Diffuse Field Modeling

The Physical and Perceptual Properties of Spatialized Reverberation

David Romblom

Music Technology Area
Department of Music Research
Schulich School of Music
McGill University
Montréal, Quebec

August 2016

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

Abstract

This dissertation addresses methods of recording and reproducing the reverberant diffuse field for musical performances; it is equally applicable to virtual reality, gaming and telecommunications. A purely physical reproduction of a sound field might attempt to replicate identical pressure fields for as large an area as possible. A purely artistic reproduction may manipulate many aspects of the physical sound field to create compelling musical presentation for existing distribution formats such as 2-channel stereo. Intermediate to these approaches is a physically and perceptually plausible virtual acoustic model that can be manipulated for the manifold objectives of artists, researchers, and engineers.

Four contributions are made. The first is a perceptual evaluation of a room microphone technique known as the Hamasaki Square that is well accepted in the sound recording community. This technique has physical interpretation as a sparse sampling of a bounding surface and, as opposed to direct sound techniques, is intended for the reverberant diffuse field. The second is a perceptual threshold experiment demonstrating that the human auditory system is sensitive to directional differences in the reverberant diffuse field and that such differences can be found in practice. The results of both of these experiments are considered in the third contribution, which is a physically-inspired strategy to generate a large number of channels of diffuse reverberation from a single B-Format Room Impulse Responses (RIR). The fourth contribution is a perceptual evaluation of the technique for a large 20-loudspeaker array and a comparison to the Hamasaki Square for a 5.1 loudspeaker array.

Abrégé

Cette thèse porte sur les méthodes d'enregistrement et de reproduction du champ réverbérant diffus lors de représentations musicales dont les résultats peuvent aisément s'appliquer aux domaines de la réalité virtuelle, des jeux et des télécommunications. D'un point de vue scientifique la reproduction physique du champ acoustique vise à simuler le champ de pression à l'identique sur une surface aussi grande que possible. Tandis que la reproduction d'un point de vue artistique vise à permettre la manipulation de nombreux aspects du champ sonore afin de créer une présentation musicale convaincante pour les formats de distribution existants, tels que la stéréophonie. Entre ces deux approches, il est intéressant de créer un modèle acoustique virtuel, à la fois physiquement et perceptivement plausible, et qui puisse être manipulable selon les objectifs divers d'artistes, de chercheurs et d'ingénieurs.

Les travaux de recherche de cette thèse ont donné lieux à quatre contributions. La première consiste en l'évaluation perceptive d'une technique d'enregistrement microphonique bien acceptée dans la communauté des ingénieurs du son et connue sous le nom de carré de Hamasaki. Cette technique qui peut s'interpréter d'un point de vue physique comme l'échantillonnage parcimonieux sur une surface limite est destinée, contrairement aux techniques de prise de son direct, à la prise du champ diffus réverbérant. La seconde est une expérience sur des seuils perceptifs démontrant que le système auditif humain est sensible à des différences directionnelles du champ réverbérant diffus, et que ces différences sont observées dans la pratique. Les résultats de ces deux expériences sont pris en compte dans la troisième contribution, qui met en place un système de synthèse de champ diffus. Ce système, fondé sur une approche physique, génère un grand nombre de canaux de réverbération diffuse à partir d'une seule réponse impulsionnelle de salle (RIS) de type B-Format. La quatrième contribution est une évaluation perceptive de cette stratégie pour un large réseau de 20 haut-parleurs ainsi qu'une comparaison avec la technique du carré de Hamasaki pour un système de diffusion sur haut-parleur en 5.1.

Dedication

To my dad.

Acknowledgements

I would like to thank my friends and advisors Philippe Depalle and Catherine Guastavino for their technical and strategic help throughout my Ph.D., as well as Richard King for being my interface to the sound recording world. I would also like to thank the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) for the outstanding facilities and technical staff. In particular, the perceptual evaluations done in this dissertation are operatic in scale, and I would to thank Yves Méthot, Harold Kilianski, and Julien Boissinot for their assistance. I would like to thank Sennheiser for funding the first two years of my Ph.D. and Natural Sciences and Engineering Research Council of Canada (NSERC) for grants with Audiokinetic, Inc and Applied Acoustic Systems. Within these collaborations, I would like to thank colleagues Juha Merimaa, Véronique Larcher, Xavier Buffoni, Benoit Alary, and Marc-Pierre Verge for a number of fruitful technical conversations.

I would like to thank my colleague Olli Rummukainen for a fun and fruitful collaboration in the perceptual evaluation of Diffuse Field Modeling (Chapter 5), Cédric Camier for a number of adventurous scientific and artistic conversations, and McGill Sound Recording students Brett Leonard and Scott Levine for their ears and feedback. I would like to thank Marshall Day Acoustics for providing the three-dimensional room impulse responses used in Chapter 3.

Most importantly, I would like to thank two central figures in my life. I would like to thank my long time partner Julie Bixby for having my back during the heavy parts of the Ph.D. My father Edward Romblom passed away in the year prior to my completion of this work. Without a decade of his help on my grade school homework, without hundreds of technical and philosophical discussions, and without his general energy and elbow grease, this dissertation would not exist.

Contributions of Authors

Chapter 2 is based on *A Perceptual Evaluation of Room Effect Methods for Multichannel Spatial Audio* by David Romblom, Richard King, and Catherine Guastavino presented at the Audio Engineering Society's 135th Convention in New York City. David designed and executed the experiment, performed the analysis of the data, and wrote the paper under the guidance of Richard King and Catherine Guastavino. We would like to acknowledge Sennheiser Technology and Innovation for funding this work and Scott Levine for helping to prepare the recordings used in the experiment. Additionally, we would like to thank CIRMMT staff Julien Boissinot, Yves Méthot, and Harold Kilianski for their assistance setting up an involved recording and reproduction system.

Chapter 3 is based on *Perceptual Thresholds for Non-Ideal Diffuse Field Reverberation* by David Romblom, Catherine Guastavino, and Philippe Depalle which was revised for the Journal of the Acoustical Society of America in the summer of 2016. David designed and executed the experiment, performed the analysis of the data, and wrote the paper under the guidance of Catherine Guastavino and Philippe Depalle. We would like to acknowledge Sennheiser Technology and Innovation for funding the initial aspects of this work and as well as CIRMMT staff Julien Boissinot, Yves Méthot, and Harold Kilianski for their assistance setting up (yet another) involved experimental system.

Chapter 4 is based on *Diffuse Field Modeling Using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part I Algorithm* by David Romblom, Philippe Depalle, Catherine Guastavino, and Richard King which was published in the Journal of the Audio Engineering Society in April of 2016. David designed the algorithm and performed the MATLAB simulations with technical assistance from Philippe Depalle. The paper was written under the guidance of Philippe Depalle, Catherine Guastavino, and Richard King.

Chapter 5 is based on *Diffuse Field Modeling Using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part II Evaluation* by Olli Rummukainen, David Romblom, and Catherine Guastavino which was published in the Journal of the Audio Engineering Society in April of 2016. The experiment and the writing of the paper were an equal collaboration between Olli and David under the guidance of Catherine Guastavino. Within this, the general direction of the experiments was defined by David, the experimental procedures were designed jointly, Olli lead the collection and analysis of the perceptual data, and David lead the preparation of the stimuli based on the algorithm presented in Chapter 4. We would like

to thank CIRMMT staff Julien Boissinot and Yves Méthot for their assistance setting up an involved reproduction system, to our colleague Olli for a fun and fruitful collaboration, and to Cédric Camier for sharing CIRMMT resources throughout this project.

Contents

Abstract ii
Abrégé iii
Dedication iv
Acknowledgements v
Contribution of Authors vi
List of Figures xiii
List of Tables xvi
Acronyms xviii

1 Introduction **1**
1.1 Overview of Problem 1
1.2 Scientific Context 3
1.3 Research Questions and Objectives 6
1.4 Structure of Dissertation 7

2 Perceptual Evaluation of Room Effect Methods for Multichannel Spatial Audio **11**
2.1 Introduction 11
2.1.1 Psychoacoustic Principles 12
2.1.2 Perceptual Attributes of Spatial Audio 13
2.1.3 Practice: 3/2 Stereo (5.1) and the Hamasaki Square 14
2.1.4 Research Questions 16
2.2 Methods 16
2.2.1 Treatments 16
2.2.2 Participants 18
2.2.3 Apparatus and Stimuli 18
2.2.4 Procedure 19

2.2.5 Analysis . . . 19
2.3 Results . . . 19
2.4 Discussion . . . 27
2.5 Conclusion . . . 28

Segue **29**

3 Perceptual Thresholds for Non-Ideal Diffuse Field Reverberation **31**
3.1 Introduction . . . 31
3.2 Experiment . . . 35
3.2.1 Participants . . . 35
3.2.2 Apparatus . . . 35
3.2.3 Stimuli . . . 36
3.2.4 Procedure . . . 39
3.2.5 Data Analysis . . . 39
3.2.6 Results . . . 40
3.3 Signal Analysis . . . 41
3.3.1 Dummy Head Recordings . . . 41
3.3.2 Three-Dimensional Room Impulse Responses . . . 43
3.3.3 Summary of Signal Analysis Results . . . 46
3.4 Discussion . . . 46
3.5 Conclusions and Future Work . . . 48

Interim Summary **51**

4 Diffuse Field Modeling using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part I Algorithm **55**
4.1 Introduction . . . 56
4.2 State of the Art . . . 59
4.2.1 Hamasaki Square . . . 59
4.2.2 B-Format Microphones . . . 60
4.2.3 Ambisonics and Wave Field Synthesis . . . 61
4.2.4 Existing Technology . . . 62
4.2.5 Non-Ideal Diffuse Fields . . . 63
4.3 Acoustic Background . . . 64
4.3.1 Sabine Reverberation Model . . . 65

4.3.2 Diffuse Field Model . . . 66
4.3.3 Modal Model . . . 68
4.3.4 Kirchhoff-Helmholtz Integral . . . 69
4.4 Diffuse Field Modeling . . . 71
4.4.1 Rationale . . . 71
4.4.2 Bank of Bandpass Filters . . . 73
4.4.3 Spatial Filtering . . . 76
4.5 Validation by Simulation . . . 79
4.5.1 Method . . . 79
4.5.2 Results . . . 81
4.6 Conclusions and Future Work . . . 85

Segue **87**

5 Diffuse Field Modeling using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part II Evaluation **89**
5.1 Introduction . . . 89
5.2 Background . . . 91
5.2.1 Recording, Rendering and Reproduction Strategies . . . 91
5.2.2 Psychoacoustics and the Perceptual Attributes of Spatial Audio . . . 92
5.2.3 Implementation of Virtual Acoustic Model using Diffuse Field Modeling 93
5.3 Comparative evaluation of different DFM treatments using a 20-channel array 95
5.3.1 Participants . . . 95
5.3.2 Apparatus . . . 96
5.3.3 Stimuli . . . 97
5.3.4 Procedure . . . 97
5.3.5 Results . . . 98
5.3.6 Discussion . . . 103
5.4 Comparative evaluation of DFM with common microphone techniques using 5.1 . . . 106
5.4.1 Participants . . . 106
5.4.2 Apparatus . . . 106
5.4.3 Stimuli . . . 107
5.4.4 Procedure . . . 108
5.4.5 Results . . . 109

5.4.6 Discussion . . . 114
5.5 Conclusions . . . 116

6 Conclusions **119**
6.1 Summary . . . 119
6.2 Research Answers . . . 121
6.3 Contributions . . . 122
6.4 Limitations and Future Work . . . 122

A Acoustics **127**
A.1 Basic Acoustics . . . 127
A.1.1 Acoustic Equations and the Wave Equation . . . 127
A.1.2 Constant Frequency Solutions . . . 131
A.1.3 Spherical Basis Functions and Multipoles . . . 133
A.2 Physical Aspects of Room Acoustics . . . 135
A.2.1 Diffuse Sound Fields . . . 135
A.2.2 Sabine-Franklin-Jaeger Theory of Reverberant Rooms . . . 136
A.2.3 Modal Theory of Room Acoustics . . . 138
A.2.4 Time Domain View of Reverberation . . . 141
A.2.5 Statistical Aspects of Room Acoustics . . . 142
A.3 Summary . . . 145

B Psychoacoustics **147**
B.1 Binaural Localization of Coherent Sources . . . 147
B.1.1 Interaural Time and Level Differences . . . 147
B.1.2 Inner Ear, Critical Bands, and Masking Effects . . . 148
B.2 Fundamental Psychoacoustics of Complex Environments . . . 150
B.2.1 The Precedence Effect and Early Reflections . . . 150
B.2.2 Coherence . . . 151
B.2.3 Distance Hearing . . . 153
B.3 Subjective Measures of Room Acoustics . . . 154
B.3.1 Auditory Source Width and Listener Envelopment . . . 154
B.3.2 Reverberance and Clarity . . . 156
B.4 Summary . . . 157

Bibliography **159**

List of Figures

2.1a Ratings for the Distance scale. The low bound was "near," the high bound was "far." 21

2.1b Main effects, interaction effects, and post-hoc tests for the Distance scale. . . 21

2.2a Ratings for the Depth scale. The low bound was "shallow," the high bound was "deep." 22

2.2b Main effects, interaction effects, and post-hoc tests for the Depth scale. . . . 22

2.3a Ratings for the Envelopment scale. The low bound was "low," the high bound was "high." 23

2.3b Main effects, interaction effects, and post-hoc tests for the Envelopment scale. 23

2.4a Ratings for the Localization scale. The low bound was "vague," the high bound was "precise." 24

2.4b Main effects, interaction effects, and post-hoc tests for the Localization scale. 24

2.5a Ratings Instrument Width scale. The low bound was "narrow," the high bound was "wide." 25

2.5b Main effects, interaction effects, and post-hoc tests for the Instrument Width scale. 25

2.6a Ratings for the Coloration/Transparency scale. The low bound was "colored," the high bound was "transparent." 26

2.6b Main effects, interaction effects, and post-hoc tests for the Coloration/Transparency scale. 26

3.1 Experimental apparatus used for the listening test. 37

3.2 Schematic diagram of lateral and frontal treatments. 38

3.3 Exemplar psychometric functions for participants in the lateral, height, and frontal conditions. 41

3.4 The left column shows the dB difference between the stimuli level above the threshold and the reference for both the right and left ears (L/R) for the height, frontal, and lateral cases. The right column shows the observed differences for IACC. The frequency axis is in Hz. A 2% MATLAB smoothing filter was used to improve readability. The HATS dummy head was placed in the position of the listener. 42

3.5 Frequency-dependent differences in dB power for opposing coincident cardioids for both the experimental stimulus (dashed line) and an RIR of a shoebox style concert hall with the early response removed (solid line). For all experimental conditions (lateral, height, and frontal), the level differences correspond to the stimulus intensity above the perceptual threshold. A 2% MATLAB smoothing filter was used to improve readability, and a description of the microphone signals is given in the text. 44

4.1 Simulated diffuse fields at 160 Hz created from the superposition of $Q = 1024$ plane waves of random direction, phase, and magnitude. 57

4.2 dB difference of the power spectral density of opposing coincident cardioids formed from a B-Format microphone placed approximately 2m high in the center of Pollack Hall at McGill University. 65

4.3 The function $\Gamma(m)$ (Equation 4.15) describes the frequency autocorrelation measured from the ensemble of I decorrelation filters, the right column shows the correspondence of the FAC value to physical RT60 values defined in Equation 77

4.4 Average channel correlation between adjacent loudspeakers for 64-channel (left panel) and 16-channel (center panel) arrays with radius 2m. 79

4.5 Simulation results for 40, 80, 160, and 320 Hz for 16 channels with an array radius equal to 3m. 82

4.6 The function $\beta_m(|\mathbf{r}_0|)$ describes the spatial autocorrelation measured in the reconstructed pressure fields. 83

4.7 Absolute value simulation results for 40, 80, 160, and 320 Hz for 64 channels with an array radius equal to 2m. 84

5.1 Loudspeaker array and listener position in the first experiment. 96

5.2 Mean scores for the extent of realism, timbral quality and spatial quality with the different treatments. 99

5.3 Mean scores for the extent of realism with the different treatments. 101
5.4 Correspondence analysis result when the ordination is constrained by the treatment. 104
5.5 Test setup in the second experiment. 107
5.6 Graphical user interface in the second experiment. 110
5.7 Mean scores for overall quality, depth of stage and sense of space. 112
5.8 Mean scores for depth of stage in the two halls with two musical excerpts. The bars represent the 95 % confidence intervals of the mean. 113
5.9 Perceived position of the cello in two halls with three methods. 114
5.10 Perceived positions of the band instruments in two halls with three methods. 115

List of Tables

2.1 Perceptual attributes of multichannel spatial audio used in this study. The definitions given to the participants are listed on the right. 14

3.1 Levels of gain reduction for three conditions. 40
3.2 Perceptual thresholds for the lateral, height, and frontal conditions at two significant digits. 41

5.1 Musical excerpts 97
5.2 Treatments and descriptions 97
5.3 Significant main effects, interaction effects, and post-hoc tests for extent of realism, timbral quality and spatial quality scales. 100
5.4 Significant main effects, interaction effects, and post-hoc tests for the extent of realism scale by excerpts. 100
5.5 Frequency of concepts 102
5.6 Musical excerpts 108
5.7 Significant main effects, interaction effects, and post-hoc tests for overall quality, depth of stage and sense of space scales. 111
5.8 Significant main effects and post-hoc tests for depth of stage scale by excerpt and hall. 111

Acronyms

ANOVA	Analysis of Variation.
ARRAY,DRY	Treatment using microphone array and no room effect.
ARRAY,HS	Treatment using microphone array and Hamasaki Square room effect.
ARRAY,SURR	Treatment using microphone array and Electronic Time Offset room effect.
ASW	Auditory Source Width.
BM	Basilar Membrane.
BRIR	Binaural Room Impulse Response.
CAD	Computer Aided Drafting.
CCA	Constrained Correspondence Analysis.
C80	Clarity.
dB	deciBel.
DirAC	Directional Audio Coding.
DFM	Diffuse Field Modeling.
EDT	Early Decay Time.
ERB	Equivalent Rectangular Bandwidth.
ETO	Electronic Time Offset (adapted from William's MMA.)
FAC	Frequency Auto-Correlation.
HATS	Head and Torso Simulator.
HRTF	Head-Related Transfer Function.
HS	Hamasaki Square.
IACC	Interaural Cross Correlation.
ILD	Interaural Level Delay.
ITD	Interaural Time Delay.
K/H	Kirchhoff Helmholtz (Integral.)

L/C/R	Left / Center / Right loudspeakers in the 5.1 (3/2 stereo) configuration.
LEV	Listener Envelopment.
LF	Lateral Fraction.
LS/RS	Left Surround / Right Surround loudspeakers in the 5.1 (3/2 stereo) configuration.
MMA	Multichannel Microphone Array.
OFC	Outward-Facing Cardioids.
ORTF	Office de Radiodiffusion Télévision Française.
PSD	Power Spectral Density.
RIR	Room Impulse Response.
RNG	Random Number Generator.
RT60	Reverberation Time (to decay 60 deciBels from initial level.)
SAC	Spatial Auto-Correlation.
SBF	Spherical Basis Functions.
SHF	Spherical Harmonic Functions.
SIRR	Spatial Impulse Response Rendering.
SPL	Sound Pressure Level.
SPOT,DRY	Treatment using spot microphones, WFS panner, and no room effect.
SPOT,HS	Treatment using spot microphones, WFS panner, and Hamasaki Square effect.
STFT	Short-Time Fourier Transform.
TS	Center Time.
VBAP	Vector-Base Amplitude Panning.
WFS	Wave Field Synthesis.
WGN	White Gaussian Noise.

Chapter 1

Introduction

1.1 Overview of Problem

The goal of spatial audio is natural, immersive, and aesthetically malleable virtual auditory experiences; it has applications in music, cinema, gaming, virtual reality, telecommunications, perceptual research, hearing research, and hearing aides. It is based on a number of sub-disciplines including physical acoustics, psychoacoustics, signal processing, and sound recording. The research presented in this dissertation pertains to the perception and reproduction of reverberation, and, in particular, the development of a signal processing technique that allows one to systematically generate a statistical approximation of reverberation for arbitrary loudspeaker arrays. While artificial reverberation has a fifty year history [75] spanning back to Schroder's first recirculating networks [68], it is the perceptual and physical attributes of reverberation's spatial properties that will be focused on here. The work was done in the context of music recording and reproduction, however it is extensible to any of the applications listed above.

Much of the research presented in this dissertation was informed by interaction with the McGill Sound Recording community. A number of different approaches can be used for the recording and reproduction of sound scenes, and sound recording engineers will often use a blend of techniques depending on the logistical constraints of a particular venue or the aesthetics of the musical work being presented.

Channel-based array techniques attempt to capture an entire "live" acoustic scene using a number of specially-configured microphones. The microphone channels may be routed directly to specific loudspeakers [29][73][79] or combinations of the microphones may be used [8][39][41]. It is almost always the case that different strategies are used for the direct sound

and the reverberation [73]. An issue with such channel-based direct techniques is that they are bound to the loudspeaker format they were designed for, adaption to other formats is non-systematic, and it is logistically forbidding to have a corresponding microphone for every loudspeaker of a large array. Scene-based array techniques such as Higher-Order Ambisonics also attempt to capture an entire acoustic scene but instead project it onto a channel-independent representation that can be "decoded" to various loudspeaker arrays [39][28][50].

Alternatively, spot microphones may be placed in close proximity to the instruments or the source may be synthesized electronically. These may then be reproduced as point sources using amplitude panning for a channel-based format such as 2-channel stereo or 5.1 (3/2 stereo)[6], using vector-base amplitude panning (VBAP) for large but sparse loudspeaker arrays [57], or using wavefront reconstruction techniques for dense arrays [1]. To avoid an unnatural anechoic presentation, early reflections can be simulated with propagation delay and the diffuse reverberation can be simulated with an artificial reverberator. Object-based representations use a geometric description in tandem with the above tools to synthesize an acoustic scene for various loudspeaker arrays.

A common technique for capturing diffuse reverberation is the use of a pair of room microphones placed near the back of the venue. This room pair can then be adapted "by ear" to various loudspeaker configurations using equalization and multi-tap delays. Recirculation-based artificial reverberators [35][68][75] are commonly used to augment the adaption. These approaches will preserve and potentially enhance the temporal aspects of the frequency-dependent reverberation time (RT60). From a physical perspective, there is considerable ambiguity in the assumptions regarding the statistical relationship between output channels (see Cook [15] and Chapter 4) and with respect to directional variation in the diffuse field (Chapter 3). From a perceptual perspective, it is not known what is gained or lost when a physical representation of the reverberant diffuse field is preserved or discarded (Chapters 2 and 5).

A systematic tool for the recording and reproduction of spatial reverberation is a major gap in the current toolset. Such a tool should allow spatial content creation in the current sound recording paradigm, be extensible to dense loudspeaker arrays such as those found in Wave Field Synthesis (WFS), and additionally be extensible to three dimensional loudspeaker arrays. More broadly, systematic adaption to different loudspeaker configurations would make it a suitable technological component of a high-definition object-based transmission format. The contributions of this dissertation are an assessment of the state of the art, an investigation into directional variation in diffuse reverberation, the algorithm development

of a systematic tool for creating physically-plausible channels of reverberation, and, finally, the validation of that algorithm.

1.2 Scientific Context

Both the art of sound recording and the engineering discipline of spatial audio are inextricably linked to physical acoustics and auditory perception. Chapters 2, 3, 4, and 5 each contain literature reviews appropriate to their publication at conferences or in journals. Appendix A gives a complete view of the physical acoustics used in this dissertation while Appendix B gives a complete view of the appropriate psychoacoustic literature. The current section gives a broad overview of the scientific underpinnings that will be used throughout this work.

Physical acoustics pertains to the measurement of small amplitude vibrations in solids, fluids or gases [22][53]; this dissertation focuses on those in the ambient atmospheric pressure due to sources bound within an enclosure. Acoustic phenomena are described mathematically as solutions to the linear wave equation which is derived from the conservation of mass, Newton's 2nd Law, and the state relationship between acoustic pressure and density (A.1.1). Two commonly encountered solutions to this equation are the plane wave and the spherical wave (A.1.2)[53].

Psychoacoustics is the study of the human auditory system's response to physical acoustics [6][48]. An example of this is the auditory system's estimate of an acoustic source's azimuth using time and level differences between the two ears. Another example is the estimate of the elevation using a priori knowledge of the spectral shaping imposed by the acoustic structure of the ear and torso (B.1.1). The localization of azimuth and elevation pertains to planar (or nearly planar) acoustic sources that are deterministic in the sense that knowledge of their value at one time and point in space very accurately describes their behavior at different points in time and space [53].

Typical acoustic sources radiate in all directions, when bound inside an enclosure some of this energy will be reflected off of the walls. The time domain view of reverberation considers acoustic energy to travel in narrow, nearly planar beams and is the gradual spreading of temporally compact energy into temporally spread energy. Absorptive materials change the frequency characteristics of each reflection while scattering materials "break up" the reflection's wavefront (A.2.4)[7][21][22][53]. The Sabine reverberation model assumes that acoustic energy is approximately the same in all spatial regions of the room and that the absorption due to the bounding walls can be treated "on average." It results in frequency-dependent

exponentially decaying energy that is described by the reverberation time RT60. It is a meaningful approximation to an extremely complicated physical scenario and was substantiated empirically (A.2.2)[53]. The modal approach to room acoustics is derived by finding solutions to the wave equation in a rigidly bounded room and results in a "considerably less approximate" substantiation of the Sabine model (A.2.3)[53].

The mathematical diffuse field model is comprised of many plane waves in many directions and approximates what happens when a source is reflected many times in an enclosure (A.2.1)[22][53]. It is perceptually related to an auditory event heard "everywhere" (B.2.2)[6] and is a foundational concept in the statistical description of reverberation (A.2.5)[53]. This stochastic description does not attempt to predict exact acoustic field values, but rather details their probability distribution and autocorrelation with respect to time, frequency, and space. The perceptual estimation of a source's distance is thought to consider a number of physical properties, one notable property being the ratio between the direct sound (deterministic) and the (stochastic) reverberant diffuse field (B.2.3)[6][48][80].

Spherical basis functions are factored solutions to the wave equation that describe variation in azimuth, elevation, and distance (A.1.3)[28][50]. An orthogonal series of spherical basis functions can be used to decompose an arbitrary pressure field in a manner similar to the Fourier Series being used to decompose a periodic time or space signal. Multipoles are a closely related concept and are commonly used to describe a source or microphone as superposition of monopoles [22][28][53]. The Kirchhoff Helmholtz Integral states that if the pressure (monopole) and pressure gradient (dipole) are known on the bounding surface then the field within the surface can be reconstructed using a secondary array of dipole and monopole sources (4.3.4)[28][53]. Spherical basis functions and the Kirchhoff Helmholtz Integral are linked by the fact that an infinite series of spherical basis functions can be used to predict the pressure on a bounding surface and can be differentiated with respect to the radius to find pressure gradient [18].

A number of limitations are encountered in practice. The first of these is the truncation of the spherical basis series to very low order which causes reconstruction error at distances away from the origin and which smooths the variation with respect to azimuth and elevation [28][50]. In the case of the Kirchhoff Helmholtz Integral, the two "layers" must be collapsed into a single layer and the surface bounding the acoustic field is reduced to discrete points of reception and reradiation. This causes error in certain regions of the reconstructed field and limits the spatial frequencies that can be reproduced accurately [1].

The observation of an acoustic field at a point (extrapolated to a bounding surface) and

on a bounding surface (reradiated from a similar surface in a listening space) underlay the major schools of sound field recording and reproduction. Various techniques hold more or less closely to a physical representation. For instance, the Kirchhoff Helmholtz Integral is the conceptual underpinning of the Steinberg and Snow's "acoustic curtain" [34] and of Wave Field Synthesis (4.2.3)[1]. While substantially more sparse, William's MMA is a "direct" microphone technique that routes a microphone in the recording space to a loudspeaker in a similar configuration in the listening space of a 5.1 system [79]. B-Format and high-order spherical microphones such as the Eigenmic attempt to capture a low-order series of spherical basis functions and can be used to reconstruct a sound field using the mode matching inverse matrix of Ambisonics (4.2.2)(4.2.3)[39][50][47]. Coincident microphones such as Blumlein, XY, and Mid / Side [8][16][71] are difficult to interpret in a strict physical sense but do share the point / surface paradigm.

It is common practice for sound recording engineers to use differing microphone strategies for the direct sound and the reverberant field [73] and can be seen as a practical manifestation of the divide between strategies for deterministic and stochastic fields. Optimized Cardioid Triangle and the Decca Tree are strategies to capture the direct (deterministic) sound [73], while the Hamasaki Square is a room effect technique that captures very little direct sound and, to emphasize the reverberant diffuse field, is typically placed some distance from the sources (4.2.1)[29]. Like William's MMA, the Hamasaki Square can be viewed as a very sparse sampling of a boundary surface in the recording space routed to similarly-placed secondary radiators in the listening space.

The point source techniques used in wave field synthesis (WFS) [1] can be seen as the simulation of deterministic acoustic values in a single-layer approximation of the Kirchhoff Helmholtz Integral. The perceptual phenomenon of summing localization is the accepted view of amplitude panning (B.2.1)[6] or vector-base amplitude panning (VBAP) [57] and can be interpreted as the perceptual simulation of deterministic values on a segment of a sparsely sampled bounding surface. Artificial reverberators such as the feedback delay network [35] have frequency-dependent spectral envelopes applied to a temporal statistical process and can be viewed as simulating the stochastic field values for common formats such as 2-channel stereo or 5.1.

The contributions of this dissertation pertain to the simulation of stochastic acoustic values on a single layer approximation of the Kirchhoff Helmholtz Integral. The first is the perceptual evaluation of the Hamasaki Square which was chosen due to its wide acceptance and because the technique can be physically interpreted as a sparse sampling of a bounding

surface in the reverberant diffuse field. The second is an auditory discrimination experiment demonstrating that the human auditory system is sensitive to directional energetic differences in the reverberant diffuse field and further showing that such physical differences can be found in practice. The results of both of these experiments are considered in the third contribution, which is a physically-inspired strategy to generate a large number of channels of diffuse reverberation from a single specially treated B-Format Room Impulse Responses (RIR). The RIR contains temporal information regarding the reverberant field. Similar to the dichotomy between spherical basis functions / multipoles and the Kirchhoff Helmholtz Integral described above, the strategy will use the limited spatial information provided by the multipole observation (B-Format) to simulate acoustic values on a bounding surface (virtual microphone array to real loudspeaker array.) The fourth contribution is a perceptual evaluation of the technique for a large 20-loudspeaker array and a comparison to the Hamasaki Square for a 5.1 stereophonic loudspeaker array.

1.3 Research Questions and Objectives

Sound recording engineers have developed strategies to capture and reproduce an acoustic environment; their focus tends to be on the aesthetic presentation of a musical work as opposed to a physical representation of the sound field. The Hamasaki Square is a "direct" microphone technique that has physical interpretation as a sparse sampling of a bounding surface. It provides plausible channel correlation, a facsimile of early reflections, and independence from the direct sound capture. German tonmeister Gunther Theile notes its broad acceptance in conjunction with a delay plan and a "natural" mixing approach [73]. Given the physical interpretation and the acceptance in practice, the first research question is to understand what the perceptual qualities of this technique are by way of a comparative perceptual evaluation.

Both the physical and perceptual literature typically assumes the reverberant diffuse field to have equal energy from all directions. It is plausible, however, that this energetic distribution could vary in acoustic scenarios with vastly differing wall materials, elongated shapes, or with opening such as windows. It is also plausible that this be part of the spatial impression of musical listening spaces such as cathedrals and would be of substantial utility for virtual reality and gaming scenarios. The second research question is to determine if the human auditory is system sensitive to directional energetic differences in the late field reverberation. It is also desired to know if these differences can be observed in existing

acoustic spaces and what the corresponding ear signals are.

The third research objective is to develop an algorithm that allows the systematic creation of physically-plausible diffuse energy. It is desired that this algorithm be well-substantiated by known physical acoustics and that the algorithm take the above findings into account. As opposed to channel-based approaches such as the Hamasaki Square, it is desired to be independent from the loudspeaker format and to be able to systematically render a large number of channels. The algorithm should be validated by confirming that the reproduced fields (or simulations of those fields) correspond to expected statistical relationships.

The fourth research question is to validate the developed algorithm from a perceptual standpoint through listening tests with trained listeners. In particular, it is desired to know if it is possible to create 20 channels of reverberation for a large loudspeaker array without objectionable artifacts such as coloration or temporal smearing. Finally, for a sparse 5.1 array, it is desired to evaluate the algorithm against standard microphones techniques including the Hamasaki Square.

