with one study showing that IID training generalizes to an ITD task, but not vice versa (Sand & Nilsson, 2014). Training in adulthood can even reverse the negative impact of abnormal developmental experience on sound localization accuracy and responses in the primary auditory cortex responses (Guo et al., 2012; Pan et al., 2011) and can improve sound localization performance in hearing-impaired populations (Firszt et al., 2015). Although training-dependent plasticity often takes place slowly, recent work indicates that feature-specific learning can occur rapidly in a task that involves spatial processing, which may reflect top-down biasing (Du et al., 2015).

Visual Influences on Auditory Spatial Plasticity

Sound sources are often visible as well as audible and the availability of visual information can improve the accuracy of sound localization estimates (Tabry, Zatorre & Voss, 2013) and even help to suppress echoes that are the consequence of listening in reverberant environments (Bishop, London & Miller, 2012). Binaural cue discrimination is also enhanced if subjects look toward the sound while keeping their head still (Maddox et al., 2014), adding evidence that eye position signals can modulate activity in the auditory system (Bulkin & Groh, 2012). Not surprisingly, auditory localization abilities can be altered if vision is impaired. The most commonly reported finding is that some blind individuals show superior auditory spatial perception relative to sighted control subjects (Hoover, Harris & Steeves, 2012; Lewald, 2013; Jiang, Stecker & Fine, 2014). Interestingly, as with adaptation to a unilateral hearing loss, more accurate sound localization in blind humans is associated with greater dependence on spectral cues (Voss et al., 2011). However, this superior use of spectral cues for localization in the horizontal plane appears to come at the cost of reduced ability to use these cues for localization in the vertical plane (Voss, Tabry & Zatorre, 2015).

Distance and Environmental Context Perception

Fundamental Cues to Sound Source Distance

The perceived distance of a sound source is mainly cued by the acoustic attributes of sound level and reverberation. Sound sources ordinarily reach a listener by not only an unobstructed wavefront (referred to as **direct sound**) but also by reflection off of and diffraction around nearby objects, such as the ground, walls or other surfaces (referred to as **indirect sound**, or