1.4 Structure of Dissertation

Chapter 2 compares the Hamasaki Square to an alternative room effect (Electronic Time Offset adapted from Williams' MMA) and to dry approaches in terms of a number of multichannel spatial audio attributes. Live recordings of musical recitals were made in Tanna Schulich Hall at McGill University and used a frontal wavefront reconstruction array that can be interpreted as a manifestation of Steinberg and Snow's "Acoustic Curtain."

The objective of Chapter 3 study is to understand listeners' sensitivity to directional variation in non-ideal diffuse field reverberation. An ABX discrimination test was conducted using a semi-spherical 28-loudspeaker array; perceptual thresholds were estimated by systematically varying the level of a segment of loudspeakers and measuring participants' performance in the different conditions. The overall energy was held constant using a gain compensation scheme and measurements of the experimental stimuli are analyzed using a HATS dummy head as well as with opposing cardioid microphones aligned on the three cartesian axes. Opposing cardioid measurements were also made in acoustic spaces and demonstrate that level differences corresponding to the perceptual thresholds can be found in practice.

Chapter 4 presents the development of a systematic tool for the recording and reproduction of diffuse sound fields. Diffuse Field Modeling uses decorrelation filters based on the

statistical description of reverberation to "virtualize" an array of outward-facing cardioid microphones from linear combinations of a B-Format microphone under the assumption that the audio is diffuse. It describes the design of the physically-inspired decorrelation filters and validates the reconstructed diffuse fields through a numerical simulation based on the Kirchhoff / Helmholtz Integral for plausible radiation assumptions. It is demonstrated that the resulting fields have the expected spatial autocorrelation, that the channels of the array have the expected frequency-dependent correlation, and that a correspondence can be established between the introduced frequency autocorrelation and the RT60 of the recorded diffuse field.

Chapter 5 presents a perceptual evaluation of DFM as a component of a physically-plausible virtual acoustic model that can be systematically adapted to various loudspeaker arrays. Two distinct experiments were conducted. The first experiment used a 20-loudspeaker array with a 16-channel lower ring (identical to the simulations presented in Section 4.5) and a 4-channel height ring. The direct path and reflections were modeled geometrically and positioned using VBAP. The objective was to evaluate various treatments of DFM and to ensure that there were no major issues in practice. A second experiment used the 5.1 loudspeaker configuration and evaluated DFM against the Hamasaki Square [29][73], which was chosen for wide acceptance among tonemeisters and its physical similarities to DFM.

Appendix A presents a complete overview of the physical acoustics that are used throughout the dissertation. It begins with a brief review of the wave equation and the underlying fluid dynamics as well as the common cartesian and spherical solutions. Because multipoles are used extensively in Chapter 4, and because spherical basis functions are regularly encountered in virtual acoustics, both topics are reviewed and the relationship between is briefly discussed. The Sabine, modal, and time domain views of reverberation are reviewed in Section A.2. Note that room modes are formally solutions to the wave equation and are reviewed in Section A.2.3.

Appendix B presents a complete overview of the psychoacoustic concepts that are used throughout this dissertation. Section B.2.2 reviews the localization of direct sources and the frequency resolution of the inner ear. This frequency resolution becomes important when interpreting the lack of coloration in the algorithm presented in Chapter 4 which notably introduces significant amplitude and phase variation into room impulse responses. Section B.2 reviews the suppression of early reflected energy due to the precedence effect, the psychoacoustics of incoherent sources, and the perception of distance. Finally, Section B.3.1 reviews the perception of concert hall acoustics which are relevant to the objective of

recreating a perceptually-transparent reproduction of a listening space.

Ethics Review

The perceptual experiments in this dissertation were conducted according to the guidelines outlined by the McGill Research Ethics Board and certified by REB file #686-0606 granted to Catherine Guastavino for the project "Auditory Perception of Space: Localization and Sound Quality."

Chapter 2

Perceptual Evaluation of Room Effect Methods for Multichannel Spatial Audio

The room effect is an important aspect of sound recording technique, and is typically captured separately from the direct sound. The perceptual attributes of multi-channel spatial audio have been established by previous authors, while the psychoacoustic underpinnings of room perception are known to varying degrees. The Hamasaki Square, in combination with a delay plan and an aesthetic disposition to "natural" recordings, is an approach practiced by some sound recording engineers. This study compares the Hamasaki Square to an alternative room effect and to dry approaches in terms of a number of multi-channel spatial audio attributes. A concurrent experiment investigated the same spatial audio attributes with regards to the microphone and reproduction approach. As such, the current study uses a 12/2 system based upon 5.1 surround (3/2 stereo) where the frontal L/C/R triplet has been replaced by a linear wavefront reconstruction array.

2.1 Introduction

In a concurrently submitted article [63], comparisons are made between the relative merits of two different spatial audio recording and rendering techniques within the context of two different multichannel reproduction systems. The two recordings and rendering techniques are "natural," using main microphone arrays, and "virtual," using spot microphones, panning, and simulated acoustic delay. The two reproduction systems are a 5.1 surround system

(3/2 stereo) [72], and a 12/2 system, where the frontal L/C/R triplet is replaced by a 12 loudspeaker linear array. A more detailed description of the recording and reproduction strategies are given in a previous publication [62].

The current experiment concerns itself only with room effects in the context of the 12/2 system. The psychoacoustic underpinnings of room acoustics are first reviewed, followed by perceptual attributes of spatial audio. The merits and practice of using a carefully designed room effect are then discussed. The methods and results of the experiment are presented. Finally, the results are discussed.

2.1.1 Psychoacoustic Principles

The precedence effect, coherence, and distance hearing are the specific psychoacoustic phenomena related to free-field perception in enclosed spaces [6]. They are reviewed for their utility in subsequent discussion. The precedence effect refers to the fact that, for a direct sound with subsequent reflections, the direct sound determines the primary direction of the auditory event [6], [48]. Subsequent wavefronts are not imperceptible, but rather they are subject to the localization dominance of the primary wavefront.

An ideal diffuse field considers plane waves to be incident from all directions, with equal power density from all directions. The multitude of reflections in such a diffuse field are no longer individually audible, but rather give rise to the impression usually referred to as a diffuse sound [6], [7]. In non-anechoic conditions, this presents partially incoherent signals to the listener's ears. Generally speaking, partially incoherent sounds are either wide, are not located, or are located "everywhere" [6].

Hearing in the horizontal plane is aided by time and level differences between the two ears. Elevation hearing relies predominantly upon spectral cues from the torso, head, and pinna. For free-field, anechoic conditions, distance hearing relies almost exclusively upon relative changes in the signal level, and, as such, it is considerably less accurate than hearing in azimuth or elevation [6]. Louder signals tend to be heard closer than less loud signals; this is independent of loudspeaker distance. Establishing the relative levels of a sound source can sharpen distance perception; cues can be taken from multiple, simultaneous sound sources, or from the movement of a single sound source. Familiarity with a source influences the perception of distance, as the listener is likely to have expectations regarding the loudness and timbre.

Moore, citing experiments by von Bekesy, discusses the influence of the ratio of direct sound to reverberant energy and the timing of the early reflections. Adjusting either param-

eter led to the impression of the auditory event moving toward or away from the subject. Further, later experiments demonstrated that subjects could make absolute distance judgements of unfamiliar stimuli based upon the direct to reverberant ratio [48].

In addition to fundamental psychoacoustic principles, a number of topics from concert hall literature are relevant to the current inquiry. The subjective measure of "Spaciousness" or "Spatial Impression" is considered to have two major components, the Auditory Source Width (ASW) and Listener Envelopment (LEV) [11]. Loosely summarizing views found in the literature, ASW pertains to the perceived width of a sound source, considering the sound source to be "perceptually fuzed" with the early aspects of the room response. Listener Envelopment (LEV) is described as "the fullness of sound images around a listener" by Bradley and Soulodre, and is influenced primarily by the late reverberation of the room impulse response [11]. Rumsey, discussed below, considers a related parameter for auditory displays, defining the "Environmental Envelopment" as the "the sense of being enveloped by reverberant or environmental sound" [66]

Concert halls and listening spaces typically include methods of absorbing or scattering early reflections. This is primarily to reduce comb-filtering artifacts, which can arise when a reflection establishes a second, delayed path relative to the direct sound. Damping this reflection reduces the level of the comb-filtering, and the (typically low-pass) frequency dependence reduces the more noticeable high frequency components. A similar reduction in comb-filter artifacts can be achieved by scattering the coherent, specular component of the reflection [21].

Summarizing, the beginning of a room response is defined by the coherent direct path and the specular reflections. Due to the precedence effect, the reflections serve to widen and color the localization defined by the direct path. The late temporal region of the room response is defined by a diffuse field, which, as opposed to a coherent source, comes from all directions. The intermediate temporal aspects of the room response are a transition between these states. Subjective measures of concert halls such as ASW and LEV give an indication of the perceptual qualities of an acoustic space. Reflections can add coloration to an acoustic space's timbre, and are typically mitigated in some way.

2.1.2 Perceptual Attributes of Spatial Audio

The perceptual measurement of sound quality is a multifaceted problem involving a number of attributes, and "it is accepted that it is possible to identify and elicit these attributes" [4]. A study by Guastavino and Katz determined emergent attributes in the perceptual

evaluation of multi-channel spatial audio [27]. At the time of the study, the vocabulary and semantic scales of spatial audio were not clearly established; linguistic analysis of spontaneous written responses gave rise to emergent attributes. The emergent semantic scales were then used in subsequent magnitude estimation rating experiments.

The scene-based paradigm for spatial audio evaluation was proposed in 2002 in a paper by Rumsey [66]. A primary objective of this work was the establishment of consensus semantic scales, such that future investigations would employ similar (if not identical) descriptive characters of a sound scene. An additional 2010 study by Le Bagousse [2] reviewed the scales presented by a variety of authors. Our study, and its parallel study, considers perceptual attributes of spatial audio related to the room effect, as well as the capture method (array vs spot microphone approaches), and reproduction method (wave front reconstruction vs. stereophonic.) The capture and reproduction attributes are considered in previous and concurrent publications [62], [63]. The scales deemed appropriate for our investigation were chosen on the basis of their parallels to non-anechoic "natural" listening conditions as might be found in a concert hall. The scales are shown in Table 2.1.

Semantic Scale	Definition
Image distance	The distance that the ensemble or solo instrument is perceived to be at.
Image depth	A range of distinct distances within ensemble's or solo instrument's image.
Sense of Envelopment	The sense of being inside an acoustic space.
Localization	The precision of localization of the ensemble or solo instrument.
Image Stability	The stability of the image to head or body movements.
Instrument Width	The width of the ensemble or solo instrument.
Coloration and Transparency	The spectral coloration of the ensemble or solo instrument.

Table 2.1: Perceptual attributes of multichannel spatial audio used in this study. The definitions given to the participants are listed on the right.

2.1.3 Practice: 3/2 Stereo (5.1) and the Hamasaki Square

Sound recording engineers have developed strategies to capture and reproduce an acoustic environment. Their objectives are both physical and aesthetic. While it is the acoustic vibrations of the air that are captured and reproduced, strategies are used to create a flattering aesthetic presentation. Typically, this will involve a division of the direct and reverberant sound, allowing some freedom of perspective and presence.

3/2 stereo (5.1 Surround) refers to the approach of a stable, well-defined frontal sound in the L/C/R triplet, in addition to an envelopment strategy [72]. In practice, it is the presence of the two rearward surround loudspeakers that enables the presentation of distance, depth, and envelopment [73]. Damaske's work, cited by both Theile [73] and Blauert [6], discusses the perceptual effects of four symmetrical loudspeakers radiating reverberant sound of variable coherence. At low coherence levels, one perceives "clouds of reverberation around the loudspeakers." At higher levels of coherence, a centralized monophonic phantom image can appear above the listener's head. At low and medium levels of coherence, the listener experiences envelopment, referring to this as "subjective envelopment."

It is possible to display lateral reflections using the Left - Left Surround and Right - Right Surround segments (loudspeaker pairs), however phantom images within these segments are highly unstable due to the wide angular spacing between the loudspeakers [72], [6]. This instability, however, seems to be particularly forgiving; due to the precedence effect their direction is not audible. Rather, they serve to widen and reinforce the direct sound, and provide some components of "spatial impression," [73], [7]. It is argued that it is their timing, rather than their incident direction, which is of the greatest importance [73].

The Hamasaki Square (HS) [29] is a room effect technique which consists of four figure-8 microphones placed 4-5m from the stage in a square pattern, with approximately 1.5m spacing. The null of the directivity pattern oriented is towards the stage, and the positive axis is oriented towards the wall. The two microphones closest to the stage are routed to the Left (L) and Right (R) loudspeakers, while the two farthest microphones are routed to the Left Surround (LS) and Right Surround (RS) loudspeakers.

The spacing between the microphones determines the correlation of the late field signal, and provides controllable listener envelopment [29]. Because very little direct sound is picked up, the recording engineer has independent, creative control of the room effect. Lateral reflections are picked up by the figure-8 directivity and reproduced by the L and LS or R and RS loudspeakers. A speculative point, based upon the authors experience designing artificial reverberators and the literature in that field [7], is that the Hamasaki Square will also capture a facsimile of the acoustic environments' behavior prior to the mixing time of the room. That is, the difficult temporal region between individual early reflections, and later diffuse energy, will be given a plausible representation by this technique. The issue facing this temporal region is that it is the transition region between the specular (deterministic) description and the diffuse (stochastic) descriptions of the sound field.

The direct sound is typically captured close to the ensemble of performers, where the

direct sound greatly outweighs the reverberant sound. The room sound is typically captured at some distance from the ensemble, where the direct to reverberant ratio heavily favors the reverberant sound. Theile [73], recommends a "delay plan" to insure temporal alignment between these two aspects in the final acoustic display.

2.1.4 Research Questions

The practitioners approach of the Hamasaki Square, in combination with a delay plan, and a "natural" mixing strategy yields a room effect that, when presented on a 5.1 (3/2 stereo) system, is physically more plausible than other approaches. We hypothesize that this physical plausibility of the Hamasaki Square will lead to increased ratings on the distance, depth, and envelopment semantic scales, when compared with dry or alternative room effect approaches. Considering the precedence effect and the subjective measure of ASW, both of which pertain to early lateral reflections, we hypothesize that the dry approach will have the most precise localization. Given the Hamasaki Square's ability to reproduce a facsimile of the early lateral reflections, we hypothesize approaches using the HS will be rated higher for the Instrument Width scale. Comb-filtering artifacts are associated with the presence of discrete, specular reflections. This, in combination with an aesthetic preference for a "large" sound, typically associated with close spot microphone distances, leads to the hypothesis that coloration will be lowest for dry approaches.

2.2 Methods

2.2.1 Treatments

The present study is part of a larger project investigating the relative merits of different multichannel spatial audio recording and rendering techniques [63]. The two recordings and rendering techniques are "natural," using main microphone arrays, and "virtual," using spot microphones, panning, and simulated acoustic delay. The two reproduction systems are the 5.1 surround system (3/2) [72], and a 12/2 system, where the frontal L/C/R triplet is replaced by a 12 loudspeaker linear array. The current experiment concerns itself only with the 12/2 system. Both spot-based and array-based microphone techniques were used. Test stimuli were recorded in Tanna Schulich Hall, McGill University. A detailed description of the recording and reproduction strategies is given in [62].

The array-based technique was derived from Steinberg and Snow's "Acoustic Curtain",

where a large, linear array of microphones was mapped directly to loudspeakers in complimentary positions. The embodiment consisted of 12 Sennheiser MKH 8040 cardioid microphones, attached to a hanging 3m bar and mapped directly to loudspeakers in complimentary positions. The spacing between each microphone and loudspeaker was 19cm on center. Given that a sufficient number of microphones and loudspeakers are used, a wavefront incident upon the microphones will be accurately reproduced by the loudspeakers [34]. The recording angle is a fixed "window" looking forward onto the stage, defined by the line from the listener to the edges of the array, and was approximately $\pm 40°$.

Closely placed spot microphones were recorded simultaneously, and were later reproduced as virtual sources using the Sonic Emotion Wave 1 processor. The spatial layout of the spot recordings was made to match the spatial layout of array-based natural recordings as closely as possible using the WavePerformer user interface to the Sonic Emotion Wave 1 processor [20]. The (x,y) position was made to match that of the spot microphone's position on stage and acoustic propagation delay was approximated using digital delay. Source level was matched to the array recordings and the recordings were mixed by graduate students from McGill's Sound Recording program. Relatively little aesthetic freedom was allowed, no artificial reverberation or compression was used, no automation was used, and the same equalizer was used for all channels to avoid phase irregularities.

Two room effect strategies were employed: the first was the Hamasaki Square, the second was a pair of cardioid room microphones mounted above the array and suitably delayed in the subsequent mixes. The HS was modified to fit the 12/2 system by reproducing the Left and Right HS signals as virtual sources using the Sonic Emotion Wave 1 processor. If this had not been done, the room effect would have emanated from a single loudspeaker in the WFS array, which would have been tremendously unnatural. Informal listening tests confirmed that the extension of the HS to the 12/2 system did not negatively impact its function.

Five treatments were mixed following a "natural" approach, where attention was given to reproducing the sensation of "you are there." Most notably, this requires the judgment of a plausible, and approximately consistent, level of reverberation. A great deal of effort was also given to insuring a consistent spatial lay out of the direct sound, as discussed in previous work [62]. SPOT,DRY comprised spots microphones, spatialization using the Wave 1 processor, and no room effect. ARRAY,DRY comprised the 12 microphones of WFR,ARRAY and no room effect. ARRAY,SURR comprised the 12 microphones of WFR,ARRAY and two cardioid room microphones. ARRAY,HS comprised the 12 microphones of WFR,ARRAY

and the HS. Finally, SPOT,HS comprised spots microphones, spatialized using the Sonic Emotion Wave 1 processor, and the HS.

A few words should be said regarding particulars of the SURR room effect. The cardioid room microphones were mounted on the array pointed into the hall, away from the ensembles, at an angle of 15° to the centerline of the room. The signals from these microphones were delayed 10 ms, as might be done in a Williams MMA configuration using electronic timing offset (ETO) [79]. For the ARRAY,HS and SPOT,HS treatments, the rear channels of the HS were held at a constant level, while the SPOT,HS had an additional 1.5dB of gain in the L/R HS signals to compensate for the spot microphones lack of frontal reverberation. The ARRAY,HS, ARRAY,SURR, and SPOT,HS cases were judged to have approximately identical amounts of reverberation. All stimuli were volume matched using a sound meter.

2.2.2 Participants

Fifteen subjects (one female and fourteen males) ranging in age from 19 to 58 years participated in the experiment. Two participants were professional sound engineers and held faculty positions in McGill's Sound Recording department. Four participants were Ph.D. students in the Sound Recording program, eight participants were Master's in the Sound Recording program, and one participant was an undergraduate in the qualifying year of the Sound Recording program. All participants were well versed in 5.1 surround sound production, two participants were versed in 12/2 surround sound production.

2.2.3 Apparatus and Stimuli

The perceptual evaluation took plane in the Critical Listening Lab of the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) at McGill University in Montreal, Canada. The room was outfitted with the 5.1 and 12/2 reproduction systems described above, and followed the ISO standard for a 5.1 listening room. The 12 loudspeaker line array consisted of Genelec 8020 loudspeakers spaced at 19cm center to center. The array was placed immediately above the LCR triplet of the 5.1 system and was identical in width. Genelec 8050 loudspeakers were used for the 5.1 system, sharing the surround loudspeakers with the 12/2 system.

Stimuli were played on a Apple Mac Pro using a RME HDSPe MADI audio interface. The playback levels of the stimuli were approximately matched by using a sound pressure level meter. The user interface was programmed in Max/MSP.

2.2.4 Procedure

This experiment was conducted in tandem with concurrent work [63] for all participants. An initial preference experiment was always performed first, while this experiment and third experiment were counterbalanced for all participants. Breaks were allowed, and subjects took between 45 minutes and 3 hours to perform the three combined experiments.

The experiment presented 21 distinct trials to the subject. Each trial consisted of one of seven semantic scales, "Distance," "Depth," "Envelopment," "Localization," "Image Stability," "Instrument Width," or "Coloration/Transparency," and 1 of 3 musical excerpts. The excerpts were taken from a modern vibraphone piece, a modern percussion ensemble, and a classical piano sonata played on a period instrument. Excerpts were chosen from each of these 3 pieces based upon musical phrasing. The user interface presented a given scale from Table 2.1 and excerpt in 5 different treatments: ARRAY,HS, ARRAY,DRY, ARRAY,SURR, SPOT,HS, and SPOT,DRY. There were 5 rating sliders corresponding to the 5 treatments, as well as a single, optional text box. The presentation of scale, excerpt, and treatment was randomized within and across trials. Upon rating each treatment according to the specified semantic scale, the subject could proceed to the subsequent trial.

2.2.5 Analysis

The analysis was done using a fully factorial ANOVA on the raw ratings, collapsed over all participants. Because a significant main effect of scale was found, each scale was analyzed separately using a $5(Treatment) \times 3(Excerpt)$ ANOVA. For most cases, a significant main effect of excerpt was found, and results are presented separately for each excerpt. Outliers beyond two standard deviations from the group means were discarded and replaced by the median value, accounting for less than 3% of the data. Post-hoc tests were performed using Tukey-Kramer multiple comparisons ($p < 0.05$) in all cases.

2.3 Results

The ratings, main and interaction effects, and post-hoc tests for the Distance scale are shown in Figure 2.1, for the Depth scale in Figure 2.2, and for the Envelopment scale in Figure 2.3.

The ratings, main and interaction effects, and post-hoc tests for the Localization scale are shown in Figure 2.4. For the Image Stability scale, we observed no significant main or interaction effects. The ratings, main and interaction effects, and post-hoc tests for the

Instrument Width scale are shown in Figure 2.5, and for the Coloration/Transparency scale in Figure 2.6. Note that higher ratings refer to greater transparency, and that lower ratings refer to greater coloration.

In most cases, either a main or interaction effect was found for *Excerpt*, and, as such, the excerpts were analyzed in isolation. For the sake of readability and brevity, the following tables present only the significant results. Statistical tendencies (nearly significant results) are noted.

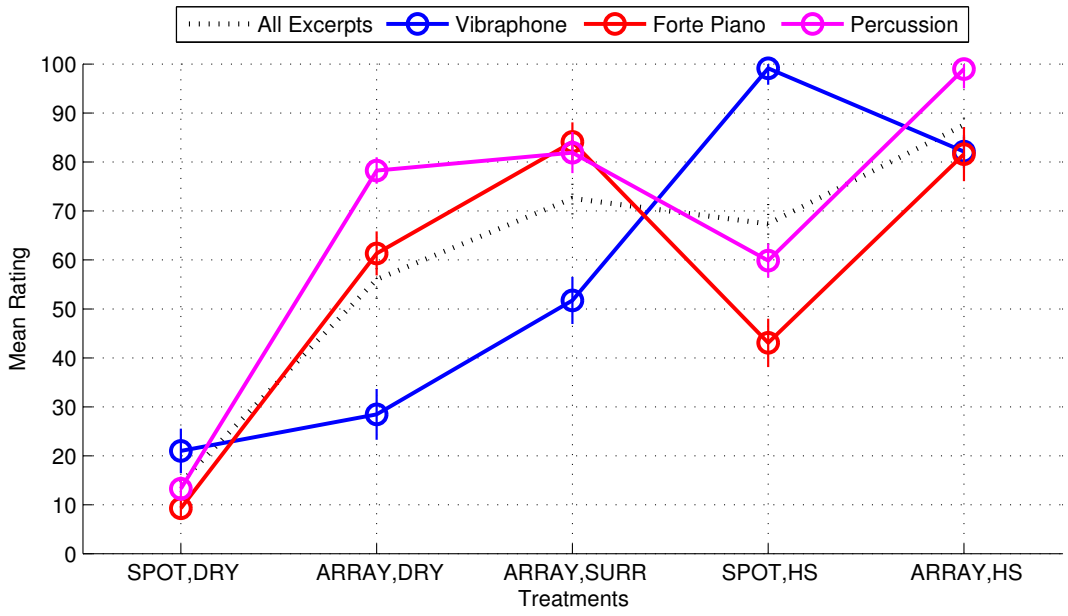

Figure 2.1a: Ratings for the Distance scale. The low bound was "near," the high bound was "far."

Factor	Significant F and p values
All Excerpts	
Treatment	$F_{(8,210)} = 139.07, p < 0.0001$
Excerpt	$F_{(8,210)} = 10.71, p < 0.0001$
Treatment × *Excerpt*	$F_{(8,210)} = 26.59, p < 0.0001$
Post-hoc tests	ARRAY,HS > All other cases; Percussion Ensemble > Vibraphone and Forte Piano.
Vibraphone	
Treatment	$F_{(4,70)} = 60.92, p < 0.0001$
Post-hoc tests	ARRAY,HS and SPOT,HS > All other cases.
Forte Piano	
Treatment	$F_{(4,70)} = 50.07, p < 0.0001$
Post-hoc tests	ARRAY,HS and ARRAY,SURR > All other cases.
Percussion	
Treatment	$F_{(4,70)} = 91.98, p < 0.0001$
Post-hoc tests	ARRAY,HS > All other cases; ARRAY,DRY and ARRAY,SURR > SPOT,HS and SPOT,DRY.

Figure 2.1b: Main effects, interaction effects, and post-hoc tests for the Distance scale.

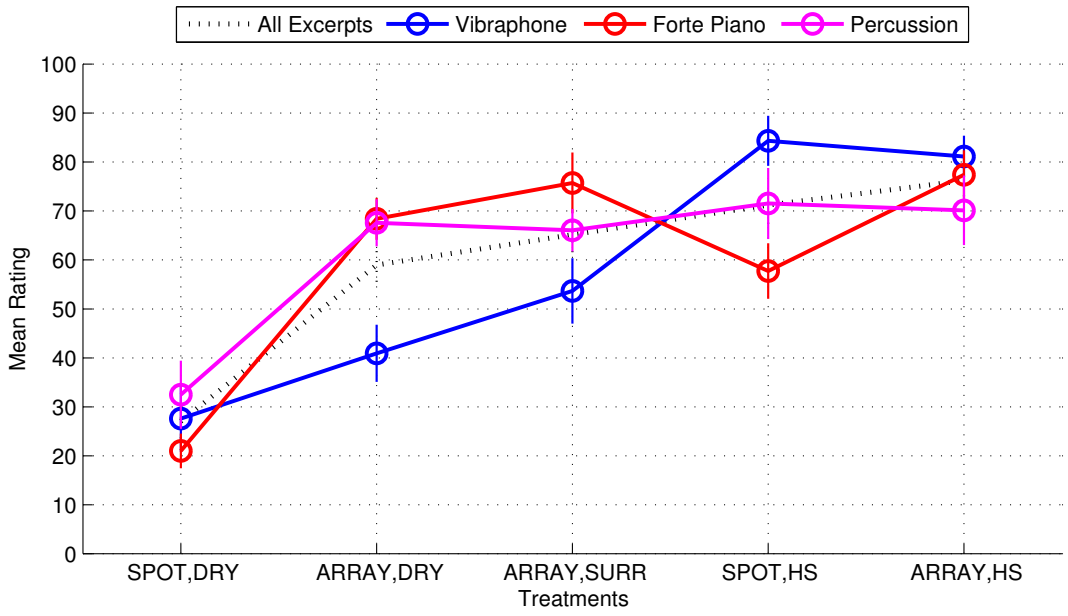

Figure 2.2a: Ratings for the Depth scale. The low bound was "shallow," the high bound was "deep."

Factor	Significant F and p values
All Excerpts	
Treatment	$F_{(8,210)} = 34.56, p < 0.0001$
Treatment × *Excerpt*	$F_{(8,210)} = 4.5, p < 0.0001$
Post-hoc tests	ARRAY,HS > ARRAY,DRY or SPOT,DRY.
Vibraphone	
Treatment	$F_{(4,70)} = 18.24, p < 0.0001$
Post-hoc tests	ARRAY,HS and SPOT,HS > All other cases.
Forte Piano	
Treatment	$F_{(4,70)} = 21.16, p < 0.0001$
Post-hoc tests	SPOT,DRY < All other cases.
Percussion	
Treatment	$F_{(4,70)} = 6.94, p < 0.0001$
Post-hoc tests	SPOT,DRY < All other cases.

Figure 2.2b: Main effects, interaction effects, and post-hoc tests for the Depth scale.

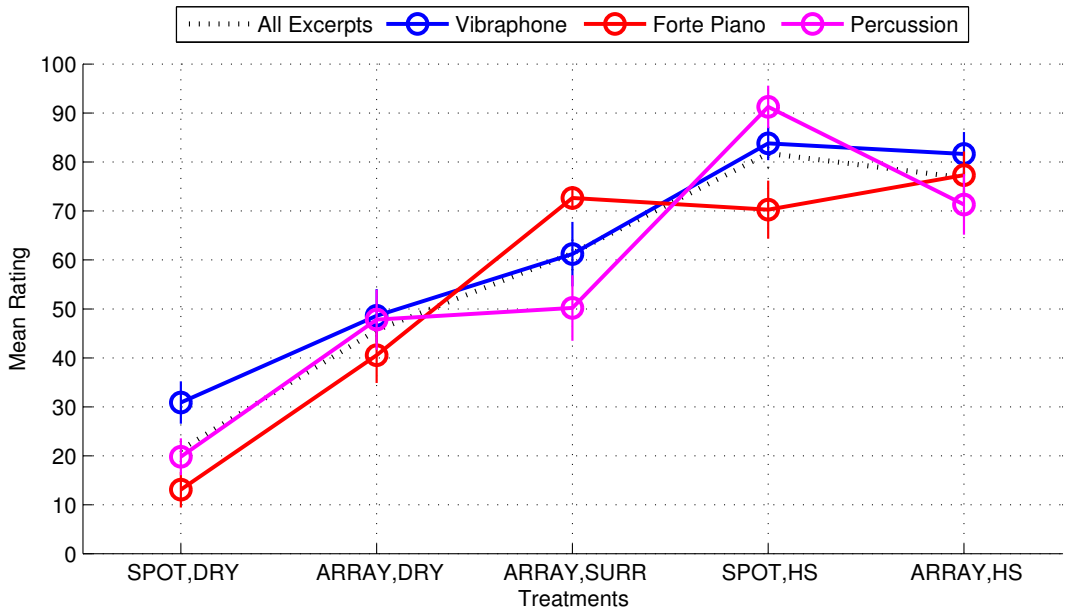

Figure 2.3a: Ratings for the Envelopment scale. The low bound was "low," the high bound was "high."

Factor	Significant F and p values
All Excerpts	
Treatment	$F_{(8,210)} = 70.59, p < 0.0001$
Treatment × *Excerpt*	$F_{(8,210)} = 2.99, p = 0.0034$
Post-hoc tests	ARRAY,HS and SPOT,HS > All other cases.
Vibraphone	
Treatment	$F_{(4,70)} = 20.33, p < 0.0001.$
Post-hoc tests	ARRAY,HS and SPOT,HS > All other cases.
Forte Piano	
Treatment	$F_{(4,70)} = 34.25, p < 0.0001$
Post-hoc tests	ARRAY,HS, ARRAY,SURR and SPOT,HS > ARRAY,DRY and SPOT,DRY.
Percussion	
Treatment	$F_{(4,70)} = 23.54, p < 0.0001$
Post-hoc tests	SPOT,HS > ARRAY,DRY, ARRAY,SURR, and SPOT,DRY; ARRAY,HS > ARRAY,DRY, and SPOT,DRY.

Figure 2.3b: Main effects, interaction effects, and post-hoc tests for the Envelopment scale.

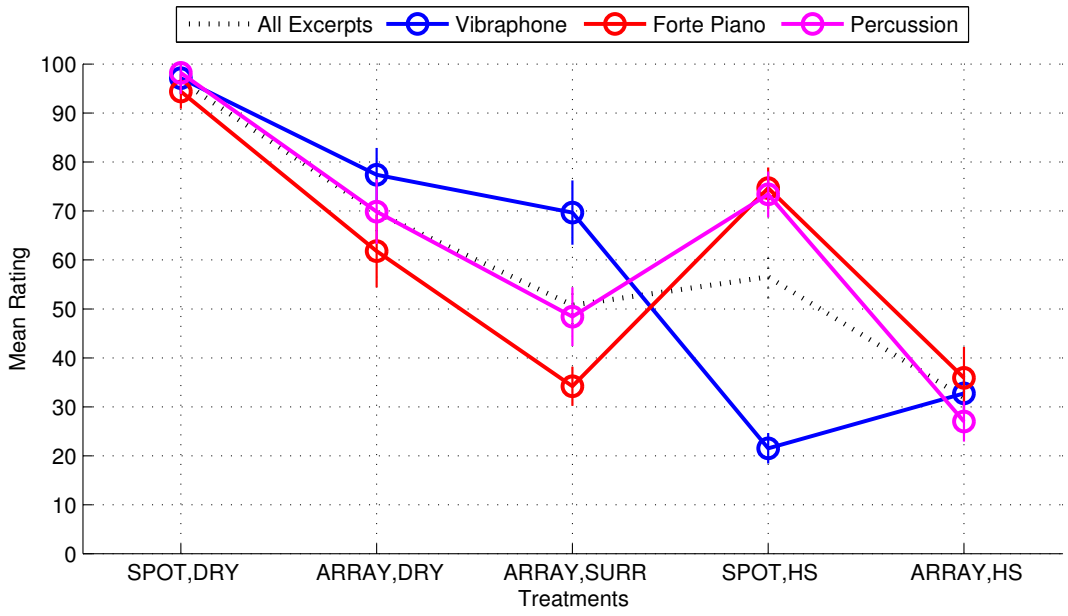

Figure 2.4a: Ratings for the Localization scale. The low bound was "vague," the high bound was "precise."

Factor	Significant F and p values
All Excerpts	
Treatment	$F_{(8,210)} = 66.99, p < 0.0001$
Treatment × *Excerpt*	$F_{(8,210)} = 12.58, p < 0.0001$
Post-hoc tests	ARRAY,HS < All other cases. SPOT,DRY > All other cases.
Vibraphone	
Treatment	$F_{(4,70)} = 44.89, p < 0.0001$
Post-hoc tests	ARRAY,HS and SPOT,HS < All other cases.
Forte Piano	
Treatment	$F_{(4,70)} = 23.55, p < 0.0001$.
Post-hoc tests	ARRAY,HS, ARRAY,SURR < All other cases.
Percussion	
Treatment	$F_{(4,70)} = 26.53, p < 0.0001$
Post-hoc tests	ARRAY,HS < All other cases; ARRAY,SURR < SPOT,HS, ARRAY,DRY, and SPOT,DRY; SPOT,DRY > All other cases.

Figure 2.4b: Main effects, interaction effects, and post-hoc tests for the Localization scale.

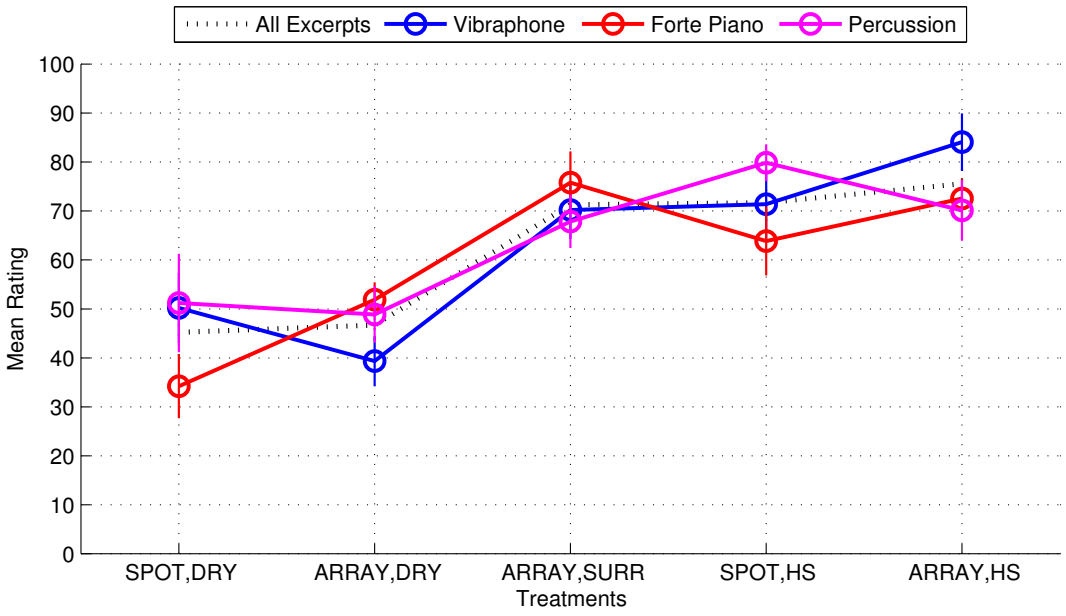

Figure 2.5a: Ratings Instrument Width scale. The low bound was "narrow," the high bound was "wide."

Factor	Significant F and p values
All Excerpts	
Treatment	$F_{(8,210)} = 17.19, p < 0.0001$.
Post-hoc tests	ARRAY,HS, ARRAY,SURR, and SPOT,HS > ARRAY,DRY and SPOT,DRY.

Figure 2.5b: Main effects, interaction effects, and post-hoc tests for the Instrument Width scale.

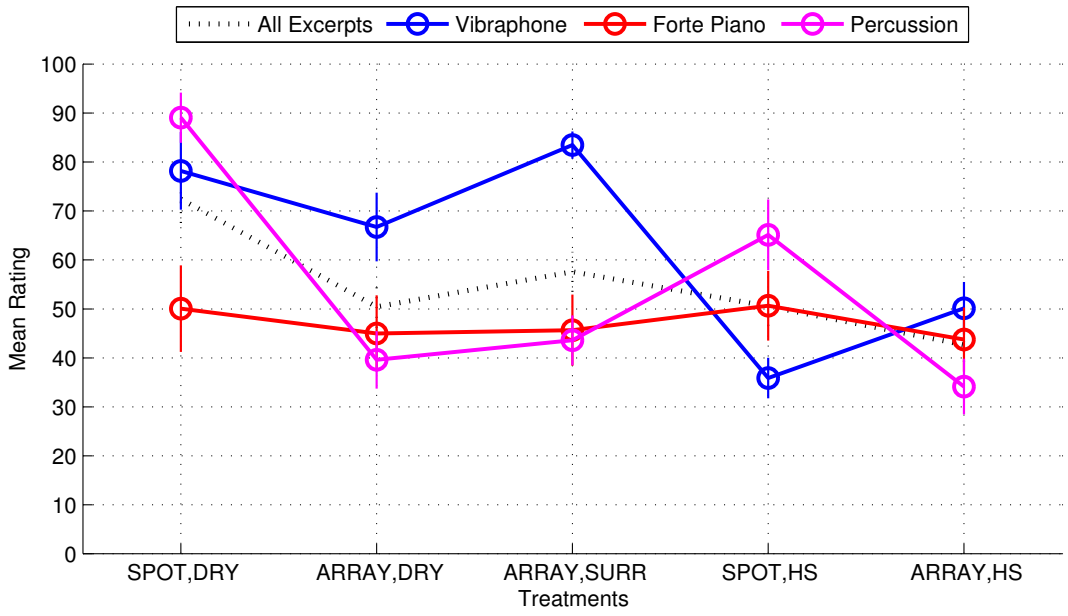

Figure 2.6a: Ratings for the Coloration/Transparency scale. The low bound was "colored," the high bound was "transparent."

Factor	Significant F and p values
All Excerpts	
Treatment	$F_{(8,210)} = 9.09, p < 0.0001$
Excerpt	$F_{(8,210)} = 7.59, p = 0.0007$
Treatment × *Excerpt*	$F_{(8,210)} = 6.5, p < 0.0001$
Post-hoc tests	ARRAY,HS < ARRAY,SURR and SPOT,DRY; SPOT,DRY > All other cases.
Vibraphone	
Treatment	$(F_{(4,70)} = 11.78, p < 0.0001)$
Post-hoc tests	ARRAY,HS < ARRAY,SURR and SPOT,DRY; SPOT,HS < ARRAY,SURR, ARRAY,DRY and SPOT,DRY.
Forte Piano	No significant main or interaction effects.
Percussion	
Treatment	$F_{(4,70)} = 15.06, p < 0.0001$
Post-hoc tests	ARRAY,HS and ARRAY,SURR < SPOT,HS and SPOT,DRY.

Figure 2.6b: Main effects, interaction effects, and post-hoc tests for the Coloration/Transparency scale.

2.4 Discussion

The hypothesis of this experiment was that different room effect strategies would lead to differing ratings of a number of spatial audio attributes. Regarding the distance scale, the finding that ARRAY,HS was rated as significantly more distant than all other cases is notable. For the depth scale, ARRAY,HS was rated as having greater depth compared to either dry approach. Considering the vibraphone excerpt in isolation, post-hoc tests showed both ARRAY,HS and SPOT,HS as being rated as having greater depth than all other approaches.

The distinction between distance and depth, that "depth is a range of distances," is somewhat subtle. It has been stated that the approach of the Hamasaki Square, in combination with a delay plan and a "natural" mixing strategy, yields a room effect that, when presented in a 5.1 or 12/2 system, is physically more plausible than other approaches. While it is expected that distance and depth will depend upon the direct to reverberant ratio, we note that this can typically be controlled by the sound recording engineers' mixing, and that the alternative ARRAY,SURR approach was judged to have a commensurate amount of reverberation. Considering this, it is speculated that there may be a dependence upon early lateral reflections, and possibly upon the time varying aspects of the late field correlation. While the argument is not physically rigorous, the Hamasaki Square will provide some facsimile of both.

It was hypothesized that envelopment, being related typically to the late reverberant energy, would also be rated higher for the ARRAY,HS and SPOT,HS treatments. Collapsing over all excerpts, this was confirmed. Similar to the depth and distance scales, there is a known dependence upon the direct to reverberant ratio. However, the consideration of the mixing levels, and the commensurate amount of reverberation in the ARRAY,SURR approach, again lead to the speculation that there may be a dependence upon early lateral reflections, and possibly upon the time varying aspects of the late field correlation.

It was hypothesized that SPOT,DRY will have the most precise localization, with decreasing localization as more non-direct energy is included in the display. This was confirmed in the fact that ARRAY,HS was rated as the least precise, while SPOT,DRY was rated as the most precise. While early reflections and the late field reverberation of ARRAY,HS and SPOT,HS may be more akin to free-field listening conditions in "natural," non-anechoic conditions, the most precise localization is loosely analogous to spatial hearing in anechoic conditions (SPOT,DRY). One can interpret this in light of the fact that diffuse energy is located "everywhere," and strategies involving a room effect balance direct and fully diffuse

sound.

Image stability, as anticipated, was not influenced by either treatment or excerpt.

Instrument width is influenced by the presence of early reflections; this was confirmed in the fact that ARRAY,HS and SPOT,HS were rated as wider than either dry case. This is consistent with literature regarding ASW and the precedence effect, and suggests that some degree of plausible early reflections are, in fact, involved in an acoustic display involving the HS. It may also be due to a balance of concisely located direct sound, and diffuse energy which is located "everywhere."

SPOT,DRY was judged to be significantly more transparent than all other cases. This may be due to comb-filtering artifacts which could be anticipated for ARRAY,HS and SPOT,HS, or to the particular timbre of the hall's reverberation.

2.5 Conclusion

The Hamasaki Square, used in conjunction with a delay plan and a "natural" recording aesthetic, was rated as having greater distance, depth, and envelopment than a alternative room effect. Instrument width was also rated as widest for this approach. To the contrary, localization was rated as most precise, and coloration/transparency most transparent, for listening conditions analogous to an anechoic chamber. Generally speaking, distance, depth, and envelopment are desirable qualities in a recording, and it is reasonable to conclude that the Hamasaki Square can provide these attributes. Instrument width may depend upon the sound engineer's particular aesthetic goals. It seems, however, that there is a trade-off between the desirable qualities of distance, depth, and envelopment, and the qualities of localization and transparency.

It has been speculated that the higher ratings of distance, depth, and envelopment may be dependent upon early lateral reflections and the time varying aspects of the late field correlation. The results of the current study merits further investigation into the relationship between these physical variables and their perceptual ramifications.

Segue

Chapter 2 is based on *A Perceptual Evaluation of Room Effect Methods for Multichannel Spatial Audio* by David Romblom, Richard King, and Catherine Guastavino presented at the Audio Engineering Society's 135th Convention in New York City; the paper received the designation of "Best Student Technical Paper" for this convention. This was done in the context of the sound recording and the focus was divided between the aesthetic presentation of a musical work and the physical representation of the sound field. The Hamasaki Square is a channel-based "direct" microphone technique that has physical interpretation as a sparse sampling of a bounding surface. Live recordings of musical recitals were made in Tanna Schulich Hall at McGill University and the Hamasaki Square was rated as having greater distance and envelopment when compared to an alternative room effect and a number of dry approaches.

Chapter 3 is a psychophysical experiment that is based on *Perceptual Thresholds for Non-Ideal Diffuse Field Reverberation* by David Romblom, Catherine Guastavino, and Philippe Depalle. This work has been accepted to the Journal of the Acoustical Society of America and will be published in the 2016 calendar year. Both the physical and perceptual literature typically assumes the reverberant diffuse field to have equal energy from all directions. It is plausible, however, that this energetic distribution could vary in acoustic scenarios with vastly differing wall materials, elongated shapes, or with opening such as windows. It is also plausible that this be part of the spatial impression of musical listening spaces such as cathedrals and would be of substantial utility for virtual reality and gaming scenarios. The objective of Chapter 3 study is to understand listeners' sensitivity to directional variation in non-ideal diffuse field reverberation, to determine if these differences can be observed in existing acoustic spaces, and what the corresponding ear signals are.

It is notable that the Hamasaki Square has only a limited ability to record and reproduce non-ideal diffuse field reverberation, the strength of this technique lies in its ability to provide plausible channel correlation, a facsimile of early reflections, and independence from the

direct sound capture. The results of Chapters 2 and 3 will be further discussed in the Interim Summary, and all of these considerations will factor into the Diffuse Field Modeling algorithm presented in Chapter 4 and validated by listening tests in Chapter 5.

Chapter 3

Perceptual Thresholds for Non-Ideal Diffuse Field Reverberation

The objective of this study is to understand listeners' sensitivity to directional variations in non-ideal diffuse field reverberation. An ABX discrimination test was conducted using a semi-spherical 28-loudspeaker array; perceptual thresholds were estimated by systematically varying the level of a segment of loudspeakers for lateral, height, and frontal conditions. The overall energy was held constant using a gain compensation scheme. When compared to an ideal diffuse field, the perceptual threshold for detection is -2.5 dB for the lateral condition, -6.8 dB for the height condition, and -3.2 dB for the frontal condition. Measurements of the experimental stimuli are analyzed using a HATS dummy head as well as with opposing cardioid microphones aligned on the three cartesian axes. Additionally, opposing cardioid measurements are made in an acoustic space and demonstrate that level differences corresponding to the perceptual thresholds can be found in practice. These results suggest that non-ideal diffuse field reverberation may be a previously unrecognized component of spatial impression.

3.1 Introduction

Physically, the reverberant diffuse field is the result of myriad reflections of an acoustic source in an enclosed space. Observed at a given location, it is incoherent energy incident from all directions equally [53] and is perceived as an auditory event heard "everywhere" [6]. Directional variations in the reverberant field can arise when a room has some area of open windows or near a particularly absorptive wall [53]. The objective of this study is

to determine if the human auditory system is sensitive to directional variations in non-ideal diffuse field reverberation and if these directional variations can be observed in measurements of an acoustic space.

The Sabine reverberation model relates the bulk exponential decay of reverberation to the absorptive properties of the room by assuming that reverberant acoustic energy is approximately the same in all spatial regions of the room and that the absorption due to the bounding walls can be estimated as an average [53]. However, absorption coefficients α describing a wall's absorption can differ by more than an order of magnitude at the same frequency. For instance, at 500 Hz a concrete floor has an α value of 0.02, while heavy drapery on a wall has a value of 0.55. An open window (no wall) has an α value of 1.0. The discrepancy between α coefficients led to the speculation that the diffuse energy may differ in a directional manner, and that these directional differences may be a component of spatial impression in both real and virtual acoustic scenarios.

A second motivation for this study originated in the development of a three-dimensional reverberation strategy referred to as Diffuse Field Modeling (DFM) described in Romblom et al.[60] and validated with listening tests in Rummukainen et al.[65]. It is based on physically-inspired decorrelation filters and specially treated B-Format room impulse response (RIR) and is intended to be used in conjunction with geometric models of the direct and reflected sound. DFM is not explicitly a physical model of an acoustic space but instead relies on the statistical description of reverberation to simulate the observations that would be made on a microphone array in a diffuse field. It considers both the RT60 of the RIR (linked to frequency autocorrelation of the decorrelation filters) as well as the loudspeaker geometry (linked to frequency-dependent channel correlation). If directional variations in the reverberant diffuse field are audible, then it ought to be appropriately modeled in the algorithm.

The pure tone diffuse field is a mathematical idealization that describes the field as a superposition of a large number of $q = 1, 2....Q$ plane waves of random direction, phase, and amplitude. This model approximates the sound field that occurs in a closed volume when a sound source is reflected many times and when the receiver is sufficiently far away from the sound source. For constant frequency fields, the complex pressure $\hat{p}$ can be described as a summation:

$$\hat{p} = \sum_{q=1}^{Q} \hat{p}_q \; e^{jk\mathbf{n_q}\cdot\mathbf{r}} \tag{3.1}$$

where $\hat{p}_q$ is the complex amplitude of the q^{th} plane wave, k is the wave number, $\mathbf{n}_q$ is the directional vector of the q^{th} plane wave, and $\mathbf{r}$ is a vector pointing to the observation [53].

Note that the distribution of acoustic values changes if the driving frequency changes [53],[33]. The modal approach to room acoustics is obtained by finding solutions to the Helmholtz equation in a rigidly bounded room - each mode is a spatial description of the pressure field that will be excited to varying degrees depending on the location of a source and the difference between the modal frequency and the source's driving frequency. Above the Schroeder cutoff frequency the modal resonances are closely spaced relative to their bandwidth and can be considered a continuous distribution of frequency components. The spectrum will be different at each location due to the varied contribution of many modes. For broadband excitation, both the spatial and modal views substantiate modeling the diffuse aspect of a room impulse response with uncorrelated white Gaussian noise (WGN) [33],[7],[53].

In the horizontal plane, the auditory system is able to localize auditory events using differences in the acoustic signals received at the two ears. Sounds not located in the median plane will have interaural time differences (ITD) due to the differing path lengths. Interaural level differences (ILD) arise due to a number of acoustic processes and are typically frequency-dependent. Diffraction, which allows sound to bend around obstructions, yields little level difference at low frequencies, but substantial head-shadowing at moderate to high frequencies. When the wavelength of the sound is on the order of the pinna, the acoustic processes of diffraction, dispersion, and reflection impose fine spectral detail upon the signals at each ear. Resonances of the ear canal and cavum conchae are excited differently based on the direction and frequency of the incident sound. Finally, the torso both reflects and shadows sound depending on the frequency and incident direction [6].

Following the definition of a diffuse field, we are concerned with the interaction of many planar random time signals interacting with the listener's head and torso. The multitude of reflections in such a diffuse field are no longer individually audible, but rather give rise to the impression usually referred to as a diffuse sound [6], [7]. The numerous reflected sound beams superimpose at the listener's ears to form partially incoherent signals, and the physical measure of cross-correlation is a means of estimating the similarity of two given signals. The cross-correlation of the time-domain ear signals ($l(t)$ and $r(t)$) is found by integrating their instantaneous product over their duration and then normalizing by the product of their root-mean-squares. This can be done for any specified time offset and is defined as:

$$\Phi_{lr}(\tau) = \frac{\int_{-\infty}^{\infty} l(t)r(t-\tau)dt}{\sqrt{\int_{-\infty}^{\infty} l^2(t)dt \int_{-\infty}^{\infty} r^2(t)dt}} \tag{3.2}$$

The maximum value of this equation is unity, and this can be achieved by signals differing

only by amplitude or a time delay. It is negated for signals differing by a 180° phase shift. Signals with values of cross-correlation near 1.0 are referred to as coherent, values near zero are incoherent [6].

A study by Hiyama and Hamasaki [31] sought to determine the minimum number of loudspeakers required to perceptually approximate a diffuse field. A horizontal ring of 24 loudspeakers was driven by uncorrelated white noise and was verified as having a frequency-dependent cross-correlation very near the theoretical ideal. This was considered to be the reference and then compared to a number of different loudspeaker configurations with fewer loudspeakers. Using a "difference grade" rating experiment, as few as 6 regularly spaced loudspeakers were found to be perceptually comparable with the reference. A 12-loudspeaker configuration showed no perceptual difference with the reference. A second experiment used band-passed uncorrelated white noise (0.1 to 1.8kHz, 1.8 to 7kHz, and 7 to 20kHz.) Again 6 loudspeakers were perceptual similar to the reference, and it is notable that the lowest band was more dissimilar for a number of the sparsest configurations. The 12-loudspeaker arrangement had difference grades of nearly 0 for all three frequency bands.

In a recent and very closely related experiment, Santala and Pulkki used uncorrelated pink noise in a 15-loudspeaker frontal arc and presented subjects with a wide variety of regular and irregular configurations of active and inactive loudspeakers. Subjects indicated the loudspeakers they thought to be active and a histogram of all of the subjects' responses was assumed to estimate the spatial impression. Notable is the fact that many configurations differing by only one loudspeaker or one pair of symmetric loudspeakers were rated similarly which suggests that the exact distribution of loudspeakers could not be perceived accurately. Their second experiment reduced the number of active loudspeakers compared to a 13 loudspeaker reference. Various bandwidths of pink noise centered at 500 and 4000 Hz were used. While only the frontal arc was used, the configurations were otherwise extremely similar to Hiyama's experiment and the resulting difference grades for 7 loudspeakers (Hiyama's 12 loudspeaker configuration) were nearly identical to the reference.

In concert hall literature, the perceptual measure of "Spaciousness" or "Spatial Impression" is considered to have two major components, the Auditory Source Width (ASW) and Listener Envelopment (LEV) [11],[51]. ASW pertains to the perceived width of a sound source, considering the sound source to be "perceptually fuzed" with the early part of the room response. Listener Envelopment (LEV) is described as "the fullness of sound images around a listener" and is influenced primarily by the late reverberation of the room impulse response [11]. The physical measures typically associated with both ASW and LEV are the

degree of interaural cross-correlation (IACC) and the degree of lateral energy described by the lateral fraction (LF.) ASW corresponds to the first 80 ms of the room response while LEV corresponds to the time after 80 ms. While ASW and LEV do influence each other, it is argued they are distinct perceptual dimensions as participants were able to independently identify conditions of ASW and LEV depending on the physical values of LF [11].

The objective of this study is to understand listeners' sensitivity to directional variations in non-ideal diffuse field reverberation. It was motivated by the observation that absorption coefficients α can vary substantially and to inform aspects of the Diffuse Field Modeling algorithm. Due to a possible interaction with the modeling of the direct sound and early reflections, and, due to the focus on the development of the above algorithm, it was decided to study the reverberant diffuse field in isolation. Given this baseline, future studies are needed to understand the perception of non-ideal diffuse field reverberation in a more plausible context involving both the direct path and the early reflections.

This paper is organized as follows. Section 3.2 describes the current experiment, data analysis, and results. Section 3.3 analyzes the experimental stimuli using a dummy head and additionally uses opposing cardioid microphones to compare the experimental stimuli to room impulse responses measured from an existing shoebox style concert hall. The experimental results and signal analysis are discussed in Section 3.4.

3.2 Experiment

3.2.1 Participants

Nineteen participants ranging from 23 to 48 years of age participated in the experiment. Six were professional sound recording engineers, eight were audio researchers, and five were professional musicians. All participants had extensive experience in listening tests and can be considered to be critical listeners.

3.2.2 Apparatus

The perceptual evaluation took place in the hemi-anechoic Spatial Audio Lab of the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) at McGill University in Montreal, Canada. The room is 5.40 m (W) 6.40 m (L) 3.60 m (H) with a measured Reverberation Time (RT60) of 0.09 seconds, and was outfitted with two rings of 14 Genelec 8020A loudspeakers (Genelec, Iisalmi, Finland, frequency range 48 - 20,000 Hz).

A photograph of the room and loudspeaker array are shown in Figure 3.1. A slightly asymmetric spacing was used due to the available mounting hardware. Care was taken that the loudspeakers were in symmetric pairs around the listening position (with the exception of the loudspeakers placed in the median plane directly in front of and directly behind the listener). Symmetric left / right pairs of loudspeakers were placed at 24, 54, 76, 104, 124, and 154 degrees. To best approximate the planar assumption of the diffuse field model, the radius was the largest that could be fit in the room. The lower ring was placed directly below the upper ring, and the height of the low-frequency driver center was 1.25m and 2.85m, respectively. A two-dimensional schematic of the array is shown in Figure 3.2.

To insure that forward axes of the loudspeakers were all oriented towards the listening position it was necessary to use a 40cm riser platform. The seat of the chair was 50cm above this platform, and the table used for the computer mouse was 80cm above the platform. The listening position had a height of approximately 1.45m for most participants and was level with the high-frequency driver of the loudspeaker. The upper ring of loudspeakers was elevated approximately 32 degrees relative to this position and any of the direct sound that did not interact with the listener was largely incident on the lowest portion of the absorptive walls. The radius of the lower ring was 2.2m from the participant's head location and the upper ring had a radius of 2.5m. To compensate for this geometric discrepancy the lower ring was delayed by 1ms and the upper ring was level compensated 1dB. Identical broadband noise was routed individually to each loudspeaker and measured to be within 1dB ($\pm$ 0.5dB) of each other using the Bruel & Kjaer Type 2250 Sound Level Meter with the 4189 omnidirectional microphone capsule and the LAF (SPL, A-weighted, Fast) setting. It was in this measurement process that the level compensation of top ring was determined.

Stimuli were played on a Apple Mac Pro using a RME HDSPe MADI audio interface. The playback levels of the stimuli were approximately 66 dB for all cases and were verified using a sound level meter. The user interface was programmed in MATLAB.

3.2.3 Stimuli

The stimuli for the test were generated in MATLAB using the white Gaussian noise generator. The 28 distinct channels of uncorrelated noise were lowpass filtered using a 2nd-order Chebyshev Type 1 filter with -3dB values of 14kHz to better approximate natural reverberation. To avoid quickly fatiguing participants with a transient attack it was necessary to use a linear attack ramp of 170ms (8192 samples at $F_s = 48$ kHz). A frequency-independent exponential decay corresponding to a 1.8 second RT60 was used for the subsequent temporal

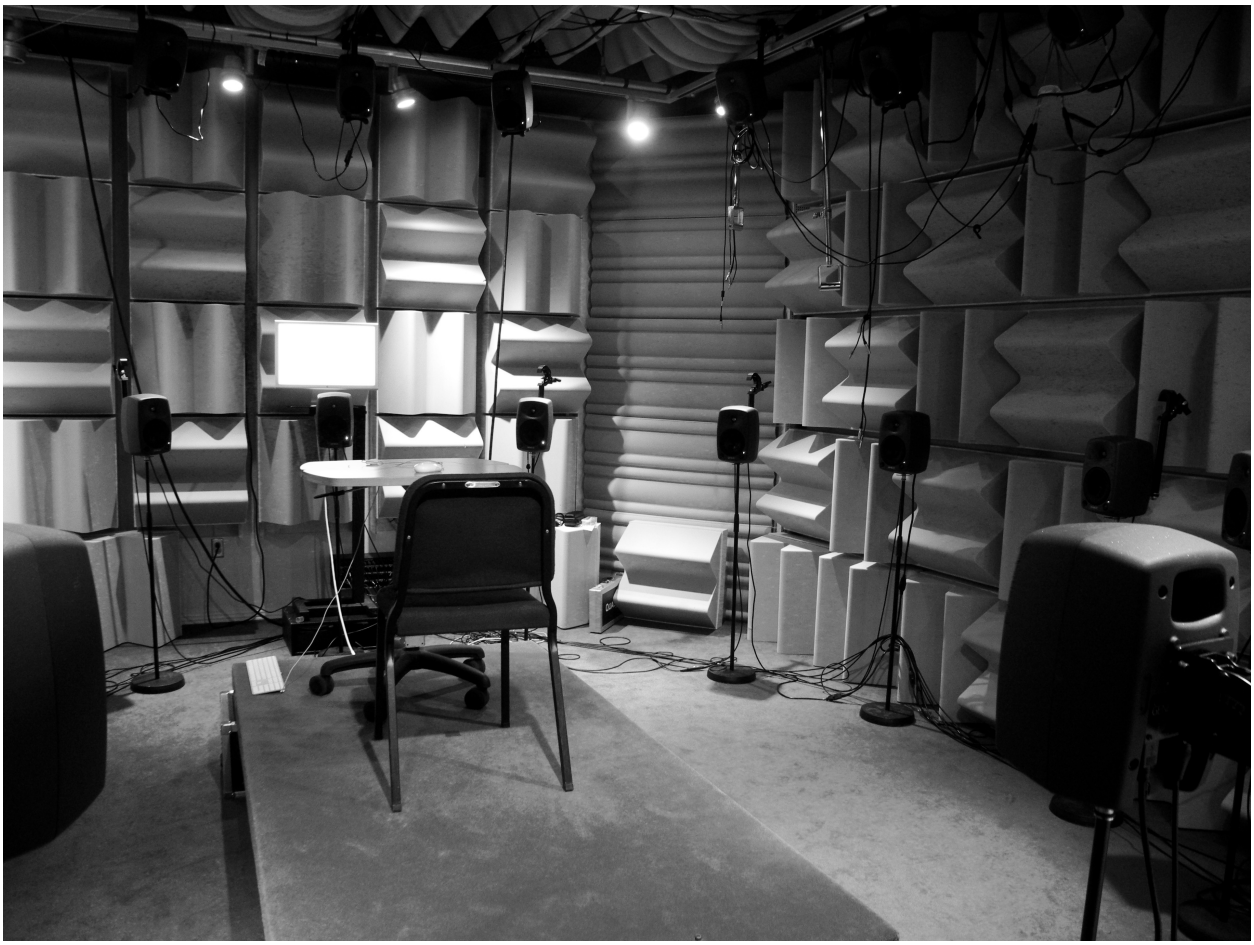

Figure 3.1: Experimental apparatus used for the listening test consisting of two rings of Genelec 8020A loudspeakers. The participants sat on a riser platform in the middle of the setup to be aligned with the high-frequency drivers of all loudspeakers. Subsequent signal measurement using a dummy head as well as a B-Format microphone were done in approximate position of the participant's head.

duration.

Segments of the 28 loudspeaker array were systematically reduced in level for three different conditions: lateral, frontal, and height. A schematic diagram of the array is shown in Figure 3.2. For the lateral and frontal conditions, identical segments of the lower and upper rings were reduced. For lateral segments (left or right), the loudspeakers at 54, 76, and 104, and 124 degrees were reduced in amplitude and the left and right stimuli were counterbalanced, described below. For the frontal condition, all loudspeakers were reduced in level relative to the five frontal loudspeakers. Both the upper and lower rings were reduced

identically. For height, the entire lower ring was reduced in amplitude relative to the upper ring. The gain of the entire array was then compensated to maintain equal power at the listening position according to the formula:

$$g_{comp} = \sqrt{\frac{N_{chn}}{(N_{chn} - N_{attn}) + (g_{attn}^2 N_{attn})}} \tag{3.3}$$

where g_{comp} is the compensation gain, g_{attn} is the attenuation gain of the reduced segment of loudspeakers, N_{chn} is the total number of channels (28), and N_{attn} is the number of attenuated channels. All scenarios and attenuation gains yielded identical SPL to within 1dB and were verified using the Bruel & Kjaer Type 2250 SPL meter.

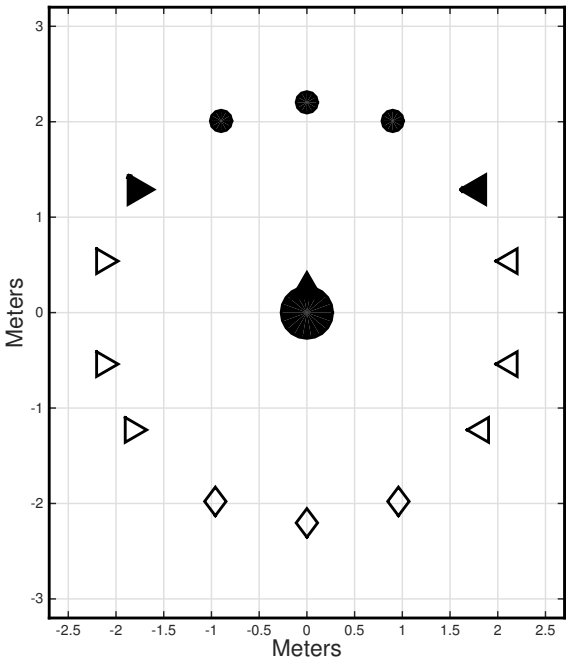

Figure 3.2: The lateral segments (left or right) are depicted using inward-facing triangles while the frontal condition is depicted using solid circles. The loudspeaker pair at 54 degrees was used in both the lateral and frontal condition and is depicted with a solid triangle. The upper and lower rings were treated identically for these conditions. In the height condition the entire lower ring was reduced in amplitude relative to the upper ring. 0 degrees corresponds to the loudspeaker directly in front of the listener.

3.2.4 Procedure

Participants were tested individually. For each participant, the entire experiment took approximately 90 minutes and was divided into a training session and three sessions corresponding to the lateral, height, and frontal conditions. Data from the training session was not included in the analysis. Each session had 72 trials and participants were allowed a short break in between. The training session used the two most disparate stimulus levels of lateral (both left and right) as well as height. The subsequent sessions were counterbalanced across participants using permutations of the three conditions - [lateral, frontal, height], [frontal, height, lateral], and [frontal, lateral, height].

The method of constant stimuli with an ABX task was used. In each trial, participants were presented with three sounds (labelled A, B and X) and asked to indicate if X=A or X=B. The reference was presented randomly as either A or B; X was also randomized to be either the reference or stimulus. In the case of frontal and height, 7 stimulus levels gave rise to 6 pairs, each of which was presented 12 times. In the lateral case, the left and right counterbalancing of 7 stimulus levels gave 12 pairs, each of which was presented 6 times. In analysis, no difference was observed between the data from the left and right conditions and the results were aggregated. In order to insure that every pair was presented an equal number of times, the trials were organized in blocks of N pairs (6 or 12 depending upon the condition) and the pairs were randomized within the block using the MATLAB random permutation function. A consequence of this organization was that all other pairs were presented at least once before a given pair would be heard again.

The levels are shown in Table 3.1 and were chosen such that the extreme case was obvious to most listeners. In pilot testing, some meticulous participants listened to the stimuli more than 40 times per trial while others listened only once. This discrepancy in attention led to erratic perceptual threshold results, and, as such, the test was limited to maximally two listens of A, B, and X.

3.2.5 Data Analysis

For each participant and each condition, the proportion of correct identification was calculated and a cumulative Gaussian psychometric functions was fit using the Palamedes MATLAB routines [55] and a maximum likelihood criterion. The perceptual threshold and slope were allowed to range freely, the guess rate was assumed to be 50%, and the lapse rate was allowed to range between 0.0 and 0.06 as advocated in [78]. Exemplar fits are shown in

Condition	Reference	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Lateral	0 dB	-0.125 dB	-0.375 dB	-0.875 dB	-1.875 dB	-3.875 dB	-7.875 dB
Height	0 dB	-0.5 dB	-1.0 dB	-2.0 dB	-4.00 dB	-8.00 dB	-16.00 dB
Frontal	0 dB	-0.5 dB	-1.0 dB	-2.0 dB	-4.00 dB	-8.00 dB	-16.00 dB

Table 3.1: Levels of gain reduction for three conditions. For the lateral condition, the loudspeakers at 54, 76, 104, and 124 degrees were reduced in amplitude for both the upper and lower rings. For the frontal condition, all loudspeakers were reduced in level relative to the five frontal loudspeakers. Both the upper and lower rings were reduced identically. For height, the entire lower ring was reduced in amplitude relative to the upper ring.

Figure 3.3. The resulting thresholds were accepted or rejected based on the criteria that the percent correct was greater than 75% for the highest stimulus level, that the threshold was not greater than the highest stimulus level in each condition, and that the deviance of the fitted psychometric function was likely to have come from the data ($p > 0.05$) based on a "ground truth" deviance distribution generated from a Monte Carlo simulation as described in [78]. For each condition, outliers greater than two standard deviations from the mean were replaced with the mean value. Following these criteria, one participant was rejected for the lateral condition (5%), four participants were rejected in the height condition (21%), and no participants were rejected for the frontal condition.

The perceptual threshold was computed for each participant in each condition as the x-axis value corresponding to the 50% point of the cumulative Gaussian function [26]. Given the 50% guess rate assumed in an ABX test, this corresponds to an ordinate of approximately 75% depending on the best-fitting lapse rate. These perceptual thresholds were then analyzed with a one-way unbalanced ANOVA and revealed significant differences between the conditions ($F(2,49) = 25.7$, $p < 0.0001$). Post-hoc Tukey-Kramer multiple comparisons ($p < 0.05$) revealed that height had a perceptual threshold that was significantly higher than either lateral or frontal. The perceptual thresholds are shown for the lateral, height, and frontal conditions in Table 3.2.

3.2.6 Results

The perceptual thresholds are shown for the lateral, height, and frontal conditions in Table 3.2. We emphasize that the perceptual thresholds reported here correspond to the dB reduction in a segment of the loudspeaker array, detailed above in Sections 3.2.3 and 3.2.4. Signal analysis in the listening position is presented next in Section 3.3.

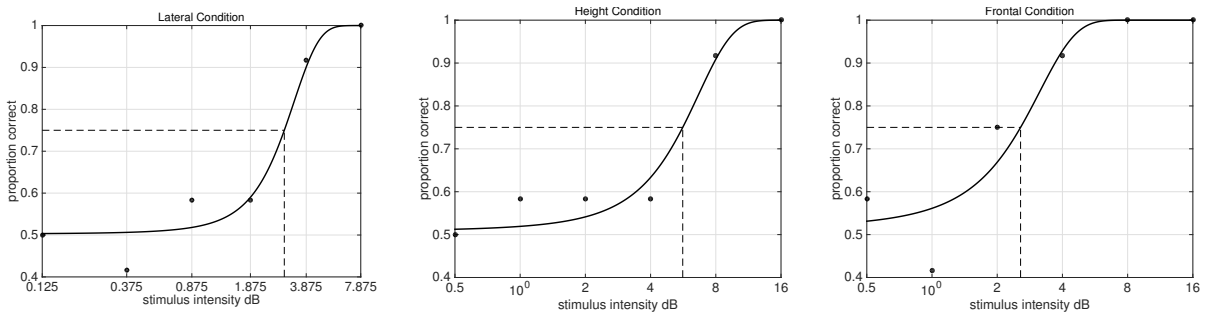

Figure 3.3: Exemplar psychometric functions for participants in the lateral, height, and frontal conditions. The dots are the percent correct for a particular stimulus level, the solid black curve is the fitted psychometric function. The dashed ordinate line correspond to the 1/2 value of the cumulative Gaussian function which is approximately 75% for the ABX task. The perceptual threshold corresponds to the dashed perpendicular line drawn to the abscissa.

Condition	Perceptual Threshold
Lateral	-2.5 dB
Height	-6.8 dB
Frontal	-3.2 dB

Table 3.2: Perceptual thresholds for the lateral, height, and frontal conditions at two significant digits.

3.3 Signal Analysis

The perceptual thresholds determined in the above listening experiment correspond to the dB level reduction in the segments of loudspeakers shown in Figure 3.2. The dB levels observed in the listening position result from the acoustic superposition of all loudspeakers and any interaction that they have with the test participant or microphone. In order to better understand the interaural cues available to the listeners, all stimuli levels were recorded using a dummy head. This analysis is presented immediately below in Section 3.3.1. Section 3.3.2 presents the analysis of both the experimental stimuli and an acoustic space using opposing cardioid microphones and demonstrates that the perceptual thresholds found in the listening experiment can be found in existing acoustic spaces.

3.3.1 Dummy Head Recordings

The signals for all stimuli levels were recorded using the Bruel & Kjaer Head and Torso Simulator (HATS) dummy head in order to better understand the interaural cues available

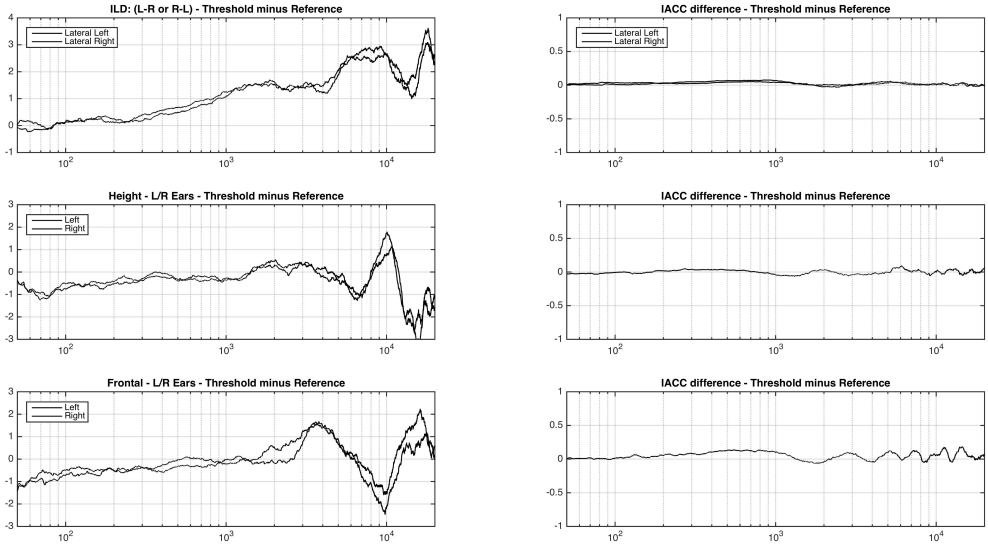

Figure 3.4: The left column shows the dB difference between the stimuli level above the threshold and the reference for both the right and left ears (L/R) for the height, frontal, and lateral cases. The right column shows the observed differences for IACC. The frequency axis is in Hz. A 2% MATLAB smoothing filter was used to improve readability. The HATS dummy head was placed in the position of the listener.

to the listeners. The signals for the stimulus level immediately above the threshold were used as a conservative estimate of the dB changes that are audible. For the lateral condition, the perceptual threshold was -2.5 dB and the next stimulus level was -3.875 dB. For the height condition, the perceptual threshold was -6.8 dB and the next stimulus level was -8.0 dB. For the frontal condition, the perceptual threshold was -3.2 dB and the next stimulus level was -4.0 dB. For each case, the Fourier Transform of the reference and perceptual threshold signals were computed and the dB power was computed for both spectra. The reference was then subtracted from the threshold to determine the audible dB difference. The top row of Figure 3.4 shows the dB level differences and IACC differences for the lateral condition, height is shown in the second row, and frontal is shown in the last row.

The lateral condition used stimuli that were asymmetric with respect to the median plane. The interaural level difference (ILD) was computed and then the difference between the threshold and the reference was taken. Below 300 Hz, there are very small differences in ILD because the long wavelengths associated with these frequencies are able to diffract around the head. Above this, a steady increase to 3 dB at 10 kHz is observed, a wide notch

at 15 kHz with a minimum of +1 dB, and a return to 3 dB at 20 kHz. The broad increase is due to the increase in level on the ipsilateral side of the HATS dummy head as well as shadowing on the contralateral side. The notch is likely due the particular features of the HATS dummy head. There is very little change in interaural cross-correlation (IACC).

The left and right ear signals from the height condition are extremely similar as is expected of symmetric stimuli. There is a modest decrease of less than -1 dB up to 1 kHz, and a mild increase of less than 1 dB near 2 kHz. This increase likely accounted for by considering the influence of the shoulder reflection as an estimated path difference of 17 cm gives constructive interference at approximately 2 kHz. This will be most influential for the loudspeakers located laterally, and, since these are only a few of many, the effect is overall quite mild. Most notable is the sharp peak of approximately 2 dB near 10 kHz and the subsequent notch of -2 dB near 12 kHz. These spectral features may be due to a differing excitation of the cavum conchae and fine details of the pinnae and head shape. It is also possible that the reduction at the highest frequencies is due to occlusion of the ear canal. There is very little change in interaural cross-correlation (IACC).

The left and right ear signals from the frontal condition are also extremely similar, as would again be expected of the symmetric stimuli. There is an increase in the dB power between 2 kHz and 6 kHz with a maximum value of 1.5 dB at approximately 3.5 kHz. There is also a notch of approximately -2 dB centered around 10 kHz, and an increase back towards +2 dB from 13 kHz to 17 kHz. These spectral features are likely due to the emphasis of forward-incident energy which will excite the cavum conchae differently than the uniform energy distribution. Finer details in the differences will depend on individual features of an individual's pinnae and head shape. Extremely modest increases (≈ 0.15) in IACC are observed around 1, 3, 5, 7, 12, and 16 kHz and are likely due to the fact that forward-incident energy will have similar ear signals and is, as a result, more correlated.

3.3.2 Three-Dimensional Room Impulse Responses

The objective of this section is to demonstrate that the perceptual thresholds found in the listening experiment can be found in existing acoustic spaces. The B-Format microphone [39] is a widely available 1^{st}-order spherical microphone that can be interpreted as a combination of an omni-directional (monopole) and bi-directional (dipole) microphones. A dipole microphone is typically phase and frequency compensated by the mass-dominated ribbon or diaphragm response ($1/jk$) which results in a signal proportional to the acoustic velocity [17]. The four channels of the B-Format microphone are:

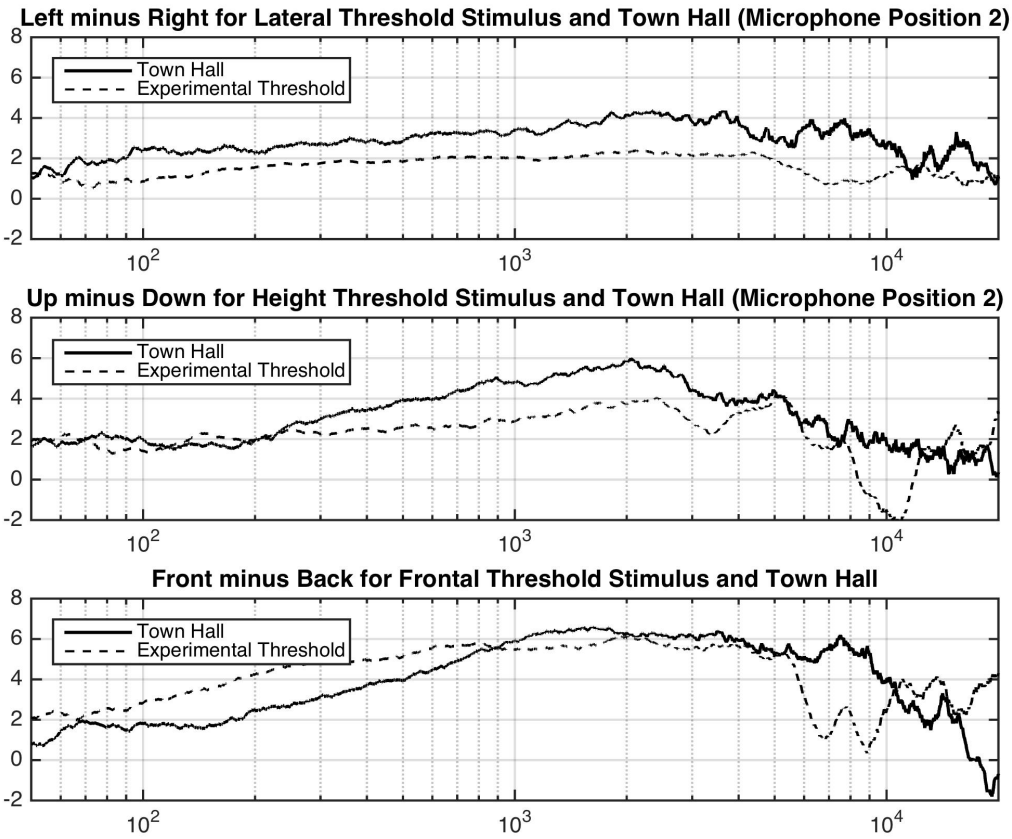

Figure 3.5: Frequency-dependent differences in dB power for opposing coincident cardioids for both the experimental stimulus (dashed line) and an RIR of a shoebox style concert hall with the early response removed (solid line). For all experimental conditions (lateral, height, and frontal), the level differences correspond to the stimulus intensity above the perceptual threshold. A 2% MATLAB smoothing filter was used to improve readability, and a description of the microphone signals is given in the text.

$$W = \hat{p} \tag{3.4}$$

$$X, Y, Z = \hat{\mathbf{v}}_{x,y,z}\rho c = \frac{-\nabla_{x,y,z}\hat{p}}{jk} \tag{3.5}$$

The velocity v or pressure gradient $\nabla\hat{p}$ can be formed in any direction by taking a linear combination of the X, Y, and Z signals. Summing $\hat{\mathbf{v}}\rho c$ with the pressure (W) creates a cardioid microphone in any given desired direction. This directivity pattern is receptive to

traveling planar energy according to $1/2(1 + \cos\theta)$ where θ is the center of the pattern. Subtracting the dB power spectrum of patterns with opposing directions gives an indication of the differences in directional energy.

The plots shown in Figure 3.5 show directional differences in diffuse fields determined by subtracting the dB magnitude square response from opposing coincident cardioid microphones formed from B-Format recordings of both the experimental stimuli and an acoustic space. This was done for the x-axis (Left - Right), the y-axis (Front - Back) and the z-axis (Up - Down). The test stimulus was considered to be the diffuse part of a room impulse response and recordings were made for the intensity level above the perceptual threshold. Because these spectra had modest features specific to the measurement set up, the threshold responses were equalized by subtracting the reference level. This is compared with the same directional differences in Auckland Town Hall with two different microphone positions. To insure that only the diffuse field reverberation was considered, the direct sound was replaced by a 3 ms rectangular window of silence and the subsequent signal was faded in using the same 170 ms ramp that was used for the experimental stimuli. Auckland Town Hall is a shoebox style concert hall and was chosen because directional differences could be found for all experimental conditions (lateral, height, and frontal). It is noted that a number of other acoustic spaces had comparable directional differences on at least one axis.

It is important to note that the dB differences observed by opposing coincident cardioids differ substantially from the perceptual thresholds from the loudspeaker array. This is due to the way in which each individual loudspeaker interacts with the microphone directivity patterns as well as the fact that only segments of the loudspeaker array were reduced in level. For the lateral condition, the opposing cardioid differences are only slightly less than the perceptual threshold. For the height condition, the opposing cardioid differences are approximately half of the perceptual threshold. For the frontal condition, the opposing cardioid differences are nearly double the perceptual threshold.

The lateral condition was evaluated by comparing the dB level of opposing cardioid microphones on the y-axis (Left - Right). The level differences at the stimulus level above the perceptual threshold are approximately 1 dB at 100 Hz, rise gently to 2 dB at 1 - 5 kHz, and gradually return to 1 dB for higher frequencies. Measurement of the same differences in the town hall RIR follow a nearly identical frequency contour but range between 2 and 4 dB.

The height condition was evaluated by comparing the dB level of opposing cardioid microphones on the z-axis (Up - Down). The level differences at the stimulus level above

the perceptual threshold are approximately 2 to 4 dB for all the spectrum and have a gentle maxima near 2.5 kHz. Measurement of the same difference in the town hall RIR follows a similar contour and has a maximum difference of approximately 6 dB at 2 kHz. The abrupt 2 dB notch at 10 kHz in the perceptual threshold difference spectrum is likely due to destructive interference from the floor or seating platform in the experimental apparatus. Neither surface is acoustically treated.

The frontal condition was evaluated by comparing the dB level of opposing cardioid microphones on the x-axis (Front - Back). The level differences at the stimulus level above the perceptual threshold are approximately 6 dB between 1 and 6 kHz. Measurement of the same differences in the town hall RIR yielded approximately 6 dB between 1 kHz and 8 kHz. The town hall is between 1 and 2 dB less than the experimental threshold below 700 Hz.

3.3.3 Summary of Signal Analysis Results

The dB differences made in segments of the loudspeaker array do not correspond directly to dB differences in the listening position. The signal changes observed by the HATS dummy head are subtle. The maximum is an ILD of 3 dB at 10 kHz for the lateral condition. Depending on the frequency, most other differences are in the range of 0 to 2 dB. IACC is largely unaffected, with the only substantial difference in a small frequency range of the frontal condition. The B-Format microphone was used to form opposing cardioid microphones on all three cartesian axes and was compared with a realistic acoustic scenario. Directional differences well above the perceptual threshold were found for the lateral and height conditions; directional differences commensurate with the perceptual threshold were found for the frontal condition.

3.4 Discussion

Both the perceptual and physical literature typically assume an equal directional distribution of diffuse energy. The findings in this paper, however, demonstrate that the human auditory system is sensitive to directional variation in the diffuse field and that commensurate differences can be found in some acoustic spaces. The perceptual threshold for the lateral condition is -2.5 dB, for the height condition is -6.8 dB, and for the frontal condition is -3.2 dB. The lateral condition had the lowest perceptual threshold, this is likely explained by the broadband level differences between the left and right ears. Both the height and frontal conditions, on the other hand, were symmetric with respect to the median plane and

instead lead to frequency dependent changes that are largely identical at both ears. This may account for the fact that the height condition's perceptual threshold was nearly three times greater than the lateral condition. However, the perceptual threshold for the frontal condition - being physically more similar to the height condition - is comparable with the lateral condition. The modest increases of IACC throughout frequency are approximately 1/3 Pollack's JND value of 0.4 (for IACC $\approx$ 0.0) [54]. However, it is possible that this cue, in tandem with frequency-dependent spectral changes, accounts for the discrepancy between the frontal and height conditions; further investigation is merited.

The dB levels observed in the listening position result from the acoustic superposition of all loudspeakers and any interaction that they have with the test participant or microphone. As a result, the dB differences made in segments of the loudspeaker array do not exactly correspond directly to dB differences in the listening position. The stimuli level above the perceptual threshold was recorded using a dummy head in order to better understand the interaural cues available to the listeners. The observed differences were found to be subtle. The maximum is an ILD of 3 dB at 10 kHz for the lateral condition. Depending on the frequency, most other differences are in the range of 0 to 2 dB. IACC is largely unaffected, with the only substantial difference in a small frequency range of the frontal condition. A B-Format microphone was used to form opposing cardioid microphones for the stimuli level above the perceptual threshold as well as for a shoebox style concert call. Directional differences well above the perceptual threshold were found for the lateral and height conditions and comparable directional differences were found for the frontal condition. This strongly suggests that non-ideal diffuse field reverberation plays at least some role in the perception of complex spaces.

The stimuli used in this experiment are similar to the diffuse component of a room impulse response. It is unrealistic in the fact that the direct and reflected sound would be present in most acoustic scenarios. While reflections do influence the color and width of the primary auditory event, their influence in localization is perceptually mitigated by the precedence effect. It is unclear if such a suppression occurs for the diffuse sound given the typical division in perception between auditory source width (ASW), typically associated with the first 80 ms of an RIR, and listener envelopment (LEV), typically associated with the RIR after 80 ms. The experiment by Bradley and Soulodre demonstrating that participants could independently identify conditions of ASW and LEV [11] leads to the speculation that non-ideal diffuse field reverberation may be audible during running music or speech. It is expected that non-ideal diffuse field reverberation would be audible in the temporal gaps of

music or speech.

The height and frontal conditions correspond to energetic distributions that would influence the lateral fraction of a concert hall. Having a greater amount of energy incident from above or the front implies that less energy is incident from the sides. The diffuse sound is typically assumed to be perceptually important after 80 ms implying that the height and frontal conditions may have perceptual relationships with listener envelopment.

The analysis by means of cardioid microphone in a diffuse field has a wide angular aperture. Similarly, the loudspeaker segments used in the experiment are relatively wide. It is not yet understood to what degree smaller angular gaps such as an open window are perceptible. The angular differences in the measurements and stimuli plausibly correspond to an acoustic scenario involving one or two walls with extreme differences in absorptive coefficients.

3.5 Conclusions and Future Work

Both the perceptual and physical literature typically assume an equal directional distribution of diffuse energy; the objective of this study was to understand the sensitivity of the human auditory system to directional variations in a non-ideal diffuse field.

The physical and perceptual results of this experiment suggest that directional energetic differences are a component of complex acoustic spaces and should be appropriately modeled. A three-dimensional reverberation strategy based on physically-inspired decorrelation filters and specially treated B-Format RIR has been developed and is used in conjunction with geometric models of the direct and reflected sound. The strategy referred to as Diffuse Field Modeling has been described in [60] and has been validated with listening tests in [65]. It has additionally been used in an experiment on auditory motion perception using both wave field synthesis and vector-base panning loudspeaker setup [13]. It is not explicitly a physical model of an acoustic space but instead relies on the statistical description of reverberation to simulate the observations that would be made on a microphone array in a diffuse field. It considers both the RT60 of the RIR (linked to frequency autocorrelation of the decorrelation filters) as well as the loudspeaker geometry (linked to frequency-dependent channel correlation). Given the existing evaluations, it appears to be an appropriate tool for perceptual research and evaluation.

Dummy head measurements of the level differences occurring in the experimental stimuli above the perceptual threshold are on the order of 1 to 2 dB for most frequencies. Mea-

surements of an acoustic space using opposing cardioid microphones demonstrates that level differences corresponding to the perceptual thresholds can be found in practice.

The results of this study suggest that non-ideal diffuse field reverberation may be a relevant factor in spatial impression. However, because the stimuli used in this experiment are similar to the diffuse component of a room impulse response taken in isolation, further evaluation is necessary to determine the perceptual thresholds in the presence of direct sound and reflections. It is additionally necessary to evaluate the effect of different types of excitations, as transients sounds are likely to differ significantly from sustained sounds.

Acknowledgements

We would like to thank Sennheiser Innovation and Technology for funding the initial phase of this research. We would also like to thank Marshall Day Acoustics for providing us with three-dimensional room impulse responses having directional differences in the reverberant diffuse field. Finally, we would like to thank CIRMMT for the outstanding facilities and technical staff Yves Méthot and Julien Boissinot for their assistance.

Interim Summary

The results of the experiments presented in Chapters 2 and 3 directly informed the direction of the third contribution presented in Chapter 4 and, for this reason, will be briefly summarized. The Hamasaki Square [29] is a room microphone technique that has interpretation as a sparse sampling of a bounding surface and is intended to recreate the reverberant diffuse field for the 5.1 (3/2 stereo) configuration. In Chapter 2 a perceptual experiment showed that this technique has two major advantages compared to the Electronic Time Offset (ETO) technique adapted from Williams' MMA [79]. When combined with a realization of the Steinberg and Snow's "Acoustic Curtain" [34], the Hamasaki Square was rated as having greater distance than all alternative techniques. When combined with either the acoustic curtain or spot microphone techniques, the Hamasaki Square was rated as being more enveloping than ETO or dry techniques.

Aside from the acoustic intensity of the direct path, the direct to reverberant ratio is the primary cue used by the auditory system to judge distance [6][48][80]. We note, however, that the alternative techniques (ETO and acoustic curtain, Hamasaki Square and spot microphones) were judged during the preparation of the stimuli to have commensurate reverberation. The high distance rating can be attributed to a number of factors. The reproduced reverberant diffuse field likely had more plausible spatial autocorrelation [53] and, as investigated by Bronkhorst and reviewed by Zahorik [80], this led to a proper balance of direct (deterministic) and reverberant (stochastic) signals at the ears of the participants. That is, the reproduced reverberation was a facsimile of true, spatial reverberation and not simply a temporal (but non-spatialized) recording of reverberation.

The presence of early reflections seems to have contributed to the perception of distance. The acoustic curtain in conjunction with Hamasaki Square would capture more early reflections than either the acoustic curtain and ETO or the spot microphones and Hamasaki Square. Moore, discussing experiments by Von Bekesy, indicates that the timing and presence of early reflections contributed to the perception of distance [48]; to the knowledge of

the author there are no other examples of this in the literature. This said, an experiment by Bronkhorst [12] demonstrated that participants' distance estimates increased for binaural room impulse responses having only 3, 9, and 27 reflections. These results are subsequently interpreted in terms of the direct to reverberant ratio, however typical physical definitions of the mixing time are much longer than this [53][7]. This research might better be interpreted as an experiment regarding early reflections. Returning to the Hamasaki Square, the presence of reflections may have served only to widen and color the primary auditory event as is suggested by literature regarding the precedence effect and Auditory Source Width [6][48][11]. In this case, the ratio of deterministic direct sound to stochastic reverberant sound may have been the primary distance cue, but the width of the auditory event better corresponded to the test subjects expectations of such a distance. More research on this topic is clearly required.

Envelopment was rated higher for the two treatments using the Hamasaki Square compared to ETO or the dry treatments. The low rating of the dry treatment is not surprising, however the ETO treatment was judged in preparation to have commensurate reverberation. With respect to the Listener Envelopment from the concert hall literature, these treatments would be expected to be similar [11]. The most likely explanation for the rating corresponds to the auditory events associated with loudspeaker arrays with modest ($k \approx 0.3$) channel coherence values discussed in Appendix B.2.2 and shown in [6]. In this case, the wide auditory event associated with the diffuse field reverberation would be associated with the acoustic space itself and lead to the perception of being enveloped in an acoustic space.

Both the perceptual and physical literature typically assume an equal directional distribution of diffuse energy [6][53]. It was hypothesized that the auditory system might be sensitive to directional energetic differences in reverberation and that such differences might be found in existing acoustic spaces. The perceptual experiment and physical measurements presented in Chapter 3 provide support for both hypothesis. The thresholds for perceptible energetic differences were estimated for a lateral condition (-2.5 dB), for a height condition (-6.8 dB), and for a frontal condition (-3.2 dB). The lateral condition had the lowest perceptual threshold, this is likely explained by the broadband level differences between the left and right ears. Both the height and frontal conditions, on the other hand, were symmetric with respect to the median plane and instead lead to frequency dependent changes that are largely identical at both ears. This may account for the fact that the height condition's perceptual threshold was nearly three times greater than the lateral condition. However, the perceptual threshold for the frontal condition - being physically more similar to the height

condition - is comparable with the lateral condition.

The dB levels observed in the listening position result from the acoustic superposition of all loudspeakers of the experimental setup and any interaction that they have with the test subject or microphone. As a result, the dB differences made in segments of the loudspeaker array do not exactly correspond directly to dB differences in the listening position. The stimuli level above the perceptual threshold was recorded using a B-Format microphone and then compared to similarly recorded room impulse responses of an existing shoebox style concert hall. By forming opposing cardioid microphones on the three cartesian axes, directional differences above the perceptual threshold were found for the lateral and height conditions and comparable directional differences were found for the frontal condition. This strongly suggests that non-ideal diffuse field reverberation plays at least some component in the perception of complex spaces.

The results of Chapter 2 suggest that the Hamasaki Square is an excellent bench mark for other techniques. This said, the channel-based limitation to 5.1 (3/2 stereo) is logistically forbidding when one considers the multitudes of loudspeaker configurations available for consumer, cinema, and academic uses. The physical and perceptual results of Chapter 3 suggest that directional energetic differences in reverberation are a component of complex acoustic spaces and should be appropriately modeled. The Diffuse Field Modeling algorithm presented in Chapter 4 was directly informed by these findings. Every attempt was made to substantiate the algorithm in terms of known physical acoustics (Appendix A), and, similar to the Hamasaki Square, the "direct" paradigm of a (virtual) microphone array in a listening space directly routed to a loudspeaker array was maintained. Chapter 5 validates the developed algorithm as a component of a physically-plausible virtual acoustic model that can be systematically adapted to various loudspeaker arrays.

Chapter 4

Diffuse Field Modeling using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part I Algorithm

While a variety of techniques exist to record and reproduce point sources, there is not a systematic tool for the recording and reproduction of diffuse sound fields. Diffuse Field Modeling uses decorrelation filters based on the statistical description of reverberation to "virtualize" an array of outward-facing cardioid microphones from linear combinations of a B-Format microphone under the assumption that the audio is diffuse. DFM is presented in two publications: this Part I presents the algorithm and Part II [65] reports the perceptual evaluation. The current paper describes the design of the physically-inspired decorrelation filters and validates the reconstructed diffuse fields through a numerical simulation based on the Kirchhoff / Helmholtz Integral for plausible radiation assumptions. It is demonstrated that the resulting fields have the expected spatial autocorrelation, that the channels of the array have the expected frequency-dependent correlation, and that a correspondence can be established between the introduced frequency autocorrelation and the $RT60$ of the recorded diffuse field. This correspondence heavily influences the spatial impression.

4.1 Introduction

Physically, the reverberant diffuse field is incoherent energy incident from all directions [53] and is perceptually related to an auditory event heard "everywhere" [6]. It is common practice for sound recording engineers to use differing microphone strategies for the specular and diffuse fields [73]. Diffuse Field Modeling (DFM) is a physically-inspired method of approximating a diffuse field and is intended to create a natural-sounding room effect for arbitrary loudspeaker configurations. It is intended to function in parallel with point source techniques such as amplitude panning [6], vector-base amplitude panning (VBAP) [57] or wave field synthesis (WFS) [1]. It can be used for spatial content creation in the current sound recording paradigm and is extensible to dense loudspeaker arrays such as those found in WFS. The ability to systematically render an approximation of a diffuse field from a four channel B-Format room impulse response (RIR) makes it a suitable component of a high-definition object-based transmission format.

A numeric simulation of a diffuse field is shown in the left panel of Figure 4.1. The objective of DFM is to systematically approximate the diffuse field of a recording space inside of a distinct listening space. The fundamental concept that enables this is a virtual microphone array linked to a real loudspeaker array at conceptually-coincident locations. Each virtual microphone and loudspeaker pair defines a discrete point of measurement and secondary radiation on a bounding surface. Through either measurement or simulation, the values of the pressure $\hat{p}$ and pressure gradient $\nabla_n \hat{p}$ on the bounding surface can be used to reconstruct an acoustic field inside the bounded volume using the Kirchhoff / Helmholtz Integral (K/H) [53][1]. The center panel of Figure 4.1 shows a double-layer K/H reconstruction using values of pressure and pressure gradient measured from the diffuse field simulation. The rightmost panel also uses measured values, however it uses a practical single-layer approximation of the K/H based on outward-facing cardioid microphones routed to inward-facing cardioid radiators. This approximation is developed in Section 4.3.4 and used in the validation by simulation in Section 4.5. The single-layer approximation of the Kirchhoff Helmholtz Integral is used throughout this paper. A dense array of actual microphones in the recording space routed to conceptually-coincident loudspeakers in the listening space will produce an approximation of the diffuse field. However, recordings made with this microphone array would be inflexible to alternative loudspeaker arrangements and would be a substantial (and likely implausible) logistical undertaking. It is desired that the recorded diffuse fields be systematically adaptable to differing loudspeaker configurations. It would be logistically elegant if the diffuse fields could be recorded from a single point in the recording space.

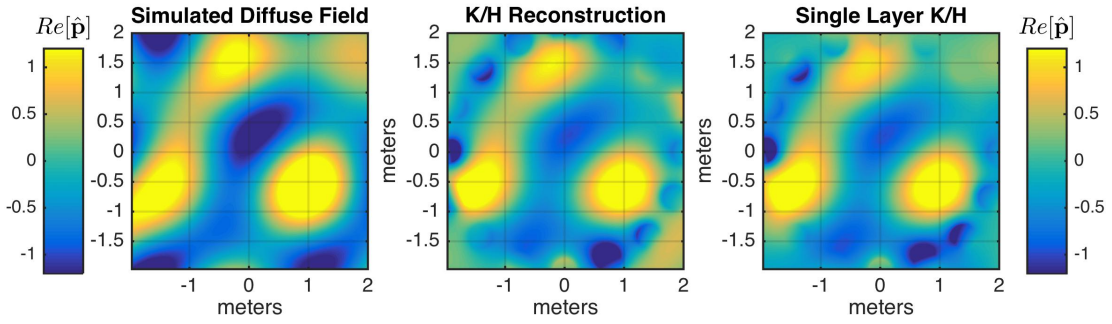

Figure 4.1: The leftmost panel shows a simulated Diffuse Field at 160 Hz created from the superposition of $Q = 1024$ plane waves of random direction, phase, and magnitude. The simulations presented in Section 4.5 are based on such a field. The center panel shows a double-layer Kirchhoff Helmholtz reconstruction of this field with values of pressure and pressure gradient measured on the bounding surface. The rightmost panel shows the single-layer cardioid microphone to cardioid loudspeaker approximation used throughout this paper. 16 spherical secondary radiators were used for the center and right panels, more detail is provided in Sections 4.3.4 and 4.5.

In DFM, room impulse responses (RIR) are recorded with a B-Format microphone [39] and are considered to be an observation of $\hat{p}$ and $\nabla_{x,y,z}\hat{p}$ at the origin. It is desired to be able to simulate a large number of observation points on the bounding surface and to simulate a diffuse field for as much of the listening area as possible. This can be done if the RIR can be assumed to be diffuse according to the physical definition of randomly incident plane waves or a superposition of exponentially decaying modes. Linear combinations of the B-Format microphone are used to form outward-facing cardioid microphones and are virtually translated to other locations in space by decorrelation filters based on the statistical description of reverberation. These filters simulate values of $\hat{p}$ and $\nabla_{x,y,z}\hat{p}$ that statistically resemble the values that would be measured from a real microphone array in an acoustic diffuse field and are informed by the $RT60$ of the RIR as well as the geometry of the virtual array. Because the outward-facing virtual cardioids are different at various points on the bounding surface, they are able to discern differences in inward-bound acoustic energy. This allows the representation of non-ideal diffuse fields where more energy is incident from certain

directions.

The assumption of diffuseness is valid after the mixing time of a room [53][7]. To achieve this constraint, the RIR must be treated to remove the direct path and any strong specular reflections. Techniques such as matching pursuit [37] could be used to isolate the direct path and reflections. Alternatively, the diffuse field can be "faded in" using the physical / perceptual metric Normalized Echo Density [32]. The latter has been used in the development of DFM in conjunction with geometrically modeled early reflections. In the future, it is desired that DFM be extended to live recording scenarios and will require a technology to remove a sufficient proportion of the specular sound. The assumption that the RIR is physically diffuse is particularly powerful as it allows multiple sources to be individually spatialized using geometric information while a single diffuse layer is used for all sources.

At low frequencies in an acoustic diffuse field the constituent random plane waves interact with neighboring microphones which leads to frequency-dependent correlation between the microphones. DFM uses a spatial filtering strategy based on the loudspeaker geometry to avoid unnecessary cancellation and to ensure physically-plausible and frequency-dependent value of correlation between channels. The degree to which the introduced variation changes with frequency is controlled by adjusting a frequency-dependent exponential decay envelope on the decorrelation filter and is tuned to a "most appropriate" value by considering the reverberation time $RT60$ of the RIR. This tuning profoundly changes the spatial impression of the modeled diffuse field and represents an engineering trade-off as the convolution of modulated exponential decays in the decorrelation filters and RIR introduces temporal distortion to the RIR. Some creative control over spatial impression is gained by adjusting the DFM filter parameters.

The companion paper [65] presents a perceptual evaluation of DFM as a component of a physically-plausible virtual acoustic model that can be systematically adapted to various loudspeaker arrays. Two distinct experiments were conducted. The first experiment used a 20-loudspeaker array with a 16-channel lower ring (identical to the simulations presented in Section 4.5) and a 4-channel height ring. The direct path and reflections were modeled geometrically and positioned using VBAP. The objective was to evaluate various treatments of DFM and to ensure that there were no major issues in practice. 16 sound recording professionals participated in the experiment and verbal analysis and semantic scales were used to evaluate the stimuli. The findings indicate that it is necessary to model reflections in conjunction with DFM, and the majority of sound recording professionals found the initial implementation to be useable in practice. A second experiment used the 5.1 loudspeaker

configuration and evaluated DFM against the Hamasaki Square [29][73], which was chosen for its physical similarities to DFM. 26 participants (8 sound recording professionals, 18 music students from McGill University) participated in the experiment and a drawing technique and semantic scales were used to evaluate the stimuli. No significant differences were found with respect to the Hamasaki Square, which indicates DFM is a viable and flexible solution for practice. DFM has additionally been used in an experiment on auditory motion perception using both wave field synthesis and vector-base panning in a 48-loudspeaker setup [13]. The direct path and reflections were modeled interactively while DFM was used for the diffuse sound and provided the impression of an auditory space without objectionable artifacts.

This paper is organized as follows. Section 4.2 discusses the state of the art including common recording strategies for the room effect, the B-Format microphone, existing technologies, and the perception of directional energetic differences in diffuse fields which DFM is able to reproduce. Section 4.3 discusses the statistical description of reverberation that is used to inform the design of the decorrelation filters. This section also discusses the Kirchhoff / Helmholtz Integral and practical radiation assumptions such that the virtual microphone array can be linked to the real loudspeaker array. Cardioid microphone to cardioid loudspeakers is proposed as a reasonable assumption for common situations. Section 4.4.1 ties together all reviewed topics and Section 4.4.2 specifies the signal processing design of the decorrelation filters. Finally, Section 4.5 validates the decorrelation filters through numerical simulation.

4.2 State of the Art

4.2.1 Hamasaki Square

A room microphone technique known as the Hamasaki Square records and reproduces both the spatial and temporal aspects of the reverberation by placing four bi-directional microphones in a square pattern with the nulls oriented to the direct sound and the forward direction oriented to the lateral walls of the recording space. These four microphones are directly routed to the left, right, left surround, and right surround of the 5.1 format (3/2 stereo). It allows independent manipulation of the direct sound and provides a reasonable facsimile of lateral reflections. Channel correlation can be adjusted by changing the microphone spacing [29]. Theile [73], discusses the broad acceptance of the Hamasaki Square and notes that broadband channel correlation values of approximately 0.3 is associated with "subjective envelopment" while lower values lead to "clouds of reverberation" around the

loudspeakers.

From a physical perspective, the Hamasaki Square can be interpreted as a sparse sampling of a bounding surface directly routed to secondary radiators in the listening space. The frequency-dependent channel correlation will depend on the wavelength of the sound and the chosen microphone spacing. For typical spacings, it is correlated below approximately 200 Hz. This view will be used extensively in DFM.

Additionally, the Hamasaki Square was evaluated in perceptual rating experiments [27][66] by 16 experienced listeners in McGill University's Sound Recording program. Statistical analysis by ANOVA confirmed that this technique was rated as having superior listener envelopment and distance perception compared to a surround microphone technique adapted from the William's MMA [64],[79]. These high ratings demonstrate that there is perceptual utility in spatialized and systematical rendered reverberation.

However, the Hamasaki Square - and all other channel-based techniques - are limited to the format they were designed for, in this case 3/2 stereophonic reproduction. Adaptions of the direct sound recorded with spot microphones are the most straight-forward, adaptions of array-based techniques are considerably less so. The aim of DFM was to generalize the positive aspects of the Hamasaki Square to arbitrary loudspeaker formats by forming a virtual outward-facing microphone array conceptually coincident with the available loudspeaker array.

4.2.2 B-Format Microphones

The B-Format microphone [39] is a 1^{st}-order spherical microphone and can be interpreted as a combination of multipoles or a truncated series of spherical basis functions (SBF.) Multipoles are directional derivatives of an acoustic field or an ideal point source (Green's Functions) and are used frequently to discuss microphones and acoustic radiators [53]. SBF are solutions to the Helmholtz Equation in spherical coordinates and can be used to decompose a sound field into an orthogonal series. This series can be determined by observations near the origin and used to extrapolate $\hat{p}$ and $\nabla_{x,y,z}\hat{p}$ for both inward-bound and outward-bound acoustic waves [28][50]. The SBF interpretation is used extensively in Ambisonics (see Section 4.2.3), however the multipole interpretation will be used throughout this paper. The relationship between SBF and multipoles is demonstrated by Gumerov and Duraiswami in [28]. For the 1^{st}-order case, multipoles and SBF are proportional within a constant.

B-Format is discussed by various authors in terms of acoustic velocity $\hat{\mathbf{v}}_n$ or pressure gradient $\nabla_n\hat{p}$, the relationship between the quantities is defined by Newton's 2^{nd} Law

$-\nabla_n \hat{p} = j\omega\rho\hat{\mathbf{v}}_n$. A dipole microphone is typically phase and frequency compensated by the mass-dominated ribbon or diaphragm response $(1/jk)$ which results in a signal proportional to the acoustic velocity [17]. In this paper, $-\nabla_{x,y,z}\hat{p}/jk$ is used in the simulations. The four channels of the B-Format microphone are:

$$W = \hat{p} \tag{4.1}$$

$$X, Y, Z = \hat{\mathbf{v}}_{x,y,z}\rho c = \frac{-\nabla_{x,y,z}\hat{p}}{jk} \tag{4.2}$$

The directional velocity v or pressure gradient $\nabla\hat{p}$ can be formed in any direction by taking a linear combination of the X, Y, and Z signals. Summing $\hat{\mathbf{v}}\rho c$ with the pressure (W) creates a cardioid microphone in any given desired direction. An identical argument can be made for a cardioid radiator pattern by summing a phase and frequency compensated dipole radiator with a monopole radiator. Monopole and dipole microphones and radiators are fundamental to the Kirchhoff Helmholtz Integral discussed in Section 4.3.4. K/H is the theoretical underpinning of the virtualized microphone and loudspeaker array discussed throughout this paper.

4.2.3 Ambisonics and Wave Field Synthesis

The B-Format microphone is typically associated with Ambisonics which is based upon the fact that a sound field can be decomposed into a weighted series of spherical basis functions (SBF), see Gumerov and Duraiswami [28] or Nicols [50]. The sound field is typically observed at a single point (assumed to be the origin) by a spherical microphone array and an encoding step determines the weights of the SBF series from the (typically cardioid) capsules of the array. The SBF series is considered as a scalable transmission format, with greater angular resolution and greater radial extent being available from higher-order series. In reconstruction, a "mode matching" decoding step considers the contributions of the loudspeakers (planar, spherical, or arbitrary) at the origin and an inverse matrix is created to weight the contribution of each SBF signal to each loudspeaker [50].

WFS is based on the fact that the knowledge of $\hat{p}$ and $\nabla_{x,y,z}\hat{p}$ on a bounding surface allows secondary sources to reconstruct an approximation of the original sound field using the Kirchhoff / Helmholtz Integral, see Section 4.3.4. Practical constraints force a simplified "single-layer" formulation of the integral with the secondary radiators typically assumed to be omni-directional. Instead of predicting both $\hat{p}$ and $\nabla_{x,y,z}\hat{p}$ on a bounding surface, virtual

acoustic sources are positioned outside the loudspeaker array, and their contribution to each loudspeaker is determined by propagation delay and a spatial windowing function.

The connection between spherical basis functions and the Kirchhoff-Helmholtz Integral is that an infinite SBF series can predict the field values $\hat{p}$ and $\nabla_n \hat{p}$ on the bounding surface. A differentiation property [28] allows the computation of $\nabla \hat{p}$, and, in general, $\nabla \hat{p}$ has an arbitrary value relative to $\hat{p}$. Given this information, the two layers of K/H will reproduce the interior acoustic field perfectly. In all practical cases, the SBF series is truncated and a single-layer radiation assumption must be made. DFM shares a conceptual similarity with Ambisonics and WFS in the fact that an observation at the origin can be used to determine values on a bounding surface. However, when B-Format is interpreted as a spherical basis series it has very little ability to predict numerous independent values of a diffuse field on a bounding surface. Because of this, DFM simulates this lost information using physically-inspired decorrelation filters that statistically resemble what would be measured on the bounding surface.

4.2.4 Existing Technology

The goal of DFM is to observe a diffuse field at the origin and to create a physically-accurate model of the magnitude and phase variation that would be found on a virtual microphone array at other locations. This is done using decorrelation filters which are a synthesis and extension of approaches found in the literature. Spatial filtering ensures the correct frequency-dependent channel correlation and avoids excessive cancelation. A link between frequency autocorrelation and RT60 is developed to allow tuning the decorrelation filters to be most appropriate for the RIR at hand. Kendall created libraries of decorrelation filters by specifying allpass magnitude and random phase $\pm\pi$ in the frequency domain [36]. Even when specifying an allpass response, the frequencies between adjacent "control values" will not be unity due to the random phase variation. To circumvent this problem, Boueri [9] presented a decorrelation strategy based upon time offsets in perceptually-spaced frequency bands.

Spatial Impulse Response Rendering (SIRR) [46][45] and Directional Audio Coding (DirAC) [59] are perceptually-motivated time / frequency processing schemes applied to B-Format microphones. SIRR uses RIR while DirAC uses continuous audio streams. Both use a definition of diffuseness that considers the ratio of acoustic velocity to the overall acoustic energy density. This is computed at each time and frequency step and is used to divide the signal into specular and diffuse buses. As few as two simultaneous plane waves

will be considered somewhat diffuse [59], such as in the early part of an RIR where specular reflections are still distinct or when there are multiple sources.

In reconstruction, the specular bus is sharpened using Vector-Base Amplitude Panning, while the diffuse bus is decorrelated using filters. In SIRR these decorrelation filters were achieved using convolution with exponentially decaying white Gaussian noise with a decay constant of 50ms. In DirAC the same technique was generalized to three distinct frequency bands. The band below 400Hz had a constant of 100ms, from 400 to 1300Hz a decay constant of 40ms and above this a decay constant of 10ms. In the SIRR algorithm, the diffuse layer was assumed to be omni-directional [46]. Subsequent work on the DirAC algorithm [77] extends the diffuse layer to various 1^{st} order directivity patterns, with the observation that directional virtual microphones produced a more authentic reproduction of spaciousness, source localization, and sound color when compared with an omni-directional diffuse layer.

Menzer and Faller [44] present a binaural method using B-Format RIR in conjunction with Head-Related Transfer Functions (HRTF). The method creates individualized binaural room impulse responses (BRIR) by matching the interaural cross-correlation (IACC) associated with the ensemble of HRTF measurements and Energy Decay Relief of the B-Format RIR. Additional work by the same authors in [43] shows that BRIR synthesized with IACC-matched White Gaussian Noise (WGN) are largely indistinguishable from a reference after approximate 20 ms beyond the direct path of the BRIR. Both papers substantiate the fact that correlation statistics largely determine the perceptual affects of reverberation.

4.2.5 Non-Ideal Diffuse Fields

Both physical and perceptual literature typically assumes diffuse field reverberation to have equal energy incident from all directions. The non-ideal case has more energy incident from certain directions. This would be expected in a partially enclosed room, a room with dramatically differing wall materials, or in a concert hall where the floor is covered with highly-absorptive seats (and patrons). It has recently been shown that listeners are sensitive to non-ideal characteristics of diffuse fields. In a discrimination test, temporally shaped decorrelated White Gaussian Noise (WGN) was presented on a semi-spherical array of loudspeakers. The level was systematically varied with direction while the overall signal power remained constant. Experienced sound engineers and audio researchers were found to have perceptual thresholds [26] of -2.5 dB of lateral discrepancy, -3.2 dB of fore-aft discrepancies, and -6.8 dB for height discrepancies [61]. These findings demonstrate the auditory system's ability to perceive spatial energetic directional differences in diffuse field reverberation, and

justifies the need to model non-uniform energy distributions.

Further substantiating the salience of non-ideal diffuse fields are the plots shown in Figure 4.2. A B-Format microphone was used to measure the impulse response of Pollack Hall at McGill University. This hall has variable acoustics and was configured to the "bright" treatment with fewer absorptive curtains. Cardioid microphones were formed in opposing directions along the $\pm x$,$\pm y$, and $\pm z$ directions. The power spectral density (PSD) was computed from a Short-Time Fourier Transform (STFT) with a 256-point Hann window, 512-point Discrete Fourier Transform, and 64-sample hop size [56]. The PSD of the opposing cardioid responses on each axis were then subtracted. The lower plot shows the directional differences in the z direction with a 27 ms (20-point) averaging filter applied across time (STFT hops) and reveals a directional energy difference for higher frequencies. The microphone had been placed approximately 2m above the seats, this difference is likely to be much greater when the microphone is closer to the seated position. The upper plot shows the lateral differences with no filtering and reveals no overall energetic trends. This is expected for a laterally symmetric hall. It is speculated that the rapid modulations of positive and negative directional differences with both time and frequency play a part in the preference for directional microphones in [77]; such differences would be lost if a decorrelated omni-directional microphone was used. More generally, it is speculated that these rapid modulations play a part in perception of spatialized reverberation and may be partially responsible for the high ratings of the Hamasaki Square in [64].

4.3 Acoustic Background

Physical room acoustics is typically given a statistical interpretation [53] and is used in the filter design of DFM in Section 4.4.2. Two statistical characterizations of physical reverberation are particularly important in DFM. Spatial autocorrelation tells us how the pressures p at various locations in space change relative to each other. For a set direction and distance, a high correlation value tells us that pressure is changing in phase and with a similar amplitude. Low correlation values tell us that the pressure is in a reactive phase and/or has differing amplitudes. Negative correlation values tell us that the pressure is changing with similar amplitude but in opposite phase. Frequency autocorrelation tells us the degree to which adjacent frequencies ω and $\omega + \Delta\omega$ change coherently. It is a description of how jagged or smooth a spectrum is; note that the spectrum is different at different locations. In tandem, the two quantities tell us how much an acoustic field changes in space with respect to frequency. Detailed treatments are found in Pierce [53].

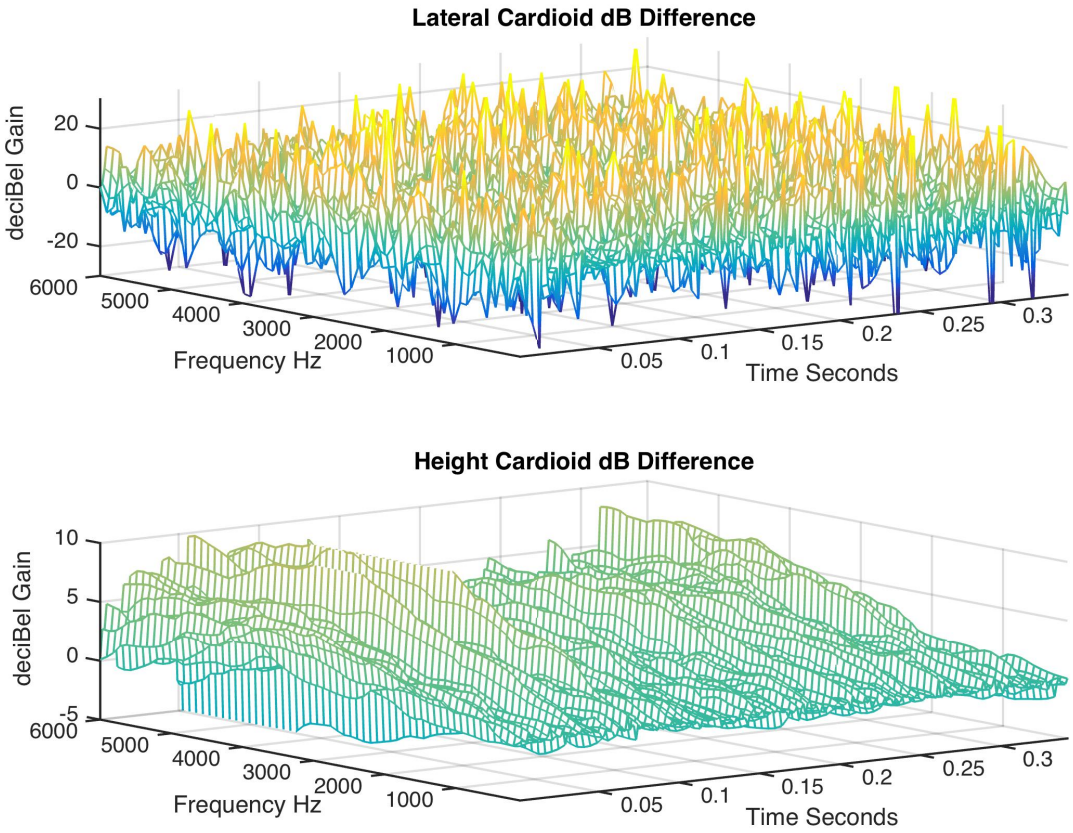

Figure 4.2: dB difference of the power spectral density of opposing coincident cardioids formed from a B-Format microphone placed approximately 2m high in the center of Pollack Hall at McGill University. The impulse response was normalized to have a maximum value of 1.0, the direct sound was removed, and a 170 ms ramp was used to fade in the response. The upper plot shows the difference between opposing cardioids oriented in the $\pm$y direction with no smoothing. Note the fast symmetrical magnitude modulation in either direction. The lower plot shows the difference between opposing cardioids oriented in the $\pm$ z direction smoothed with a 27 ms averaging filter. While modulation is still apparent, note the asymmetrical bias toward the positive z direction, indicating that more sound intensity was incident from above than below. This is likely due to the absorptive seating.

4.3.1 Sabine Reverberation Model

The Sabine reverberation model assumes that reverberant acoustic energy is approximately the same in all spatial regions of the room and that the absorption due to the bounding walls can be treated on average. It is derived in terms of the conservation of acoustic energy, the

acoustic power P that is inputted to the room, and the acoustic power P_d dissipated by the walls bounding the room. The acoustic energy density w is the sum of the potential acoustic energy due to pressure, and the kinetic acoustic energy due to particle velocity. Assuming spatial uniformity of $\overline{w}$ allows a simplified differential equation, and the solution to this is a decaying exponential:

$$\langle p^2(t)\rangle/\rho c \;\approx\; \overline{w}(t) = \overline{w}_{initial} e^{-t/\tau}$$

where $\langle p^2(t)\rangle$ is the spatial average of the pressure squared. The quantity τ is the characteristic decay time and will be used in the design of the decorrelation filters in Section 4.4.2. It is more common, however, to discuss reverberation in terms of $RT60$, which is the amount of time needed for $\langle p^2(t)\rangle/\rho c$ to decay 60dB from its initial value. The quantities are related by $\tau = RT60/6.9$. Sabine's empirical work was later verified by the modal approach to room acoustics [53] discussed below in Section 4.3.3.

4.3.2 Diffuse Field Model

The diffuse field can be represented as a mathematical idealization that describes the field as a superposition of $q = 1, 2, ...Q$ plane waves of random direction, phase, and amplitude. This idealization is physically accurate after the mixing time of the room, that is, after an acoustic source has been reflected many, many times. For constant frequency fields, the complex pressure $\hat{p}$ can be described as a summation:

$$\hat{p} = \sum_{q=1}^{Q} \hat{p}_q e^{\mathbf{n_q}\cdot\mathbf{r}} \tag{4.3}$$

where $\hat{p}_q$ is the complex amplitude of the q^{th} plane wave, $\mathbf{n}_q$ is the directional vector of the q^{th} plane wave, and $\mathbf{r}$ is a vector pointing to the observation. This idealization is used in the numerical simulation in Section 4.5 and was used to generate the left panel of Figure 4.1. For a constant frequency pressure field, the autocorrelation is given by expected value with respect to space of the product of $p(\mathbf{r})$ and $p(\mathbf{r} + \mathbf{\Delta r})$ where r is the observation point and Δr is a fixed offset. This offset will cause the arguments of each constituent plane wave q to proceed differently, in the limit of infinite plane waves this allows us to treat the pressure as a random variable. Mathematical manipulation leads to the familiar description of spatial autocorrelation [53][15]:

$$\langle p(\mathbf{r})p(\mathbf{r} + \Delta\mathbf{r})\rangle = \langle\overline{p^2}\rangle \frac{\sin k|\Delta r|}{k|\Delta r|} \tag{4.4}$$

where k is the wave number and $\langle\overline{p^2}\rangle$ is the mean square pressure averaged over space. Lower frequencies will be more correlated for greater amounts of space, higher frequencies less so. For a given frequency, autocorrelation is periodic with space, though greater distances are considerably less correlated. The same equation arises if we instead fix our observation to a single point in space and consider the sphere defined by the magnitude vector $|\Delta r|$. This formulation is used in the simulation in Section 4.5, a similar approach was used by Elko in [19].

The derivation of Equation A.53 is based on the value of pressure and corresponds to measurement by an omni-directional microphone. Additional considerations regarding the microphone directivity pattern and orientation are shown [19]. Parallel bi-directional microphones at various spacings are correlated to higher frequencies than are omni-directional microphones. Cardioid microphones have correlation determined by the pattern orientation as well as the spacing. For identical orientation, the spatial autocorrelation for bi-directional, cardioid, and omni-directional microphones differ by their frequency shape. For coincident microphones, Faller [23] shows that correlation ranges from 1/3 for opposing cardioids, 2/3 for a 90 degree orientation, and 1 for the identical orientation.

The diffuse field approximation yields an important constraint that can be used to inform the introduction of randomness in DFM. Consider the pressure observed along the vector $\mathbf{r}_{1,2}$ connecting two arbitrary observation points $\mathbf{r}_1$ and $\mathbf{r}_2$. For a given spatial frequency $k = \omega/c$, the maximum phase progression corresponds to the plane wave propagating along $\mathbf{n_q}$ exactly parallel to $\mathbf{r}_{1,2}$. Due to their projection onto $\mathbf{r}_{1,2}$ the arguments of all other diffuse plane waves proceed more slowly and will contribute a lesser value of k. If a collinear array of equally spaced points is considered, the observed spatial frequencies on $\mathbf{r}_{1,2}$ are all k up to an ideal brick-wall cutoff of $k = \omega/c$. If we consider the observations at each point to be a random process, the autocorrelation is a transform pair with the power spectral density [30]. That is, the observed spatial frequencies on a collinear array (or nearly collinear array) in a diffuse field are linked to the spatial autocorrelation of the observations through the spatial Fourier Transform. The use of dipole or cardioid capsules changes the weight of the observed spatial frequencies and thus the spatial autocorrelation of the observations. This is used extensively in the filter design in Section 4.4.2 to low-pass filter the introduced randomness. This, in turn, ensures a physically-plausible and frequency-dependent value of channel correlation.

4.3.3 Modal Model

The modal approach to room acoustics is obtained by finding solutions to the constant frequency wave equation (Helmholtz equation) in a rigidly bounded room. The structure of the modes is determined by the room's geometry. Qualitatively, each mode is a spatial description of the pressure field that will be excited to varying degrees by the location of a source and the difference between the modal frequency and the source's driving frequency. There are many discrete modes spaced over frequency, and the modes typically overlap depending upon their frequency bandwidth.

Rooms with short RT60 values have modes that decay quickly and which have large bandwidths. A given driving frequency will influence a relatively large range of modes. Rooms with long RT60 values have modes which ring for a greater duration of time and which have more narrow bandwidth. As a result, distant driving frequencies have less influence. When the modal resonances are closely spaced relative to their bandwidth, multiple modes are excited for a given input. The Schroeder cutoff frequency is defined as the frequency beyond which the modal spacing is less that 1/3 the modal bandwidth, and, in this case, the summation of discrete modes can be considered a continuous distribution of frequency components.

For a constant frequency the complex pressure field is the superposition of all excited modes and is calculable for a given source and observation point. Similar to the plane waves of the diffuse field idealization, however, the contribution of each individual mode changes if either the source or observation move. Above the Schroeder frequency the field can be described statistically in terms of the frequency autocorrelation and the mean square pressure distribution. The frequency autocorrelation is the expected value of the product of the mean square pressure at two adjacent frequencies averaged over all spatial observation points. For the time average of the squared acoustic pressure $\overline{p^2}(\omega)$ and source frequency ω, this is given by:

$$\langle \overline{p^2}(\omega)\overline{p^2}(\omega + \Delta\omega)\rangle = \langle \overline{p^2}(\omega)\rangle^2 \left[1 + \frac{1}{1 + (\tau\Delta\omega)^2}\right] \tag{4.5}$$

The fact that the frequency autocorrelation depends on the characteristic decay time τ is extremely important in DFM. When the term $(\tau\Delta\omega)^2$ is large, the right hand term is near 1.0. This indicates that the mean squared pressure at adjacent frequencies differs, as would be expected for large frequency spacings $\Delta\omega$ or for long-ringing modes with large values of τ and narrow bandwidths. Conversely, when $(\tau\Delta\omega)^2$ is vanishingly small, the right hand term is near 2.0, which indicates that the mean square pressure at adjacent frequencies is largely

identical, as would be expected for small frequency spacings $\Delta\omega$ or for shorter modes with small values of τ and wide bandwidths. The smoothness of the spectrum plays a large role in the spatial impression, and plays a substantial role in the design of the DFM decorrelation filters. This is addressed in Section 4.4.2.

The mean squared pressure in the modal approximation of a nearly-rigid walled room follows an exponential distribution with probability distribution function $\mu(\overline{p^2})$ of mean squared pressure $\overline{p^2}$.

$$\mu(\overline{p^2}) = \lambda e^{-\lambda \overline{p^2}} \text{ for } \overline{p^2} \geq 0 \tag{4.6}$$

with $\lambda = \frac{1}{\langle \overline{p^2} \rangle}$. The value of $\overline{p^2}$ is always positive, however it is approximately twice as likely for a given value of $\overline{p^2}$ to be below the mean $\langle \overline{p^2} \rangle$ than above the mean. Because of this, the average sound pressure level (SPL) is 2.5 dB lower than the SPL level corresponding to $\langle \overline{p^2} \rangle$. The exponential distribution is used in DFM to define the magnitude squared values of the decorrelation filters, and the 2.5 dB SPL is used to compensate the gain of the filters.

The real and imaginary parts of a modal sound field are uncorrelated, zero-mean Gaussian random variables. As such, all phase values between $-\pi$ and π are equally likely. That is, phase is uniformly distributed on the interval $[-\pi \;\; \pi\,[$. This fact is used in DFM to define the phase of the random components of the decorrelation filters.

4.3.4 Kirchhoff-Helmholtz Integral

An acoustic monopole is defined by the Green's function:

$$G(\mathbf{r}) = \frac{e^{jkr}}{4\pi r} \tag{4.7}$$

The Kirchhoff-Helmholtz Integral Theorem (K/H) relates the acoustic field inside or outside of a bounding surface to values of the acoustic field on that surface [53][28].

$$\hat{p}(\mathbf{r}) = -\iint_S \Big(\; G(\mathbf{r}_s)\nabla_n \hat{p} - \hat{p}\nabla_n G(\mathbf{r}_s) \; \Big) \; \overrightarrow{dS} \tag{4.8}$$

$\mathbf{r_s}$ is the local vector from an arbitrary point on the surface to a point in space $\mathbf{r}$ and ∇_n is the dot product of the gradient operator and the surface normal $\nabla \cdot \mathbf{n}_s$. K/H can be considered an infinite surface array of dipole microphones $\nabla_n \hat{p}$ routed to monopole sources $G(\mathbf{r_s})$ and omni-directional microphones $\hat{p}$ routed to dipole sources $\nabla_n G(\mathbf{r_s})$. The convention of $\mathbf{n}_s$ determines if the integral describes an inward or outward space, and, in the theoretical

limit of a continuous surface, there is no reproduced field in the contrary space. Since we are interested in reproducing the field in the listening area, we choose $\mathbf{n}_s$ to be directed inside the surface. A useful approximation of the dipole is:

$$\nabla_n G(\mathbf{r}) \approx \; jkG(\mathbf{r}) \cdot \mathbf{n}_s \tag{4.9}$$

due to the fact that the gradient results in terms that decay with $(1/r)$ and $(1/r^2)$. In the far field, the $(1/r^2)$ term is typically discarded because it quickly vanishes as we move away from a source. Considering the dot product with the surface normal - and momentarily ignoring the phase shift j - the monopole and dipole are nearly identical when observing the radiators on a vector parallel to $\mathbf{n}_s$ (inward space) and nearly opposite on a vector antiparallel to $\mathbf{n}_s$ (outward space).

A great deal of intuition can be gained by considering the simple case of a plane wave entering a flat surface parallel to the surface normal $\mathbf{n}_s$. The monopole and dipole driving functions have proportional information and respectively drive the dipole and the monopole radiators. There is a phase shift associated with the dipole microphone, and, on the other layer, there is a phase shift corresponding to the dipole radiator. For this simple scenario, the layers are in-phase. In the inward space the two sources are nearly identical, and in the outward space the sources nearly cancel. This effectively forms a cardioid-to-cardioid reception and radiation pattern. If the plane wave were instead exiting the surface (antiparallel to $\mathbf{n}_s$), the driving functions are opposing, and no inward traveling wave would be radiated (though an outward traveling wave would be.) For an arbitrarily shaped surface, the dipole and monopole microphones have differing values and will excite the two radiation layers to differing degrees, making visualization difficult without the aide of computer simulation.

In all practical cases, we do not have the luxury of a double layer Kirchhoff / Helmholtz reconstruction and must form a single driving function for a single radiator. Plane wave assumptions are used in the "classic" view of Ambisonics when determining a given loudspeaker's contribution to the reconstructed field. This has been recently generalized to arbitrary radiation patterns [50]. Omni-directional radiators are typically used for virtual sources in the wave field synthesis literature [1], the non-zero reconstruction outside the K/H surface is ignored. DFM uses cardioid radiation in simulation, this has the benefit of not reproducing as much energy outside of the array, and has slightly lesser lateral radiation. In practice, the available loudspeakers dictate the radiation assumption. The driving function for these radiators is formed from the linear combination of X, Y, Z to form the normal velocity $\hat{\mathbf{v}}_n \rho c = -\nabla_n \hat{p}/jk$ and the surface pressure $\hat{p}$. Single-layer cardioid to cardioid radiation

is defined as:

$$\hat{p}(\mathbf{r}) \approx -\frac{jk}{2} \iint_S \left(\hat{p} + \frac{-\nabla_n \hat{p}}{jk} \right) \left[G(\mathbf{r}_s) + \frac{\nabla_n G(\mathbf{r}_s)}{jk} \right] \overrightarrow{dS} \tag{4.10}$$

The dipole radiator $\nabla_n G(\mathbf{r}_s)$ has been phase-aligned with the monopole radiator $G(\mathbf{r}_s)$ by the term $1/jk$. The additional jk multiplier outside of the integral is needed to maintain approximate equality and the overall phase relationship with Equation 4.8. The product of the cardioid driving function and the cardioid radiator cause extra terms not found in K/H that depend upon the incident field and the observation vector $\mathbf{r}$. The two major terms $\hat{p}\ \nabla_n G(\mathbf{r}_s)$ and $\nabla_n \hat{p}\ G(\mathbf{r}_s)$ are present. The error terms depend upon the direction of the incident acoustic waves and the radiation direction $\mathbf{r}_s$. The division by 2 is required because these error terms are largely redundant with the primary terms for inward-bound radially-directed plane waves.

4.4 Diffuse Field Modeling

4.4.1 Rationale

DFM forms a virtual cardioid microphone array that links the recording space and the loudspeakers of the listening space. It uses physically-inspired decorrelation filters to simulate the variation that would be observed had we made measurements at other points in space. These observations are simulated by assuming the structure of the signal to be diffuse. An enclosing Kirchhoff Helmholtz surface is defined by real-world loudspeakers in the listening space. Linear combinations of the B-Format microphone are translated by way of the decorrelation filters to locations conceptually-coincident with the loudspeakers. This forms a virtual outward-facing cardioid microphone array in the recording space that is linked to the real loudspeakers in the listening space. This virtual linkage is formally described by the cardioid-to-cardioid approximation of the Kirchhoff / Helmholtz Integral developed in Section 4.3.4. It is particularly elegant that outward-facing cardioids can discern inward-traveling acoustic waves in a manner analogous to K/H.

Translating the B-Format microphone to locations on the virtual array relies on the statistical interpretation of reverberation. It is not deterministic values of field quantities that are important but rather the ensemble averages of these field quantities. This interpretation has been put to great use in the frequency-dependent temporal decays of artificial reverberators such as in Jot [35]. DFM extends this to include acoustic field values on a bounding surface. The introduced variation is tuned to have the correct frequency autocorrelation and

profoundly influences the spatial impression of the modeled field. The frequency-dependent channel correlation is ensured to be correct by spatially filtering the introduced randomness. Reconstruction by way of K/H demonstrates that the spatial autocorrelation of the modeled diffuse field is correct.

The introduced variation is based on the statistical descriptions reviewed in Section 4.3.2. The introduced magnitude squared variation is derived from an exponential distribution. The introduced phase variation is equally distributed over $[-\pi \;\; \pi\,[$. The degree to which this variation changes with frequency can be tuned by adjusting frequency-dependent exponential decay constants in the decorrelation filters. This decay constant influences the frequency bandwidth, which in turn influences the frequency autocorrelation. This can then be linked to the physical reverberation time RT60 (Section 4.3) and set to a most appropriate value for the RIR at hand. This is formalized below in Section 4.4.2.

In a diffuse field, the projection of each constituent plane wave will necessarily be different at the individual virtual array locations, and, because there are a very large number of plane waves, the pressure at each location is simulated as a decorrelated random variable. Artificial reverberators are typically discussed as having broadband decorrelation between output channels [35]. In DFM, this corresponds to a virtual microphone array having a very large radius where the constituent random plane waves of the diffuse field to make very different contributions to neighboring microphones. As the radius is decreased, each constituent random plane wave makes more similar contributions to neighboring microphones which leads to frequency-dependent correlation between the microphones. For this reason, DFM uses a spatial filtering strategy based on the loudspeaker geometry to avoid unnecessary cancellation and to ensure physically-plausible and frequency-dependent value of correlation between channels. This is formalized in Section 4.4.3.

An assumption used in DFM is that all of the microphone observations (at the origin) are identical before introducing the variation that would be found at other points in space in an acoustic diffuse field. For adjacent loudspeakers in dense arrays (corresponding to largely overlapping cardioid patterns) this is very nearly true and corresponds to loudspeakers that have important interaction in the acoustic field. The assumption is questionable for distantly spaced points on the array, with the worst case being opposing cardioids. However, we expected the corresponding locations on the virtual array to be largely uncorrelated and the spatial filtering does little to change a distant location. Further, these loudspeakers interact very little. For this reason, we consider our assumption to be appropriate for practice.

The introduction of spectral magnitude variation is analogous to that which would be

found if an array of room microphones were set up in an actual acoustic diffuse field; a similar point is made by Schroeder in his seminal paper on artificial reverberation [68]. It is the case that the introduced spectral magnitude variation causes coloration of the individually synthesized channels, however this coloration tends to disappear as the number of loudspeaker channels increases because the mean of the exponential distribution specifying the magnitude squared is unity. Coloration is also related to how well the criterion of diffuseness is enforced - specular components should not be given diffuse field properties.

4.4.2 Bank of Bandpass Filters

The objective of DFM is to systematically approximate the diffuse field of a recording space inside of a distinct listening space. The concept that enables this is a virtual microphone array linked to a real loudspeaker array at conceptually coincident locations. The filters presented in this section provide a physical basis for decorrelation strategies and are a synthesis and extension of the decorrelation filters presented by Kendall [36], Merimaa [46], and Pulkki [59] reviewed in Section 4.2.4. The loudspeaker driving signals for the array of $i = 1, 2, ...I$ evenly-spaced loudspeakers are formed by applying an individual, physically-inspired decorrelation filter to each of the I virtual cardioid microphones. The "action" of the decorrelation filters is to simulate the field values that would have been measured by a real microphone array in an acoustic diffuse field.

Each of the I decorrelation filters is created by superimposing a large number of randomly weighted modulated exponential decays, the transform pair of this function is a bandpass filter. The exponential decay is an excellent temporal shape because the convolution of a long exponential decay (in this case a room mode) with a shorter exponential decay (a constituent bandpass filter) results in a function with a gradual "bloom" that is then dominated by the longer exponential decay. That is, the decorrelation filters do not change the frequency-dependent RT60 and there is no abrupt "gating" at the end of the RIR tail. The gradual bloom is temporal distortion, however it is believed to be acceptable for diffuse audio. This can be validated empirically in MATLAB and the formal mathematics will be presented in a future publication.

The goal of DFM is to minimize temporal distortion while introducing physically-inspired variation that yields the correct spatial autocorrelation (SAC) and frequency autocorrelation (FAC) values. The smoothness of both the RIR's and decorrelation filter's spectra can be described by FAC. Creating filters with FAC equivalent to the FAC of the RIR allows us to introduce frequency-dependent variation as would be found in the diffuse field of the RIR.

The multiplication of the filters time sequence by an exponentially decaying temporal envelope will smooth the spectrum because of the time / frequency duality of multiplication and convolution [56]. It is this duality that gives a means of controlling the introduced FAC by adjusting the decay of the filter envelope. This same adjustment gives perceptually salient control over spatial impression, providing a trade-off between temporal distortion and a modeled diffuse field.

Minimum-phase decorrelation filters designed to introduce a physically-plausible magnitude squared variation do not synthesize a diffuse field because the filters introduce too little phase variation. The shaped noise used by Merimaa [46] and Pulkki [59] does not allow for the spatial filtering at low frequencies, nor did it allow for the specification of the desired magnitude and phase distributions. The method based upon randomly weighted complex exponential decays was chosen because it allows straight-forward control over spatial impression and the introduced temporal distortion.

The design method was formulated based on $l = 1, 2, ...L$ equally-spaced frequency points. Note that these filters could be made to use asymmetric spacing if desired. At each of these L points, a temporal exponential decay ν_l is specified and then modulated by the frequency Ω_l. This results in a bandpass filter with center frequency Ω_l with impulse response:

$$b_l[n] = \ e^{-(1/\nu_l - j\Omega_l)n} \ u[n] \tag{4.11}$$

where $u[n]$ is the step function. The z-transform of the l^{th} normalized gated exponential is:

$$B_l(z) = \frac{1 - a_l}{1 - a_l z^{-1}} \tag{4.12}$$

with $a_l = e^{-(1/\nu_l - j\Omega_l)}$. The frequency response of B_l is evaluated at $z_m = e^{j\Omega_m}$ where the subscript $m = 1, 2, ...M$ denotes the evaluation frequency as opposed to the bandpass center frequency l. Shorter exponential decays have wider bandwidths, longer decays have more narrow bandwidths. The individual band-pass filters are then superimposed to form a "bank of bandpasses" decorrelation filter for an individual channel i:

$$d_i = \sum_{l=1}^{L} A_{i,l} \ b_l \tag{4.13}$$

where $A_{i,l}$ is a complex gain containing both randomized magnitude drawn from the exponential distribution and randomized phase value drawn from a uniform distribution. This

results in frequency response:

$$D_i(e^{j\Omega_m}) = \sum_{l=1}^{L} A_{i,l} B_l(e^{j\Omega_m}) \tag{4.14}$$

Note that all L bandpass filters contribute to every observation frequency Ω_m, and each bandpass is scaled by a random weight $A_{i,l}$. The resulting spectral randomness described by $D_i(e^{j\Omega_m})$ is necessarily different from the originally introduced values $A_{i,l}$. Remarkably, it does not substantially change the mean or variance of the distribution. The exponential decays ν_l are specified such that low frequencies have longer decays and higher frequencies have shorter decay using logarithmically specified control values and the linear interpolation function in MATLAB. That ν_l is reduced according to the logarithm of frequency is justified by the fact that many physical systems have increased damping with frequency and follow an identical pattern. Two examples are frequency-dependent RT60 and the damping of the cochlea, approximated by Equivalent Rectangular Bandwidth (ERB) [48]. The frequency autocorrelation can be measured at the M observation frequencies from the ensemble of the I decorrelation filters spectra $D_i(e^{j\Omega_m})$ (abbreviated $D_i(m)$.) Following the physical definition of frequency autocorrelation in Equation A.54, the quantity $\Gamma(m)$ is defined:

$$\Gamma(m) = \frac{\sum_I |D_i(m)|^2 |D_i(m+1)|^2}{(\sum_I |D_i(m)|^2)^2} \tag{4.15}$$

where the mean square pressure has been replaced by the magnitude squared of the frequency response and the terms have been rearranged for normalization. It is desired that the introduced normalized FAC be likened to physically arising normalized FAC of the room or RIR:

$$\Gamma(m) \approx \frac{\langle \overline{p^2}(\omega) \overline{p^2}(\omega + \Delta\omega) \rangle}{\langle \overline{p^2}(\omega) \rangle^2} \tag{4.16}$$

for the case when the discrete frequency Ω_m corresponds to the continuous frequency ω and the bin spacing $\omega_{bin} = F_s/M$ corresponds to the continuous frequency offset $\Delta\omega$. This can be likened to a frequency-dependent RT60 by further manipulating Equation A.54 and by considering the relationship between τ and $RT60$:

$$RT60 = 6.9\tau = \frac{6.9}{\Delta\omega} \sqrt{\left[\frac{\langle \overline{p^2}(\omega) \overline{p^2}(\omega + \Delta\omega) \rangle}{\langle \overline{p^2}(\omega) \rangle^2} - 1 \right]^{-1} - 1} \tag{4.17}$$

Substituting the normalized FAC measured from the ensemble of filters yields:

$$RT60(m)_{est} = 6.9\tau = \frac{6.9}{\omega_{bin}}\sqrt{[\Gamma(m) - 1]^{-1} - 1} \tag{4.18}$$

It is the relative smoothness of the spectrum of the RIR that allows the shorter decorrelation filter to introduce variation appropriate for a given $RT60$. By manually adjusting ν_l and observing the corresponding $RT60_{est}$ the decorrelation filters can be adjusted to a most appropriate value for the RIR at hand. It is emphasized that $RT60_{est}$ is not intended to influence the reverberant decay time of the room, but rather to serve as a guide: when the filters are adjusted to the RIR at hand, the variation across space will change with frequency in a way analogous to a (real) microphone array recording a diffuse field in that acoustic space. The parameterization of ν_l is the primary adjustment in DFM, the most accurate physical approximation is obtained when the left and right sides of Equation 4.16 are equal. Preliminary tunings for the room impulse responses of Redpath Hall and Tanna Hall at McGill University are shown in Figure 4.3.

4.4.3 Spatial Filtering

The spatial filtering in DFM pertains to low frequencies and adjacent loudspeakers; it is particularly relevant to loudspeaker arrays with close spacing such as a wave field synthesis system. It also pertains to sparse loudspeaker arrays such as 5.1 in approximately the lowest two octaves of the audible frequency range. Considering a densely sampled acoustic curtain [34],[63], it is known that there is a limit to the possible change in magnitude and phase between two adjacent sampling points in a diffuse field. Said differently, the observed spatial frequencies between two points in a diffuse field are all k up to an ideal brick-wall cutoff of $k = \omega/c$. The reasons for this are discussed in Section 4.3.2.

Artificial reverberators are typically discussed as having broadband decorrelation between output channels [35]. In DFM, the concept of a virtual microphone array is used to find the output channels for a real loudspeaker array. In this interpretation, broadband decorrelation corresponds to microphones spaced at an effectively-infinite distance or to microphones with an extremely narrow pattern oriented in different directions. This would be the case for a virtual microphone array using outward-facing cardioid microphones and having a very large radius. As the radius is decreased, each constituent random plane wave of the diffuse field begins to interact with the neighboring microphone which leads to frequency-dependent correlation between the microphones. It is this frequency-dependent channel correlation that

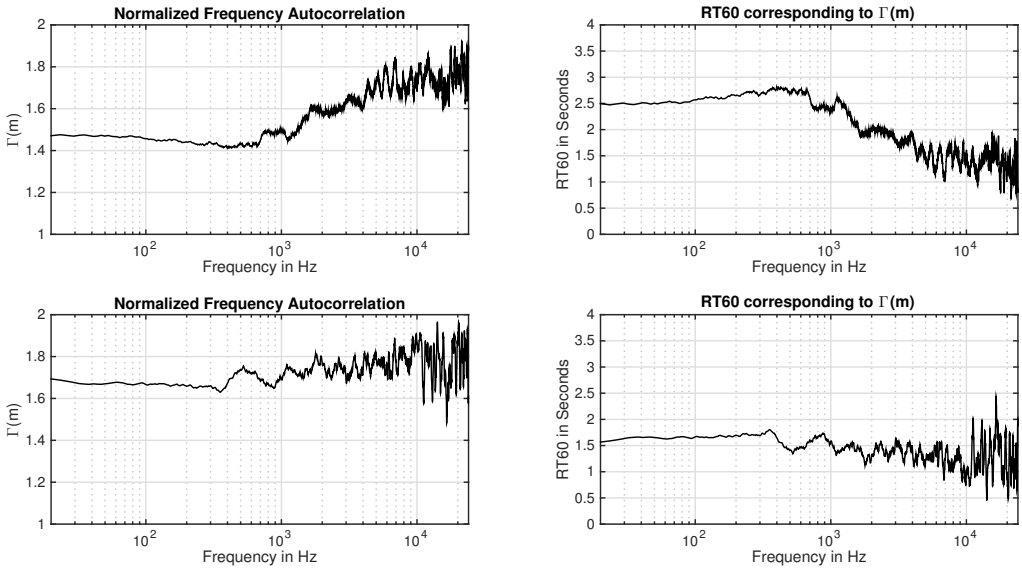

Figure 4.3: The left column shows examples of the function $\Gamma(m)$ defined in Equation 4.15 and which describes the frequency autocorrelation measured from the ensemble of I decorrelation filters. The right column shows the correspondence of the FAC value to physical RT60 values defined in Equation 4.18. The filters are preliminary tunings for the room impulse responses of Redpath Hall (top) and Tanna Hall (bottom) at McGill University; they are obtained by manually manipulating ν_l until the corresponding $RT60_{est}$ matches that of the RIR.

the spatial filtering in DFM attempts to preserve for arbitrary loudspeaker spacings. The major advantages are 1) it avoids unnecessary cancellation for dense loudspeakers at low frequencies, 2) it avoids uneven energy distribution among locations and frequencies where there is not cancellation, and 3) it ensures a physically-plausible and frequency-dependent value of correlation between channels. For aesthetic decisions, the virtual array distance can be adjusted from "physically-plausible" to broadband decorrelation (large virtual array distances.)

The spatial filtering assumes regular loudspeaker spacing. For a circular geometry this would be computed using the array distance and number of channels. A more sophisticated two-dimensional approach would be necessary for a spherical loudspeaker array. It is likely sufficient to approximate a spherical array as a series of circular rings, in which case the presented method can be used.

After the introduction of uncorrelated random variation $A_{i,l}$, these values must be filtered

across i (space) such that the maximum spatial frequency corresponds to the bandpass center frequency $k_l = \Omega_l/c$. Because the virtual array is discretized to loudspeaker locations, mathematics identical to discrete temporal filtering can be used. The continuous spatial frequency is $k = \omega/c$ radians/meter, the spatial sampling period is X meters/sample and the spatial sampling frequency is $G_s = 1/X$ samples/meter. The normalized frequency is $\kappa_l = k_l X$ radians/sample and the Nyquist frequency is $G_s/2$ samples/meter. The ideal lowpass is:

$$h[n] = \frac{\sin \kappa_l n}{\kappa_l n} \tag{4.19}$$

The lowpass filter is applied by circular convolution across the channels of the circular array:

$$A'_{i,l} = \sum_{\iota=1}^{I} A_{\iota,l} h[i - \iota] \tag{4.20}$$

where ι is the convolution index and the channel count I truncates the lowpass. In practice, this is compensated by moderately scaling κ. The filtering is applied for discrete (temporal) frequencies Ω_l up to the spatial Nyquist $G_s/2$ of the loudspeaker array.

The power spectral density of $A_{\iota,l}$ is broadband with respect to k. The power spectral density of $A'_{i,l}$ is defined by the k-domain Fourier Transform of $h[i]$. The inverse Fourier Transform of the power spectral density is the autocorrelation of the random process [30], which is to say that $h[i]$ will also determine the spatial autocorrelation of the channels. Figure 4.4 shows the frequency-dependent channel correlation for circular arrays of 64 (left panel) and 16 loudspeakers (center panel.) This graph was created with a set spacing r_i between the loudspeakers for frequencies $2\pi\omega_l$. It could have equivalently been graphed in terms of $k_m = \omega_l/c$ or $k_m r_i$. The left panel of Figure 4.4 shows that the channel correlation plotted in terms of $k_m r_i$ is in very good agreement with the theoretical spatial autocorrelation defined by Equation A.53. The spatial filtering interacts with the FAC calculations in a non-obvious way. The physical derivation of FAC assumes averaging over all points in a room. Spatial filtering is applied to a constrained geometry with a finite number of discrete points. The FAC is higher than desired because spatial filtering causes the values of $A'_{i,l}$ to be more similar across space (channels i) which causes the FAC to be somewhat higher across evaluation frequencies Ω_m. In practice there are always a limited number of channels I, though FAC does approach the non-filtered value for very high channel counts. To mitigate this, slightly longer filter decay constants ν_{long} are used for all spatially filtered frequencies because their narrow bandwidth better preserves the desired FAC.

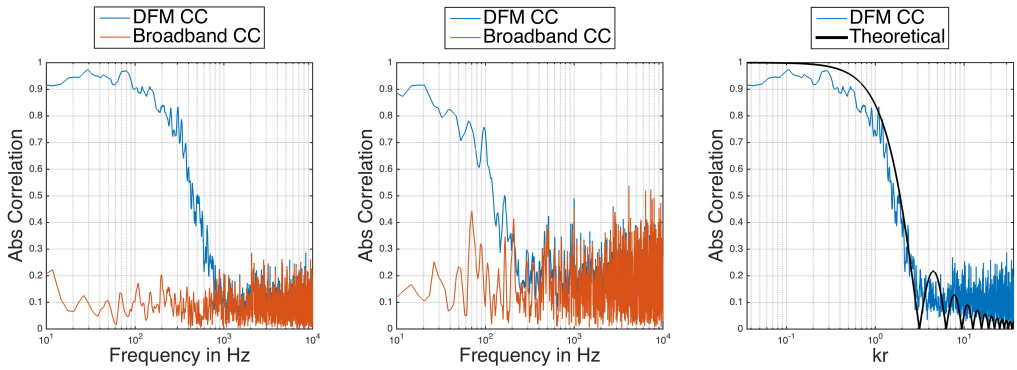

Figure 4.4: Average channel correlation between adjacent loudspeakers for 64-channel (left panel) and 16-channel (center panel) arrays with radius 2m. The blue line is with spatial filtering, while the burnt orange line is broadband decorrelation. The rightmost panel shows the average channel correlation for the 64-channel array plotted with respect to kr as well as the theoretical spatial autocorrelation according to Equation A.53.

4.5 Validation by Simulation

4.5.1 Method

The theoretical discussion in Section 4.3.4 was presented in spherical coordinates as it better reflects the reality of loudspeakers of various heights in three-dimensional space. For viewing, the numerical MATLAB simulation is evaluated on the horizontal plane and makes use of the stationary phase approximation borrowed from the wave field synthesis literature [1]. This approximation requires a multiplication by $\sqrt{\frac{-2\pi}{jk}}$ to compensate for the fact that there are only radiators on a two-dimensional ring as opposed to a three-dimensional surface. Note that the sign convention of the Green's Function (Equation 4.7) causes a change in sign relative to [1] and that we do not adopt their amplitude correction term. The discrete formulation of the Kirchhoff-Helmholtz Integral is obtained from Equation 4.8 by replacing the integrals with summations over a virtual array with fixed radius R and I equally spaced channels. An arc of the circumference Δs is equal to $2\pi R/I$. Δs is left inside the summation to maintain

the relationship to the continuous space formulation, $\mathbf{n}_i$ is the local inward surface normal, and $\mathbf{r}_i$ is the local vector from one of the I locations.

A diffuse field was simulated with $Q = 1024$ plane waves of random direction and complex amplitude. Plots of the real component of the complex pressure field were computed for a dense cartesian grid, while spatial autocorrelation was computed using an array of concentric circles, described below. A virtual array was formed using the cardioid microphone to cardioid loudspeaker approximation of K/H defined in Equation 4.10 of Section 4.3.4. The driving functions for the radiators are driven by linear combinations of pressure p and normal velocity $\mathbf{v}_n \rho c = -\nabla_n p / jk$ and are described by the functions $E(i)$ and $C(i)$ (discussed below.)

The major results of the simulation are shown in Figure 4.5 for 40, 80, 160, and 320 Hz and $I = 16$ channels with an array radius equal to 3m. $k_m = (F_s \Omega_m)/c$ is the spatial frequency corresponding to discrete temporal frequency Ω_m. The top row is the reference field using the pressure p_i and the normal pressure gradient $\nabla_i p_i$ "measured" at I locations in the simulated diffuse field:

$$E(i) = \hat{p}_i + \frac{-\nabla_i \hat{p}_i}{jk_m} \tag{4.21}$$

The non-zero reconstruction error is due to the limited number of radiators, the spherical vs. cylindrical radiation assumption, and by errors introduced by the single-layer approximation of the Kirchhoff Helmholtz Integral. In Figure 4.1, it can be seen that reconstruction is quite similar to the reference diffuse field. For the second and third rows, the multipole observations at the origin are transposed to the I locations of the virtual array (Section 4.4.1.) The pressure $\hat{p} = W$ and the x and y pressure gradients $(-\nabla_{x,y}\hat{p})$ / $jk = X, Y$ are observed at the origin. For every channel i, an outward facing cardiod $C(i)$ is derived from a linear combination of X and Y based upon the angle ϕ_i of vector $\angle \mathbf{r}_i$:

$$C(i) = W + \cos(\phi_i)X + \sin(\phi_i)Y \tag{4.22}$$

In the second row, nothing further is done and the loudspeaker driving signal is $E(i) = C(i)$. In the third row the decorrelation filters are applied $E(i) = D_i(m)C(i)$. The equation used to obtain the simulated results is:

$$\hat{p}(\mathbf{r}) = -\frac{jk_m}{2}\sqrt{\frac{-2\pi}{jk_m}} \quad \sum_{i=1}^{I} E(i) \left[\; G(\mathbf{r}_i) + \frac{\nabla G(\mathbf{r}_i) \cdot \mathbf{n}_i}{jk_m} \; \right] \Delta s \tag{4.23}$$

The constants outside the integral are not reduced such that their justifications are evident.

Figure 4.6 shows the spatial autocorrelation computed for concentric rings $|\mathbf{r}_0|$ around the origin by considering the origin reference pressure $P(0)$ and the field pressure $P(|\mathbf{r}_0|, \phi_i)$. The quantity $\beta_m(|\mathbf{r}_0|)$ is defined as a function of distance:

$$\beta_m(|\mathbf{r}_0|) = \frac{\sum_I P(0)P^*(|\mathbf{r}_0|, \phi_i)}{\sqrt{\sum_I |P(0)|^2 \sum_I |P(|\mathbf{r}_0|, \phi_i)|^2}} \tag{4.24}$$

following the definition of normalized cross-correlation [30] and using the origin reference point $|\mathbf{r}_0| = 0$ and all angles $\angle\mathbf{r}_0$ as described in Section 4.3.2. It is desired that $\beta_m(|\mathbf{r}_0|)$ resemble the theoretical spatial autocorrelation shown in Equation A.53, discussed immediately below.

4.5.2 Results

Figure 4.5 shows simulation results for frequencies corresponding approximately to the lowest three octaves of music. Figure 4.6 shows the spatial autocorrelation described by the function $\beta_m(|\mathbf{r}_0|)$ defined in Equation 4.24 averaged over 16 distinct realizations using different MATLAB random number generator (RNG) seeds. This is necessary because any given sample from a random number generator is no longer random, and, for a given frequency, some RNG seeds lead to slightly lower or higher $\beta_m(|\mathbf{r}_0|)$ values. When averaged, the $\beta_m(|\mathbf{r}_0|)$ values are very close to the reference. Lower frequencies were chosen because they are easier to interpret visually, and because low frequency decorrelation requires the largest microphone spacing and can be a logistical issue for sound recording engineers.

The top row of Figure 4.5 shows results for a double-layer K/H reconstruction driven by measured values from the simulated diffuse field at those locations. There is some error due to the modest number of channels, however this becomes minutial as the channel count is increased. In Figure 4.6 this reconstruction is the reference spatial autocorrelation.

The second row of Figure 4.5 shows the results for the center observations translated to the edge of the virtual array without the decorrelation filters applied. This case is analogous to decoding too many Ambisonic channels from a limited series of SBF. It is visually obvious that the modeled diffuse field is too patterned. In Figure 4.6 this is the dotted line with very high spatial autocorrelation values.

The third row of Figure 4.5 shows the results for the center observations translated to the edge of the virtual array with the DFM decorrelation filters applied. The modeled fields are visually similar to the reference K/H fields. The particulars of the patterning are entirely different. In Figure 4.6 this is the dashed line and is very close to the reference K/H

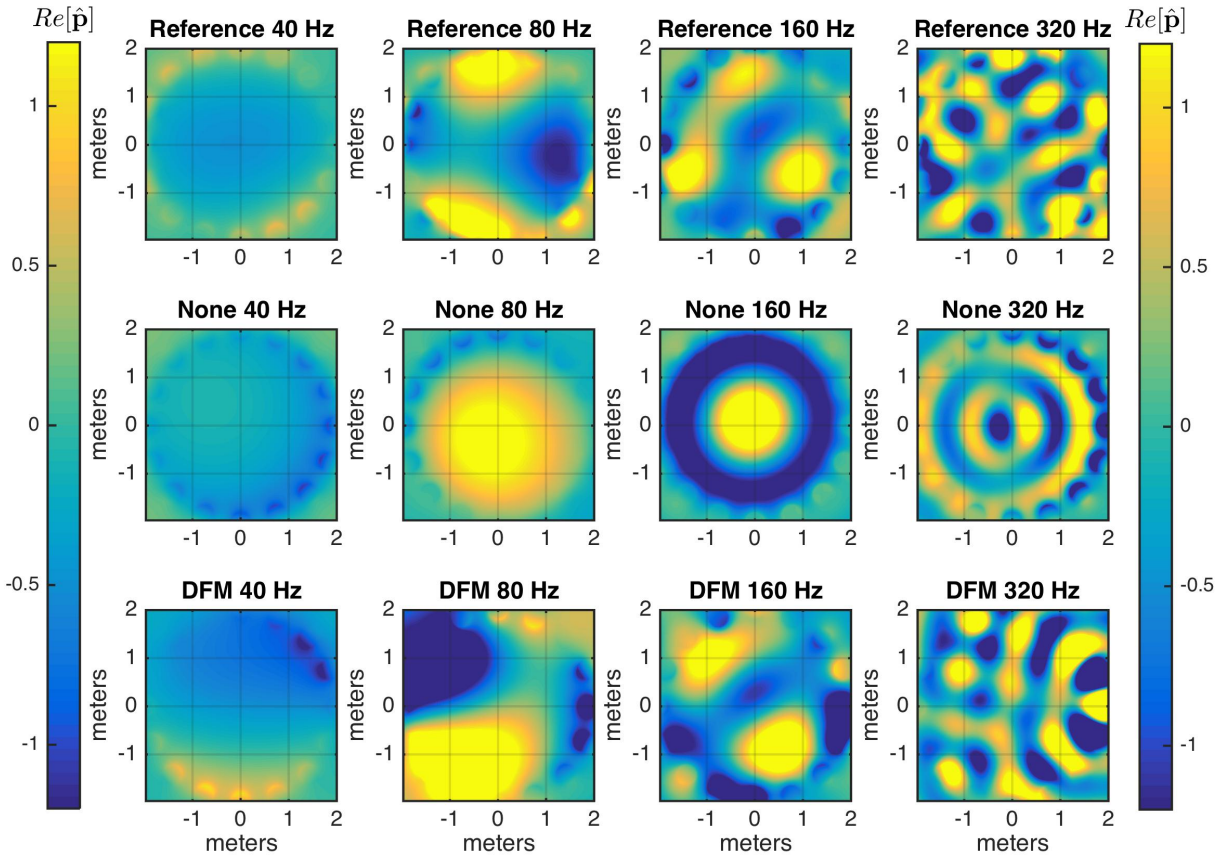

Figure 4.5: Simulation results for 40, 80, 160, and 320 Hz for 16 channels with an array radius equal to 3m. The top row is the cardioid to cardioid array using measured values of pressure and pressure gradient from spatial distinct points in the field. The second row uses transposed observations made at the origin without decorrelation, and third row is with DFM decorrelation.

reconstruction.

Figure 4.7 shows simulation results for the absolute value of pressure from a 64 channel array. The top row is the cardioid to cardioid array using measured values of pressure and pressure gradient from spatially distinct points in the field. The second row uses transposed observations and the DFM decorrelation strategy, and third row uses broadband decorrelation. It can be seen that there is substantial cancellation in the broadband decorrelation case and that erratic "hotspots" of diffuse energy occur when adjacent loudspeakers do not cancel.

If the random number generator is specified for the filter generation, the results of the

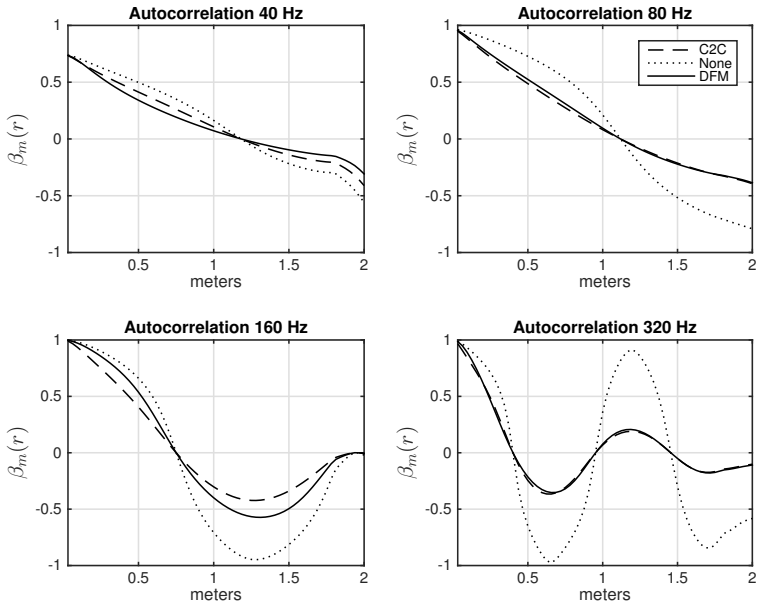

Figure 4.6: The function $\beta_m(|\mathbf{r}_0|)$ describes the spatial autocorrelation measured in the reconstructed pressure fields. The dashed line is the reference using values of pressure and pressure gradient measured from the diffuse field. The dotted line is the transposed virtual microphone array without decorrelation filters, and solid line is the DFM algorithm. Because any given sample from a random number generator is no longer random, some RNG seeds lead to slightly lower or higher $\beta_m(|\mathbf{r}_0|)$ values. DFM converges to the reference when averaged over 16 different random fields.

simulation will still vary depending upon the particular diffuse field that is simulated. For the reference case this is of little importance, but it points out a short coming of DFM that may lead to coloration - the p and $\nabla_{x,y,z}p$ values observed at the origin will influence the modeled edge values. For a given frequency, the introduced variation is centered around these values. And identical phenomenon will happen acoustically - a B-Format microphone used to recording a RIR will have spectral variation depending upon the location it was placed in the recording space. Even in the case that the filters introduce variation which has a mean of 0 dB, the coloration at the observation point will influence the entire modeled diffuse field. This strongly suggests that the RIR should be equalized to have smooth magnitude response and is left as future work.

The virtual array of outward-facing cardioid microphones used in DFM is able to discriminate between directional energy in the diffuse RIR. The angular differences in directional energy are limited by the angular resolution of the cardioid microphone. In development, a non-ideal diffuse field was simulated by placing a linear multiplier on any of the Q plane

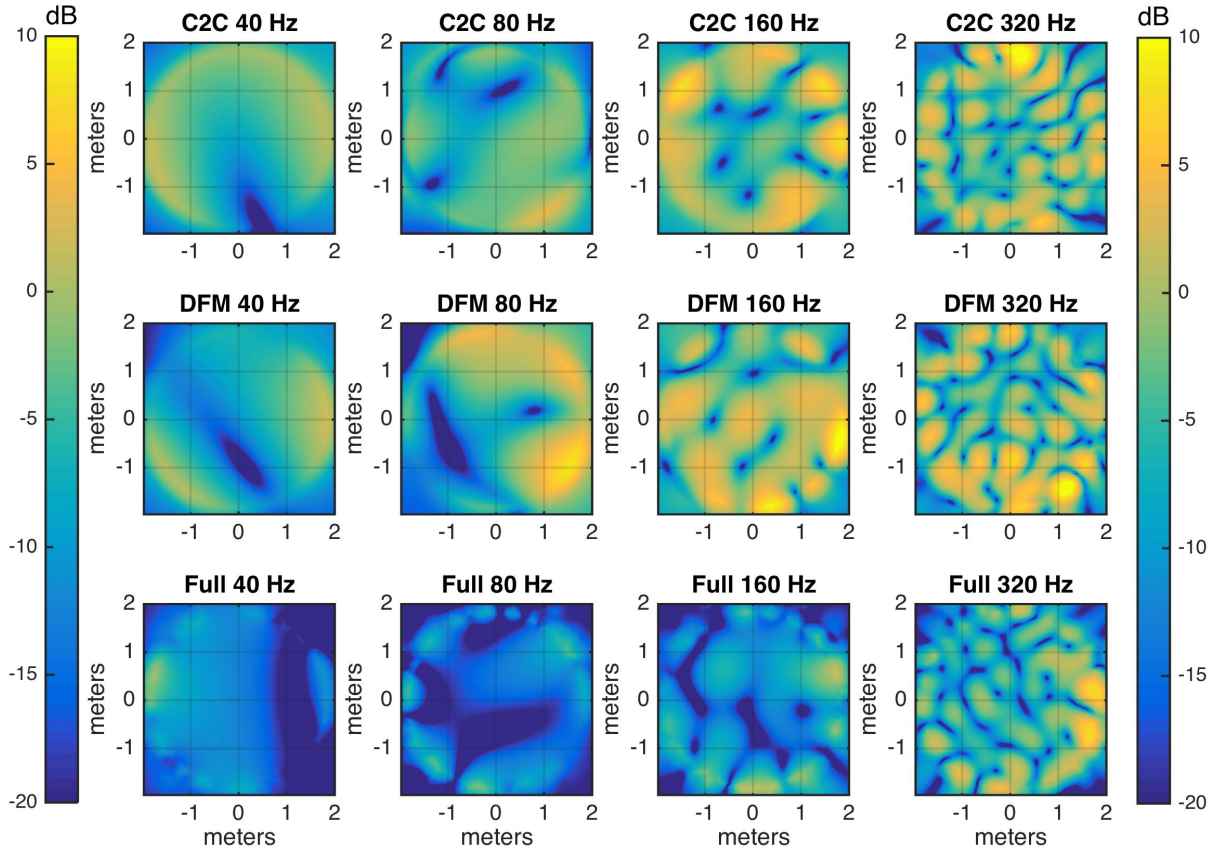

Figure 4.7: Simulation results for 40, 80, 160, and 320 Hz for 64 channels with an array radius equal to 2m. The absolute value of the pressure field is plotted in this example. The top row is the cardioid to cardioid array using measured values of pressure and pressure gradient from spatial distinct points in the field. The second row uses transposed observations and the DFM decorrelation strategy, and third row uses broadband decorrelation. It can be seen that there is substantial cancellation in the broadband decorrelation case and that erratic "hotspots" of diffuse energy occur when adjacent loudspeakers do not cancel.

waves oriented in the positive x-direction and the resulting driving functions clearly show directional biases in the loudspeaker driving signals of the array. The decorrelation filters introduce variation into this directional bias, however the mean of $D_i(m)$ tends toward unity over both i and m. That is, when considering a large "area" of space and frequency, the greater directional energy is preserved to a low angular resolution.

4.6 Conclusions and Future Work

Diffuse Field Modeling (DFM) is a systematic method for the reproduction of diffuse sound fields using decorrelation filters based on the statistical description of reverberation. This paper described the design of decorrelation filters based on physical acoustics and validated the reconstructed diffuse fields through a numerical simulation based on the Kirchhoff / Helmholtz Integral. It was demonstrated that the resulting fields have the expected spatial autocorrelation and that the filters can be tuned to introduce random variation that has physically-plausible frequency autocorrelation. The latter heavily influences the spatial impression.

The companion paper [65] presents a perceptual evaluation of DFM as a component of a physically-plausible virtual acoustic model that can be systematically adapted to various loudspeaker arrays. The findings indicate that it is necessary to model reflections in conjunction with DFM, and the majority of sound recording professionals found the initial 20-loudspeaker implementation to be useable in practice. A second experiment used the 5.1 loudspeaker configuration and found no significant differences with respect to the Hamasaki Square. Finally, DFM has been used in an experiment on auditory motion perception using both wave field synthesis and vector-base panning in a 48-loudspeaker setup [13] without objectionable artifacts.

DFM allows the creation of arbitrary observation points on a virtualized array in the recording space from a 4-channel, B-Format microphone array. The channels of the microphone array must be treated such that they can be assumed to be diffuse. Sound engineers typically used differing strategies for the direct and room sound. DFM allows greater logistical flexibility for the diffuse sound, however it must be used in parallel with a point source techniques the direct sound and specular reflections. This appears to be a good fit with the existing sound recording workflow, and opens the possibility of using very concise spatialization techniques for the early response. This is beneficial from the view of both physical approximation and aesthetic possibilities.

Constraints upon the random number generator used to create the filters are a clear next step. It is desired that similar randomness be introduced, however it is also desired that all frequencies have a mean magnitude of 0 dB for the exact number of loudspeakers at hand. Non-regular spacing of the bandpass center frequencies will ensure that the band edges of the filters interact minimally in the frequency domain. It was pointed out in Section 4.5.2 that the spectral coloration of the diffuse field at the point of the microphone will influence the entire DFM acoustic display, and strongly suggests an equalization scheme to minimize

this controllable impact to sound quality. Finally, the ability to automatically generate decorrelation filters from an impulse response based upon the frequency autocorrelation would be a substantial economization of workflow compared to the manual tuning currently required. This would yield physically-plausible values of frequency autocorrelation such that it easy for a sound recording engineer to start with a very good approximation of acoustic space.

Segue

The results of Chapter 2 suggest that the Hamasaki Square is an excellent benchmark for other techniques; the physical and perceptual results of Chapter 3 suggest that directional energetic differences in reverberation are a component of complex acoustic spaces and should be appropriately modeled. Both were taken into account in Chapter 4 which presented the development of a systematic tool for the recording and reproduction of diffuse sound fields. Diffuse Field Modeling (DFM) uses decorrelation filters based on the statistical description of reverberation to "virtualize" an array of outward-facing cardioid microphones from linear combinations of a B-Format microphone. It described the design of the physically-inspired decorrelation filters, validated the reconstructed diffuse fields through a numerical simulation, demonstrated that these fields have the expected spatial autocorrelation and that the channels of the array have the expected frequency-dependent correlation, and finally established a correspondence between the introduced frequency autocorrelation and the RT60 of the recorded diffuse field.

Similar to the Hamasaki Square, DFM maintains the "direct" paradigm of a (virtual) microphone array in a listening space directly routed to a loudspeaker array. A limitation of the Hamasaki Square is that it cannot represent three dimensional fields. Due to the microphone pattern it cannot discern between fore and aft differences in diffuse energy and will awkwardly distort lateral differences due to the negative lobe of the bi-directional microphones. In light of physical acoustics (Appendix A), the microphones on the bounding surface must be able to discern between inward and outward-bound acoustic waves to approximate the Kirchhoff Helmholtz Integral. It is the choice of virtual cardioid microphones formed from a B-Format Room Impulse Response that allows DFM to discern inward-bound acoustic waves, to approximate the Kirchhoff Helmholtz Integral, and to generate channels for three-dimensional loudspeaker configurations.

Chapter 5 presents a perceptual evaluation of DFM as a component of a physically-plausible virtual acoustic model that can be systematically adapted to various loudspeaker

arrays. A first experiment will use a 20-loudspeaker array with a 16-channel lower ring and a 4-channel height ring. The objective is to assess the necessary computational complexity of DFM and to ensure that there are no major issues in practice. A second experiment will use the 5.1 loudspeaker configuration and evaluate DFM against the Hamasaki Square [29][73].

Chapter 5

Diffuse Field Modeling using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part II Evaluation

Diffuse Field Modeling (DFM) is a systematic means for simulating and reproducing a diffuse field for arbitrary loudspeaker configurations. DFM is presented in two publications: Part I [60] presents the algorithm and this Part II reports the perceptual evaluation. Two experiments were conducted: in Experiment 1, sound recording professionals were to rate different treatments of DFM presented on a 20-channel array. The treatments under evaluation included the geometric modeling of reflections, strategies involving the early portion of the B-Format Room Impulse Response, and a comparison between 0^{th} and 1^{st}-order RIR. Results indicate that it is necessary to model the earliest reflections and to use all four channels of the B-Format room impulse response. In Experiment 2, musicians and sound recording professionals were asked to rate DFM and common microphone techniques presented on 5.1 (3/2 stereo) setup. DFM was found to be perceptually comparable with the Hamasaki Square technique.

5.1 Introduction

The logistical and economic constraints of high-quality loudspeakers and microphones necessitate that spatial sound reproduction is always an approximation. Additionally, a number of

different approaches can be used for the recording and reproduction of sound scenes. Array-based recordings attempt to capture the entire scene using a number of specially-configured microphones. The microphone channels may be routed directly to specific loudspeakers or combinations of the microphones may be used. Alternatively, spot microphones may be placed in close proximity to the instruments or the source may be synthesized electronically. In this case, spatial impression is created using panning, simulated propagation delay, and room effect. The reproduction of sound scenes can be done in a stereophonic format such as 2-channel stereo or 5.1 (3/2 stereo), using vector-base amplitude panning (VBAP) for larger but sparse loudspeaker arrays, or using wavefront reconstruction techniques for dense arrays.

The approach used in this study is a physically-plausible virtual acoustic model that can be viewed as a hybrid approach combining array-based and spot-based approaches in the sense that isolated sources are presented in an acoustic approximation of the scene. The model-based approach allows manipulation of the scene crucial to the presentation of music, while the acoustic approximation of the scene attempts to be similar to experiencing the original sound field. This model replaces the panning, delay, and reverberation that would be typically used and can be viewed as a precursor to a scene-based mixing paradigm where aesthetic decisions are made by moving objects in a virtual acoustic scene. Source and listener geometry are used for the direct sound and reflections. The reverberation was rendered using Diffuse Field Modeling (DFM), which is a systematic means for reproducing a diffuse field for arbitrary loudspeaker configurations. This is done using specially treated B-Format Room Impulse Response (RIR) and decorrelation filters that take into account the loudspeaker array geometry and the properties of the RIR. The topic of DFM spreads across two papers: formal technical presentation is given in Part I [60] and perceptual evaluation here in Part II.

Two experiments are presented. The first aims to identify the best treatment of the DFM algorithm, and the second compares the best treatment to standard microphone techniques in spatial sound reproduction. Experiment 1 uses a large, three-dimensional array of 20 loudspeakers. It is desired to know if a physically-plausible virtual acoustic model is able to produce a perceptually-plausible reproduction of a concert hall, and if expert sound recording engineers at McGill believe that DFM is a useable reverberation effect. Furthermore, perceptual effects related to the computational cost and complexity of DFM and the virtual acoustic model are investigated. Specifically, what happens when early reflections are modeled, omitted, or modestly distorted, when the perspective of reflections and late field are randomly altered, and when the room impulse response is reduced from 1st to 0th

order? Experiment 2 compares DFM, the Hamasaki Square (HS) and a virtual microphone technique with highly correlated loudspeaker signals. While DFM is more adaptable than HS, it is desired to know if it is comparable to a spatialized technique using unprocessed microphone signals in the standard 5.1 configuration.

5.2 Background

5.2.1 Recording, Rendering and Reproduction Strategies

A great deal of popular music is recorded using spot microphones. Spatial impression is created using amplitude panning, simulated propagation delay, and (typically) artificial reverberation. It would be incredibly tedious and time consuming for a sound recording engineer to accurately simulate an acoustic space using these tools, and, not surprisingly, this is rarely done. The advantages of using these tools and techniques include the ability to freely position sources, to change the level of each source individually, to change the properties of the artificial reverberation, and to individually manipulate the sources using audio effects such as equalization and dynamic range compression.

Concert hall music is frequently recorded with a main microphone array and individual sections of the ensemble are reinforced using zone microphones. There are a number of different arrays for stereophonic reproduction systems. Coincident microphone techniques such as XY and Mid / Side produce level differences between channels and have stable imaging [71, 16]. They are often criticized because the overlap in the pattern causes any recorded diffuse sound to be too correlated. The Blumlein technique [8] is similar to XY and M/S, but has more desirable reverberation properties because the bi-directional microphones are "open in the back" [25]. Near-coincident and spaced cardioid techniques such as ORTF, William's MMA, and Optimized Cardioid Triangle produce both time and level differences between the channels [79, 73]. Spaced techniques such as AB or Decca Tree [71] produce predominantly phase differences between the recorded channels and can have erratic imaging. They have very desirable spatial properties because they sample the diffuse field at a number of separated points.

An example of a non-stereophonic array for large and dense loudspeaker arrays is the "Acoustic Curtain" proposed in 1934 by Steinberg and Snow and also used in [63] and [64]. Spherical microphone arrays allow a full physical capture of a sound field from the perspective of a single point. Higher-order microphone arrays are able to capture greater variation in azimuth and elevation, and, if the order of the array is sufficiently high, sound scenes are

translatable to many different loudspeaker array configurations [50].

The advantages of array-based approaches is that they capture an entire acoustic scene including the placement of the instruments, the radiation properties and extent of the instruments, as well as the early reflections and reverberation of the recording space. The disadvantage of an array-based technique is that this capture is fixed and allows little manipulation at the mixing console. The Mid/Side technique is an exception to this. Zone microphone reinforcement allows level reinforcement of the musical parts, and time alignment of the zone microphones with respect to the main array will largely preserve the spatial impression.

It is almost always the case that a separate strategy is used to capture the sound of the reverberation, and it is desirable to have independent control over this aspect. A microphone technique known as the Hamasaki Square records and reproduces both the spatial and temporal aspects of the reverberation by placing four bi-directional microphones in a square pattern with the nulls oriented to the direct sound and the forward direction oriented to the lateral walls of the recording space. These four microphones are directly routed to the left, right, left surround, and right surround of the 5.1 format (3/2 stereo). It allows independent manipulation of the direct sound, the frequency-dependent channel correlation can be adjusted by microphone spacing, and it provides a reasonable facsimile of lateral reflections [29]. However, the Hamasaki Square - and all other channel-based techniques - are limited to the format they were designed for, in this case 5-channel reproduction. Adaptions can be made, but they are not systematic and potentially distort the acoustic properties of the recording such as the early reflections or late field correlation.

From a physical perspective, the Hamasaki Square can be interpreted as a sparse sampling of a bounding surface directly routed to secondary radiators in the listening space. The aim of DFM was to generalize the Hamasaki Square to arbitrary loudspeaker formats by forming a virtual outward-facing microphone array conceptually coincident with the available loudspeaker array. By specially treating the RIR and assuming the audio to be diffuse, physically-inspired decorrelation filters can be designed that simulate the random variation we would expect to see on this virtual array.

5.2.2 Psychoacoustics and the Perceptual Attributes of Spatial Audio

In room acoustics, early reflections are the first cues about the spatial properties of the sound field. The timing and direction of arrival of early reflections are important psychoacoustic

cues related to auditory source width (ASW) [11] and the accurate perception of distance [48]. The first few early reflections are the most important in forming the spatial impression, and the effect of an individual reflection reduces gradually until only a collective effect, the diffuse field, is perceived [3]. Physically, the reverberant diffuse field is incoherent energy incident from all directions [53], and is perceptually related to an auditory event heard "everywhere" [6] and a listener's sense of being enveloped in an acoustic space [11]. Reverberant energy facilitates sound source localization, realism, and externalization [69], but it can also impair the source width discrimination [11]. The frequency-dependent exponential decay of the reverberant diffuse field is related to the perceptual dimensions of reverberance and timbre [35], and the statistical correlation of the pressure field at the listener's ears, measured by the inter-aural cross-correlation (IACC), is related to the perceived spaciousness of the auditory image and the plausibility of the reproduction [6, 42]. Low IACC and high lateral energy are needed for wide ASW in concert halls, which, in turn, is directly related to preference in concert hall acoustics [51]. Low IACC is also found to be the best predictor for listener envelopment [24].

The perceptual measurement of sound quality is a multifaceted problem involving a number of attributes, and "it is accepted that it is possible to identify and elicit these attributes [4]." To some extent, spatial audio attributes (width, envelopment, spaciousness, distance) can be predicted by objective metrics, such as IACC and lateral fraction, but timbral attributes cannot [14], giving rise to the need for subjective testing when evaluating sound quality. Perceptual attributes of spatial sound quality have been addressed by many authors [5, 27, 40], for a review see [2]. In general, the elicited sound quality attributes can be divided under broader topics of spatial quality, timbral quality and annoyance [38]. Our test procedures were motivated by the scene-based paradigm for spatial audio evaluation proposed by [66]. The scene-based paradigm separates descriptions of sources, groups of sources, environments, and global scene descriptions in order to clarify the complex concepts under evaluation. Moreover, the selection of test participants also has an effect on the outcome: professional listeners tend to focus on the quality of the frontal image, whereas untrained listeners are biased by the wow-factor of the surrounding sound [67].

5.2.3 Implementation of Virtual Acoustic Model using Diffuse Field Modeling

The Diffuse Field Modeling algorithm is presented in detail in the separate technical publication (Part I) [60]. Description of DFM as a part of a virtual acoustic model and the

implementation details are given here.

The geometric perspective and temporal alignment of the direct path, reflections, and late field were carefully preserved. The direction and time of arrival of the reflections was computed using a scene-based image-source model based on measurements of Tanna Schulich and Redpath Halls at McGill University. The positions of the sources were chosen to be plausible arrangements of performers on the stages of these halls. The level of the reflections was computed by assuming that the sources obeyed the spherical spreading law and the direct path had unity gain. For instance, if a reflection's propagation distance was twice that of the direct path, the amplitude of the reflection was halved. A modest amount of further gain reduction (1-2 dB) was allowed to account for acoustic scattering, discussed below. The direct path originally present in the room impulse response (RIR) was used for time alignment of the diffuse energy and the recording orientation of the RIR was maintained such that any energetic differences in the diffuse field would be preserved. The B-format RIRs were captured at distances of 6 m and 9 m from the center of stage in the Tanna Schulich and Redpath halls, respectively.

The direct path and early reflections were rendered within Max/MSP by playing the source audio file. The direct path was positioned on the horizontal plane using a vector-base amplitude panning patch [58]. The reflections were delayed and lowpass filtered at $\approx$ 13kHz. The number of reflections that need to be modeled for accurate distance perception has been found to be as low as the three first reflections [12]. Here, the first five reflections were modeled: two wall reflections, two stage wall reflections and a ceiling reflection. In the first experiment the reflections were "hard panned" to specific loudspeakers. In the second experiment, the lateral reflections were displayed in the left and right loudspeakers due to the fact that summing localization is highly unstable in the lateral segments of a 5.1 display [73]. The ceiling reflection was not modeled. The timing and amplitude of reflections was computed using a virtual cardioid microphone conceptually-coincident with the loudspeaker in an effort to be consistent with the assumptions of the DFM algorithm.

The channels of diffuse energy were generated in MATLAB and played back as a multi-channel audio file within Max/MSP along with the direct path and reflections. The assumption of diffuseness is valid after the mixing time of a room [53, 7]. To achieve this constraint, the direct path and any strong specular reflections were removed using a MATLAB script. This can be achieved in a number of ways, in this paper two approaches were used. In the first, the direct sound was removed and nothing further was done (Treatments named *Clipped, Clipped scrambled & Clipped 0th*). In the second, the diffuse field was "faded in"

with a 21 ms logarithmic gain ramp (-12 to 0 dB) using the physical / perceptual metric Normalized Echo Density [32] (Treatments named *Reflections off & Reflections modeled*).

The RIR are then convolved with source audio to form a four-channel stream, and linear combinations of these streams are used to form outward-facing virtual cardioids coincident with the inward-facing loudspeakers. Each virtual cardioid is subject to a single, distinct decorrelation filter generated in MATLAB using the array geometry and RT60 of the RIR. The fact that the direct sound and any desired early reflections are treated separately is particularly powerful as it allows multiple sources to be individually spatialized with precise geometric information while a single diffuse layer is used for all sources. Additionally, the "fade in" of the diffuse RIR is overlapped in time with geometrically modeled early reflections and forms a coarse approximation of acoustic scattering for the early response (Treatment: *Reflections modeled*). This temporal spreading of energy is the reason that an additional 1-2 dB of attenuation was allowed in the reflection tuning.

The treatments with direct sound clipped include three different versions for the room effect. The first, *Clipped*, removes the direct sound and leaves the early reflections and late field unchanged in the RIR. In the second, the perspective of reflections and late field are randomly altered by changing the directions of the virtual array (Treatment: *Clipped scrambled*). This treatment explores the effect of maintaining the overall perspective in the room effect reproduction. In the third, the order of the RIR is reduced from 1st to 0th order (Treatment: *Clipped 0th*) to explore the need for the added computational complexity of using 1st order RIR.

5.3 Comparative evaluation of different DFM treatments using a 20-channel array

5.3.1 Participants

A total of 16 people participated in the first experiment of our study. One of the participants was female and the average age was 34.4 years (SD=12.7). The participants were sound recording professionals or sound recording PhD students in the McGill sound recording program, who can be viewed as experts in evaluating sound reproduction quality. The participants were paid $20 CAD for their contribution. A written informed consent was obtained from the participants prior to the experiment and the procedure was approved by the McGill Ethics Review Board (See Chapter 1.4).

5.3.2 Apparatus

The perceptual evaluation took place in the hemi-anechoic Spatial Audio Lab of the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) at McGill University in Montreal, Canada. The loudspeaker array consisted of sixteen Genelec 8030 loudspeakers equally-spaced on the horizontal plane in a circle with diameter of 3.7 m at ear-level. Four additional equally-spaced loudspeakers were used for height at elevation of 2.5 m, and a Genelec 7070A subwoofer was placed immediately below the forward-most horizontal plane loudspeaker. Figure 5.1 shows a photograph of the loudspeaker array and listening position. The interface was written in Max/MSP version 6.1 and played back on a Macbook Pro using an RME Madiface interface.

Figure 5.1: Loudspeaker array and listener position in the first experiment. Only the larger loudspeakers in the upper circle (Genelec 8030) were used, the smaller loudspeakers are part of another experiment. In addition, four height channels were used, which are not shown in this photograph.

Excerpt	Author	Track
Cello	J.S. Bach	Allemande (Suite No. 3 in C)
Band	Broken Crank	Red to Blue

Table 5.1: Musical excerpts

Treatment	Description
Reflections modeled	Direct sound clipped off and the RIR faded in after the early reflections. Five first early reflections modeled based on room geometry after convolution with source material and the DFM processing.
Reflections off	Direct sound clipped off and the RIR faded in after the early reflections.
Clipped	Direct sound clipped off from the RIR. Early reflections left intact.
Clipped scrambled	Direct sound clipped off from the RIR. Orientation of the virtual cardioids scrambled.
Clipped 0th	Direct sound clipped off from the RIR. Only omni-channel used for deriving loudspeaker signals.

Table 5.2: Treatments and descriptions

5.3.3 Stimuli

The stimuli were constructed starting with B-format RIRs from Tanna Schulich and Redpath Halls at McGill University. The halls can be described as a medium-sized wooden hall and a large hall with distinctively different RT60 times: Tanna RT60 = 1.4 s and Redpath RT60 = 2.3 s. Two musical excerpts, a cello recording and a band recording, were chosen to get different excitations for the halls. The excerpts were trimmed to 20-s loops. Table 5.1 gives detailed information about the excerpts. The recordings were done by close microphone technique and they contain only little reverberation from the recording space. Five different versions of the DFM method were used, called here the treatments. The different treatments are summarized in Table 5.2.

5.3.4 Procedure

The participants were tested individually at the centre of the loudspeaker array in the hemi-anechoic room. They were instructed to imagine themselves sitting in a concert hall and listening to a live performance on stage. This experiment was divided into two sessions. In the first session, for each treatment, they were asked to "Rate the extent to which this sounds like a live performance in a medium-sized wooden / large hall". The rating was done

via a continuous slider with a scale from 1 to 100, with the end points labeled as "not at all" to "very much". In addition to the slider there was an open text field where the participants were asked to write reasons affecting their rating, *i.e.* why the treatment was or was not able to elicit the feeling of a live concert hall experience. This free-format was used to identify factors contributing to the experience of listeners as described in their own words.

Each session was divided into four trials corresponding to the two concert halls and two musical excerpts. Each trial contained five renderings of the same content corresponding to the five DFM treatments, and the user interface allowed the participants to switch seamlessly between the treatments and create shorter loops to facilitate comparison of them. Furthermore, the participants were given an option to listen to a 5-channel Hamasaki Square rendering of the excerpts in each trial. The goal was to offer a direct microphone based memory refresher, that was not intended to be perceived as a reference. The order of presentation was randomized within and across trials. Each of the treatments had to be rated and commented before the participants were allowed to proceed to next trial. In the second session, the participants were asked to rate the same treatments in terms of spatial quality and timbral quality via two sliders. These rating scales were chosen based on the classification of spatial sound quality attributes in previous research [38]. Here, again, an open text field was provided, but commenting was not mandatory. A questionnaire concerning demographic information, the participants' professional background, and whether they found some of the treatments applicable in the context of their own work was filled out after the test. Average duration of the test was 55 minutes.

5.3.5 Results

Quality of the treatments

Figure 5.2 shows the mean scores for the extent of realism with different treatments collapsing across participants, excerpts and halls. Eight outliers out of 320 data points were detected by the inter quartile range method, where the detection criterion is defined as $Q_1 - 1.5 \cdot (Q_3 - Q_1)$ or $Q_3 + 1.5 \cdot (Q_3 - Q_1)$, where Q3 is the third quartile and Q1 is the first quartile. The detected outliers were marked as missing values. A three-way ANOVA (5 treatments $\cdot$ 2 halls $\cdot$ 2 excerpts) revealed a significant main effect of the treatment and also a significant interaction between the treatment and the excerpt. No other effects were observed. Post-hoc tests were conducted using Tukey's HSD with a significance level of 0.05. The treatment *Reflections modeled* was rated as significantly better than all the other techniques except

Clipped for all 3 scales. *Clipped 0th* was significantly worse than any other treatment. The relevant ANOVA and post-hoc results are presented in Table 5.3. In addition, the mean

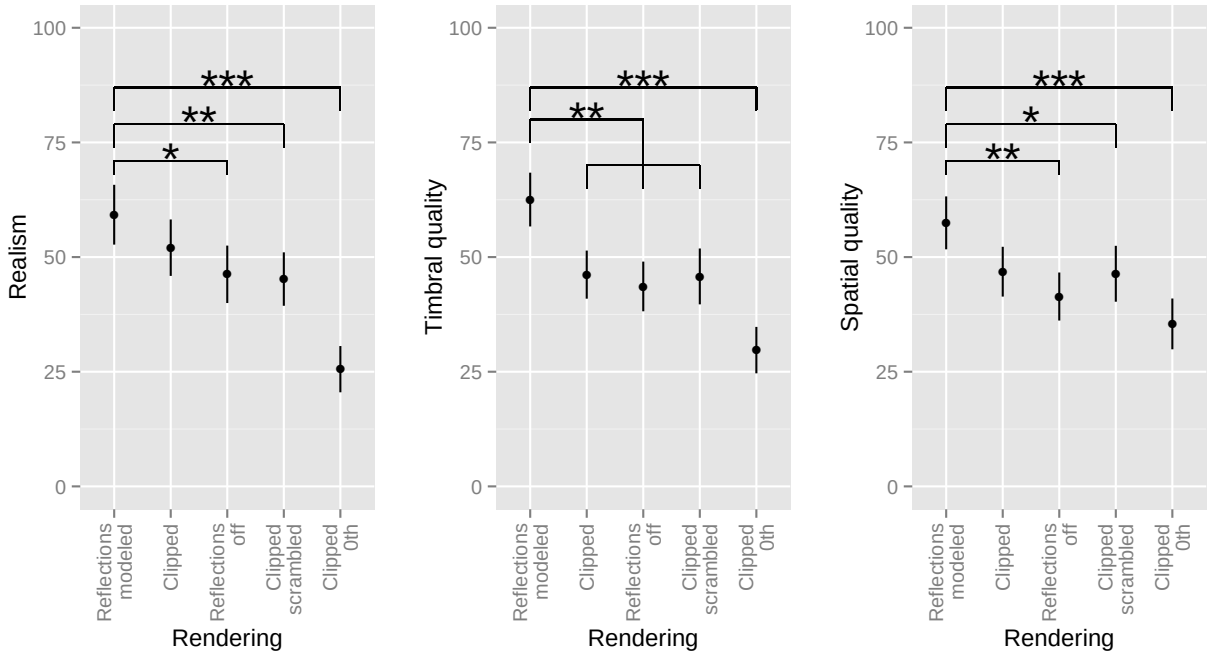

Figure 5.2: Mean scores for the extent of realism, timbral quality and spatial quality with the different treatments. Data from different halls and excerpts is collated in this analysis. The bars indicate the 95 % confidence intervals of the mean. The annotations: $***: p = 0, **: p < 0.01, *: p < 0.05$

scores for timbral and spatial quality are shown in Figure 5.2 and summarized in Table 5.3. Outlier detection found seven outliers for the timbre scale and three outliers for the spatial scale, which were removed from the data. A three-way ANOVA revealed a significant main effect of the treatment in the timbral quality and the spatial quality. No interactions were observed. A post-hoc Tukey's pairwise test revealed the *Reflections modeled* to be significantly better than the other treatments and the *Clipped 0th* to be significantly worse than the rest in timbral quality. In spatial quality the *Reflections modeled* was significantly better than *Reflections off*, *Clipped scrambled* and *Clipped 0th*, but not significantly better than *Clipped*, although there was a strong tendency.

Figure 5.3 shows the mean scores for extent of realism by excerpt. The largest difference in the means is in the *Reflections off* treatment, where the solo stimuli is rated higher than the ensemble. The visible tendency did not reach statistical significance (p=0.17). An opposite, but not significant, effect is observed in the *Clipped scrambled* treatment, where

Realism	
$Treatment$	$F_{(4,292)} = 18.115, p < 0.0001$
$Treatment \times Excerpt$	$F_{(4,292)} = 3.028, p = 0.018$
Post-hoc tests	Reflections modeled > Reflections off, Clipped scrambled, Clipped 0^{th}; Clipped > Clipped 0^{th}
Timbral quality	
$Treatment$	$F_{(4,293)} = 16.938, p < 0.0001$
Post-hoc tests	Reflections modeled > All other cases; Clipped 0^{th} < All other cases
Spatial quality	
$Treatment$	$F_{(4,297)} = 8.258, p < 0.0001$
Post-hoc tests	Reflections modeled > Reflections off, Clipped scrambled, Clipped 0^{th}

Table 5.3: Significant main effects, interaction effects, and post-hoc tests for extent of realism, timbral quality and spatial quality scales.

the ensemble receives higher rating than the solo. The ANOVA and post-hoc results are shown in Table 5.4.

Cello	
$Treatment$	$F_{(4,143)} = 7.748, p < 0.0001$
Post-hoc tests	Reflections modeled, Clipped, Reflections off > Clipped 0^{th}
Band	
$Treatment$	$F_{(4,149)} = 13.863, p < 0.0001$
Post-hoc tests	Reflections modeled > Reflections off, Clipped 0^{th}

Table 5.4: Significant main effects, interaction effects, and post-hoc tests for the extent of realism scale by excerpts.

Constituents of perceived quality

Open text field related to the treatments under evaluation was analyzed by searching the text data for often occurring expressions that can be grouped under a higher abstraction

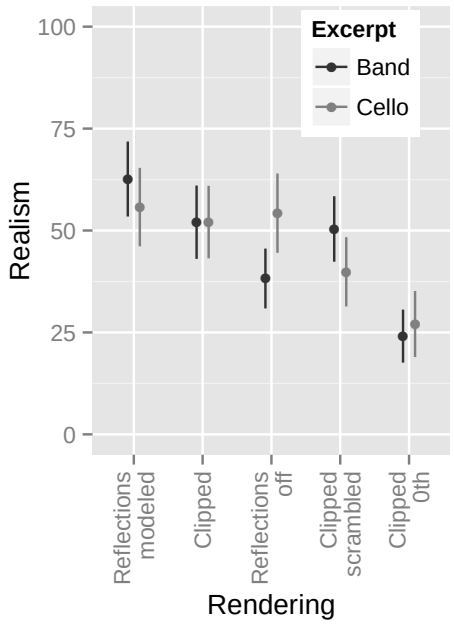

Figure 5.3: Mean scores for the extent of realism with the different treatments. Data from different halls is collated in this analysis. The bars indicate the 95 % confidence intervals of the mean.

level concept. A good example is the concept *Timbre problems* in our analysis, which holds expressions such as *"the cello sounds filtered and unnatural"*, *"the timbre is greatly affected"*, and *"certain frequencies are way too emphasized"*. The grouping process was repeated on a subset of the text data by another researcher and the two independent groupings were then compared and discussed by the researchers until a consensus was found. Table 5.5 presents the 19 most frequent concepts found in the text data.

A given concept is always connected to segments of the text data that are related to specific test conditions. To examine the correspondence of the concepts and treatments we performed Constrained Correspondence Analysis (CCA) on a contingency matrix formed based on the frequency of concepts in the test conditions. The counts in the contingency matrix are transformed into Chi-squared values based on the ratio of the expected and observed number of references to concepts in each test condition. Total inertia of the contingency matrix is computed by summing the Chi-squared values and dividing the sum by the number of observations. Total inertia of our contingency matrix is 6.549. We constrained the analysis to display only the variation caused by the 5 different treatments. In constrained analysis the axes are formed as linear combinations of the explanatory variables, which in our case

Concept	Count
Timbre problems	40
Realistic	29
Unrealistic	24
Unnatural reverberation	19
Too reverberant	16
Poor sense of space	16
Too distant	16
Poor early reflections	15
Unnatural	14
Artificial	14
Not enough reverberation	14
Smaller than expected	14
Natural	13
Phase problems	13
Being there	13
Sense of space	13
Poor balance direct to reverberant	11
Comb filtering	9
Good balance direct to reverberant	8

Table 5.5: Frequency of concepts

are the treatments. The constrained model explains 29 % of the full inertia in the contingency matrix, and the two first axes explain 36 % and 28 % of the constrained inertia. A permutation test confirmed the constrained solution to be statistically significant after 199 permutations ($F_{(4)} = 1.5120, p < 0.005$).

The two-dimensional solution is shown in Figure 5.4. Adding a third dimension did not provide any additional information because the treatments are already separated in the two first dimensions enabling us to interpret what the participants were paying attention to in the given treatment. The best treatment based on the extent of realism scale was *Reflections modeled* and the text data seems to confirm the result with concepts *realistic, natural, being there* and *sense of space* located closest to the cluster of test cases with that treatment. There is, however, one negative concept, *poor balance direct to reverberant*, blended together with the positive ones. The corresponding positive concept, *good balance direct to reverberant*, is located closer to the cluster of *Clipped* treatments implying that the choice of modeling the reflections or leaving them as they naturally are is affecting the integration of the source and space. However, the perceived poor balance does not seem to impair the overall experience

given the large number of positive concepts.

The worst treatment according to the extent of realism scale was *Clipped 0th*, which receives a large number of negative concepts: *unrealistic, comb filtering* and *timbre problems* among others. Moreover, the participants perceived the reproduced space to be confusing and smaller than they would expect. The treatment *Clipped scrambled* is most clearly associated with concepts concerning the reverberation: either the reverberation is unnatural or there is not enough reverberation. The treatment completely without early reflections, *Reflections off*, is located further away from the other four methods, implying that the concepts associated with it were mostly used only there, and it is described as *too distant* and *artificial*.

5.3.6 Discussion

The DFM algorithm allowed the generation of 20 channels of diffuse energy from a B-Format RIR without major coloration or imaging effects nor the unpleasant artifacts typically associated with excessive correlation. Modeling the early reflections along with the DFM algorithm was found the best treatment to reproduce a concert hall experience in a large loudspeaker array. The difference to the second best treatment, clipping the direct sound and leaving the early reflections subject to DFM decorellation filters, was not significant in the overall extent of realism scale, but a significant difference was found in the quality of timbre. This is expected as the early reflections are smeared in time due to the decorrelation process, which is likely to have an effect on the timbre of the stimulus. However, based on the data, the participants found the balance between the direct sound and reverberation better with the clipping method compared to modeling the early reflections. This finding may be due to the low number of modeled reflections used in our study, which may result in a perceivable divide between the stage and the room. In future implementations of the DFM algorithm the number of modeled reflections should be increased to avoid such effects. Ultimately, the additional computational effort of modeling the early reflections is justified based on our data with the *Reflections modeled* collecting 3 positive realism concepts and one positive space concept. The treatment *Clipped* is attributed with both positive and negative space-related concepts, and a negative timbre-related concept. The treatment without any early reflections, *Reflections off*, collects only negative concepts about both realism and space.

Another decision regarding computational load was if the 1^{st}-order microphone signals are necessary, or if there are perceptually similar results using only the 0^{th}-order omni-signal. Our results are clearly in favor of using the higher order microphone signals with the

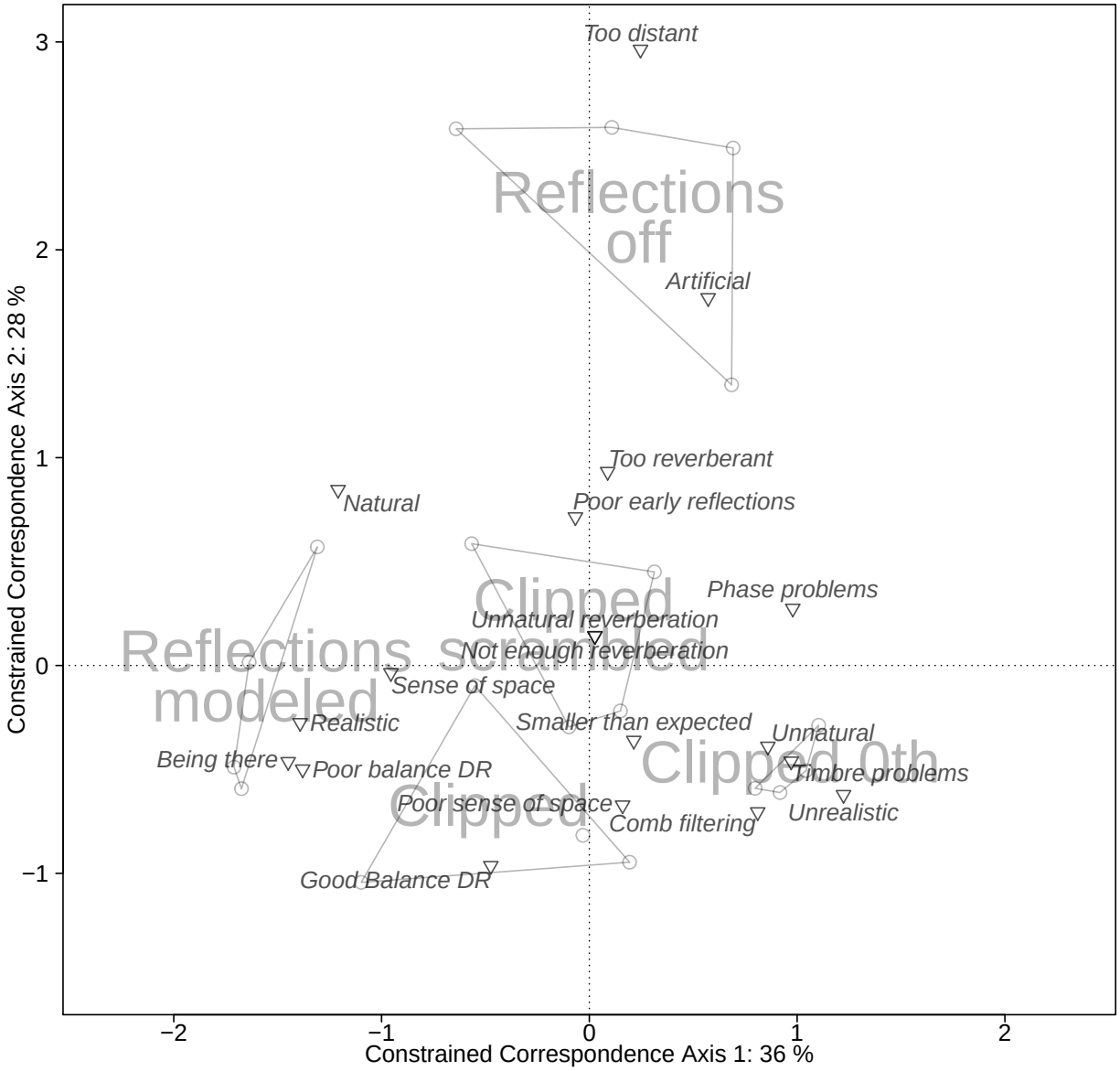

Figure 5.4: Correspondence analysis result when the ordination is constrained by the treatment. The four connected dots are the test cases (2 halls · 2 excerpts) in each treatment. The grey text denotes the treatment and the black text with triangles denotes the concepts. The map displays the concepts and treatments with more interconnections closer to each other.

Clipped 0th yielding lowest scores on every scale. A piece of evidence in favor of maintaining geometric perspective in the room effect, including early reflections and diffuse field, is the poor performance of the *Clipped scrambled*. If the perspective of the early reflections and late

field are randomly altered (scrambled) the overall quality and spatial quality are significantly worse compared to a treatment with an undistorted reproduction. The distorted case can be likened to a novice sound recording engineer being sloppy with the directionality of reflections and reverberation in his or her work. Reverberation is both temporal and spatial, and our results are consistent with evidence from concert hall acoustics [51].

The RIR used in this experiment had only mildly non-ideal behavior due to the absorptive seating; acoustic environments with far greater discrepancies are likely. This further substantiates the need to maintain geometric perspective of the B-Format RIR and to maintain the 1^{st}-order microphone signals. Because the outward-facing virtual cardioids are different at various points on the bounding surface, they are able to discern differences in inward-bound acoustic energy. This allows the representation of non-ideal diffuse fields where more energy is incident from certain directions and has been shown to be audible in [61].

Furthermore, the concepts elicited here indicate that the observed differences between the treatments were not due to aesthetic preferences alone. Rather, there were substantial perceptual issues caused by some of the treatments. The lowest scoring treatments were described with concepts *timbre problems*, *artificial*, and *comb filtering*, which, when used by professionals, cannot be considered as aesthetic choices.

Altogether the scores given for the extent of realism were quite modest, keeping in mind that the upper end of our scale was labeled as being in a concert hall. There are a number of possible reasons for the low score: the participants knew they were in a laboratory and they could see their surroundings which was contradicting the auditory impression; the participants were professionals who are used to critical listening and are not easily convinced; the task required them to listen to music in a new way, and not along the lines of current sound recording practice. Moreover, the finger noise from the cello was distracting from truly immersing into a concert hall experience. Those sounds are something that you would not hear when sitting further away from the stage. This feedback is taken into account when designing the second experiment of our study.

When asked whether the participants would use the best treatments in their own work, 11 out of 16 found the algorithm usable. Some found the best version realistic or hyper-realistic, suitable for classical music productions, and nice in general, but others thought the treatments were unrealistic and impaired by artifacts. Given that the DFM is a new method to reproduce the room effect there was a desire to compare it to existing technologies and to be able to tune it to some specific applications before making an assessment of its usability. However, the majority of the participants had a positive attitude towards what they heard

and were ready to accept DFM as a tool for future use.

As further validation, in [13], DFM was used as the reverberation effect for a 48-loudspeaker ring. The direct path and reflections were modeled with an interactive wave field synthesis system being used for perceptual evaluation of auditory trajectories. DFM was not the subject of this experiment, however the presence of the reverberation effect gave the impression of an auditory space and did not cause objectionable artifacts.

Experiment 1 aimed at characterizing various treatments of DFM. The best rated was *Reflections modeled*. Experiment 2 was designed to evaluate this DFM treatment in comparison to standard microphones techniques, namely the Hamasaki square and outward-facing cardioids, on a standard 5.1 system.

5.4 Comparative evaluation of DFM with common microphone techniques using 5.1

5.4.1 Participants

A total of 26 people participated in the second part of our study. 10 of the participants were female and the average age was 29.8 years (SD=9.8). 18 of the participants were music students from the McGill Schulich School of Music and eight of the participants were sound recording professionals or sound recording PhD students in the McGill sound recording program, who can be considered as professionals. Seven of the sound recording participants also took part in Experiment 1. The participants were paid $20 for their contribution. A written informed consent was obtained from the participants prior to the test and the test procedure was approved by the McGill Ethics Review Board.

5.4.2 Apparatus

Our second experiment took place in the Critical Listening Room at CIRMMT conforming to the ITU-R BS.775-1 standard. A standard 5.0 setup build with Bowers &Wilkins 802D loudspeakers was utilized in the test. The frequency response of the loudspeakers starts from 34 Hz, which is why a separate subwoofer was not considered necessary. An acoustically transparent curtain was mounted in front of the listening position to cover the frontal loudspeakers and the front section of the room in order to remove visual cues about the distance and extent of the true sound sources. Figure 5.5 shows a photograph of the setup. The

Figure 5.5: Test setup in the second experiment. The frontal loudspeakers were hidden by the acoustically transparent curtain. The lights in the room were dimmed so that the walls and ceiling behind the curtain were not visible from the observation position.

interface was written in Max/MSP 6.1 and the experiment was run on a Mac Pro computer.

5.4.3 Stimuli

The same two halls as in the previous experiment, a medium-sized and a large hall, were used here. The musical excerpts were an anechoic cello recording and two different band excerpts with close-microphone recordings of three instruments: drums, bass and a synthesizer. The excerpts were chosen based on feedback from the first experiment. There the finger noise and fast bowing in the cello excerpt were perceived, and this time the cello recording was chosen to minimize those effects. Similarly, the band excerpt was chosen to match the medium-sized hall and the large hall separately by selecting a slow section of the track for the large hall and fast section for the medium hall. Table 5.6 summarizes the excerpts.

Three methods were used: *DFM with reflections modeled*, *Hamasaki Square (HS)* and

Excerpt	Author	Track
Cello	G. Fauré	Sicilienne (Opus 78)
Band	Falcon Punch	Surface People (slow and fast sections)

Table 5.6: Musical excerpts

Outward-facing cardioids with modeled reflections (OFC). DFM with modeled reflections was chosen based on the first experiment due to its highest rating. HS is a popular microphone technique to reproduce a room effect and we were interested to see how DFM compares to it. HS is created by placing four bi-directional microphones in a square pattern with the nulls oriented to the direct sound and the forward direction oriented to the lateral walls of the recording space. These four microphones are directly routed to the left, right, left surround, and right surround of the 5.1 format. OFC was chosen due to expected poor performance to provide a low-end anchor in our test. Coincident microphones in a diffuse field have correlation which ranges from 1/3 for opposing cardioids, 2/3 for a 90 degree orientation, and 1 for the identical orientation [23]. OFC derives such microphones from B-Format, and, when these microphones are used to reproduce a diffuse field, the resulting spatial autocorrelation is far too high. This is shown in the simulations of the technical companion paper [60]. A professional sound engineer tuned each excerpt to match the hall and equalized obvious flaws in the frequency response of the hall. The sound engineer was allowed to adjust the relative levels of the direct sound, early reflections and reverberation, and to create an equalization filter for each of the 12 stimuli. The tuning and equalization were kept as subtle as possible (within 2 dB); the rationale was to provide stimuli that are not offending to the professionals participating in our test, while maintaining a physical approximation.

5.4.4 Procedure

The participants were tested individually in the Critical Listening Room. As in Experiment 1, they were instructed to imagine themselves sitting in the audience in a concert hall and listening to a live performance on stage. The test was divided into four trials according to the two halls (medium and large) and two musical excerpts (cello and band). In each trial an interface was shown where the three renderings of the same content could be compared side by side. Figure 5.6 shows a screen capture of one panel in one screen. There were always three panels like this side by side corresponding to the 3 methods. The order of presentation

was randomized within and across trials. The goal of the interface was two-fold: firstly to ease the participants to visualize the spatial composition of the sound environment, and secondly to get information about the individual sound sources in the scene. Similar idea of a drawing interface has been successfully applied by [74] in studying source and reverberance localization.

The task was to position the instruments in every panel as they were perceived. The instruments were instructed to be placed in relation to the listener and the curtain in front of them. Furthermore, the perceived width of each instrument could be controlled with a slider: if the instrument was perceived as point-like and precisely localized in space, the width should be zero. Any ambiguity in the localization should result in more perceived width. In addition, the participants rated three global scene parameters along semantic scales for the overall quality of the different methods (poor - excellent), the depth of stage (flat - deep), and the clarity of the listening environment (confusing - clear). Specifically, the depth of stage was concerned with how the frontal image of the rendering was perceived and the listening environment was concerned with the room impression. The overall quality slider encompassed any aspect the participants considered relevant. On average the test took 41 minutes to complete.

5.4.5 Results

Analysis of the ratings

Figure 5.7 shows the results for the overall quality, depth of stage, and sense of space scales. Outliers were detected as in the first experiment resulting in seven outliers from the overall quality scale, eight outliers from the depth of stage scale and one outlier from the sense of space scale being removed out of 312 data points in each scale. No difference was found between groups of participants (professionals vs. musicians), hence the results are presented collapsed over all participants. A three-way ANOVA (3 methods $\cdot$ 2 halls $\cdot$ 2 excerpts) revealed no significant main effect of the method in the overall quality, but a significant main effect of the hall was observed in addition to a significant three-way interaction of the method, hall and excerpt. A non-significant tendency was observed in the overall quality scale, where the DFM and HS methods are nearly rated higher than the OFC. In the depth of stage scale a significant main effect of the method was observed. Post-hoc testing revealed that both DFM and HS are rated higher than OFC. In addition, a significant three-way interaction of the method, hall and excerpt is observed. Regarding the sense of space scale,

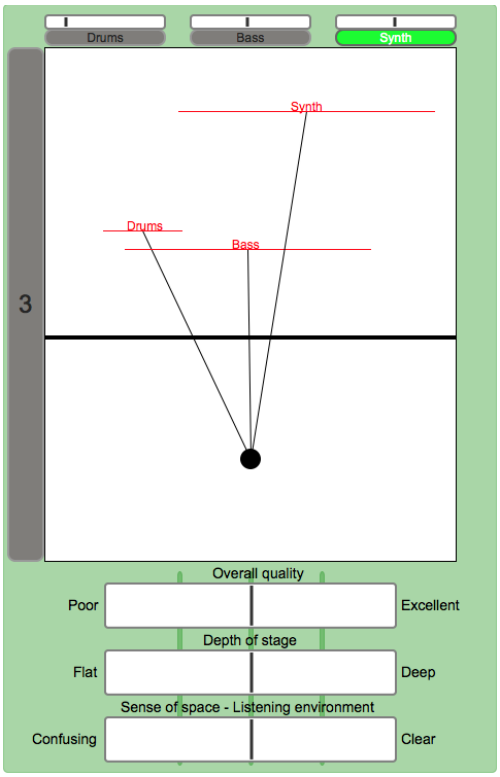

Figure 5.6: Graphical user interface in the second experiment consisted of three panels, one of which is shown above. The participant is depicted by the black circle and the horizontal line is the curtain in front of them. Each panel contained one rendering of the same stimulus content. The task was to place the instruments as they were perceived, assign a width to them, and rate the overall quality, depth of stage and sense of space elicited by the method.

no significant effect of the method was observed, but a significant main effect of the excerpt was found, where the cello was rated higher than the band. Table 5.7 summarizes the ANOVA and post-hoc results. The mean scores for depth of stage for each hall, excerpt and method are displayed in Figure 5.8. DFM and HS were significantly better than OFC with the band excerpt and the cello excerpt in the large hall. In the medium hall, with the band excerpt, the HS method is rated significantly better than OFC, but not significantly better than DFM. There is no significant difference between DFM and OFC. The cello excerpt in the medium hall, however, was rated significantly better with DFM than with either HS or OFC. Table 5.8 presents the ANOVA and post-hoc results.

Looking into the correlation of the depth of stage and listening environment scales with the overall quality reveals that the two scales have $R^2 = 0.46$ with overall quality. Further investigation of the relative importance of regressors reveals that the listening environment

Overall quality	
$Hall$	$F_{(1,293)} = 5.542, p = 0.019$
$Method \times Hall \times Excerpt$	$F_{(2,293)} = 6.555, p = 0.002$
Post-hoc tests	Large hall > Medium hall
Depth of stage	
$Method$	$F_{(2,292)} = 21.080, p < 0.0001$
$Hall$	$F_{(1,292)} = 10.045, p = 0.017$
$Method \times Hall$	$F_{(2,292)} = 3.081, p = 0.047$
$Method \times Excerpt$	$F_{(2,292)} = 5.948, p = 0.003$
$Method \times Hall \times Excerpt$	$F_{(2,292)} = 4.287, p = 0.015$
Post-hoc tests	DFM, HS > OFC; Large hall > Medium hall
Sense of space	
$Excerpt$	$F_{(1,299)} = 7.724, p = 0.006$
Post-hoc tests	Cello > Band

Table 5.7: Significant main effects, interaction effects, and post-hoc tests for overall quality, depth of stage and sense of space scales.

Cello, Medium hall	
$Method$	$F_{(2,75)} = 7.781, p = 0.001$
Post-hoc tests	DFM > HS, OFC
Cello, Large hall	
$Method$	$F_{(2,71)} = 12.3, p < 0.0001$
Post-hoc tests	DFM, HS > OFC
Band, Medium hall	
$Method$	$F_{(2,73)} = 5.992, p = 0.004$
Post-hoc tests	HS > OFC
Band, Large hall	
$Method$	$F_{(2,73)} = 9.14, p = 0.0002$
Post-hoc tests	DFM, HS > OFC

Table 5.8: Significant main effects and post-hoc tests for depth of stage scale by excerpt and hall.

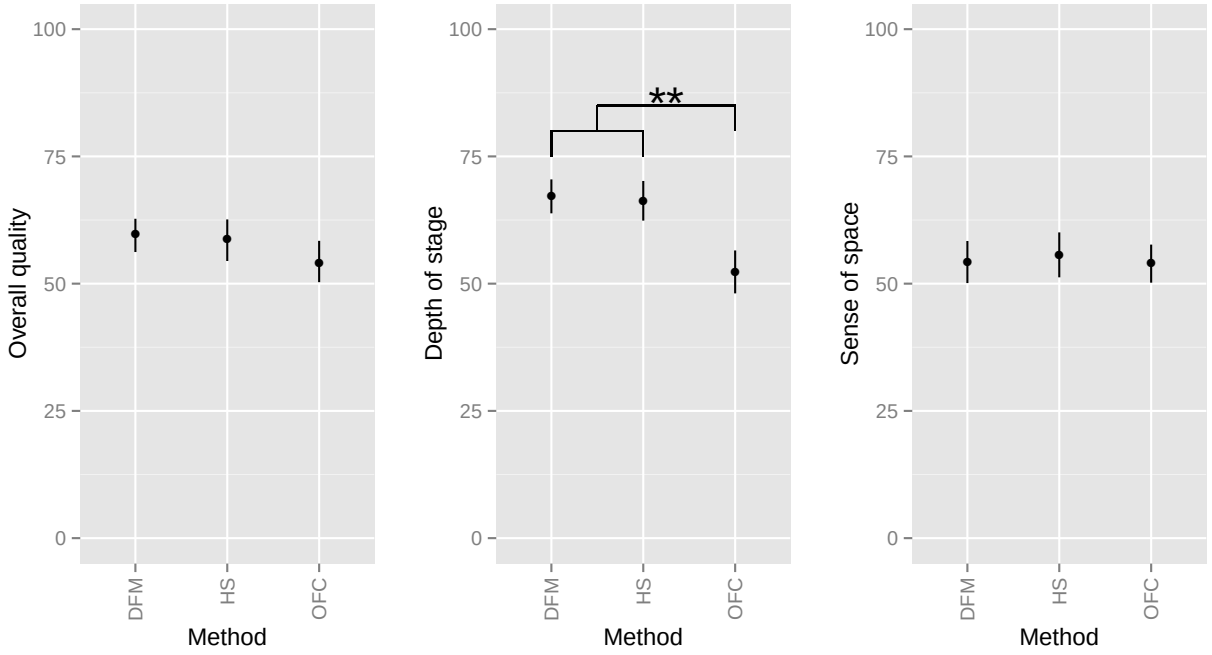

Figure 5.7: Mean scores for overall quality, depth of stage and sense of space. The data from the two halls and two excerpts is collated in this analysis. The bars represent the 95 % confidence intervals of the mean. The annotations: $***: p = 0, **: p < 0.01, *: p < 0.05$

clarity contributes 82 % and the depth of stage 17 % of the R^2 when averaged over orderings. There is no correlation between the depth of stage and listening environment ($R^2 = 0.04$) - clarity of the listening environment was rated highly for both flat and deep stages depending on the listener.

Visualizing the spatial composition

The drawing task data is visualized in Figures 5.9 and 5.10. A two-way ANOVA (method · hall) was conducted by excerpt to analyze the width and distance of the auditory event, and in the case of the band excerpt, the ensemble spread in distance. For the band excerpt the width was defined as the distance from the left-most instrument to the right-most instrument measured with taking the instrument widths into account, and the distance was the defined as the mean distance of the three instruments. A significant main effect of the hall was observed in the width of the band, where the band in the large hall was perceived wider than in the medium hall ($F_{(1,150)} = 10.63, p = 0.001$). Average distance of the band yielded a significant main effects of the method ($F_{(2,150)} = 6.48, p = 0.002$) and hall ($F_{(1,150)} = 10.53, p = 0.001$).

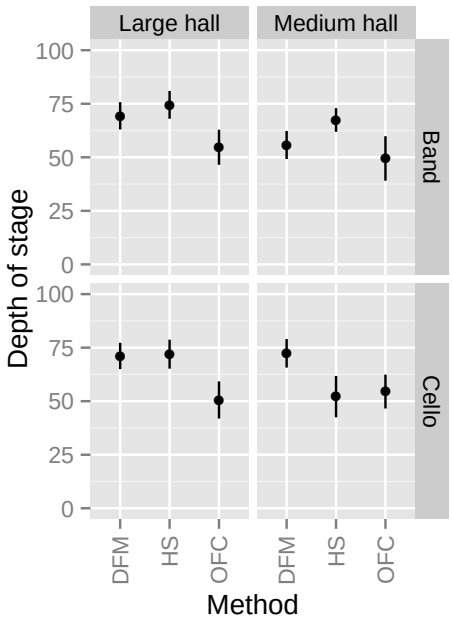

Figure 5.8: Mean scores for depth of stage in the two halls with two musical excerpts. The bars represent the 95 % confidence intervals of the mean.

Post-hoc testing revealed the HS being rated significantly more distant than the OFC, and the large hall resulting in more distant auditory events compared to the medium hall. No other effects were observed.

Despite the lack of significant results for the cello excerpt, inspecting the visualizations gives more insight about the numerical results presented in the previous section. In the figures the positions of the circles are the positions each participant assigned to the auditory event, and the size and transparency of the circles refer to the perceived width. Larger and more transparent circles imply more perceived width. The participant is positioned at [0, 0] pixels and the curtain is at distance of 100 px.

The cello in the large hall (Figure 5.9) appears to be mostly positioned close to the curtain with OFC. With HS there is a cluster near the curtain but also another cluster further away. DFM results in a more distant auditory event that is also quite evenly spread in distance across participants. The depth of stage scale (Figure 5.8) shows similar results. The medium hall gives similar results, with the exception of HS producing localizations spread to the right from the centre line in addition to the cluster closer to the curtain. This localization error may have resulted in the drop in the depth of stage score versus DFM.

The band visualization shows the positions of each instrument. In the large hall DFM

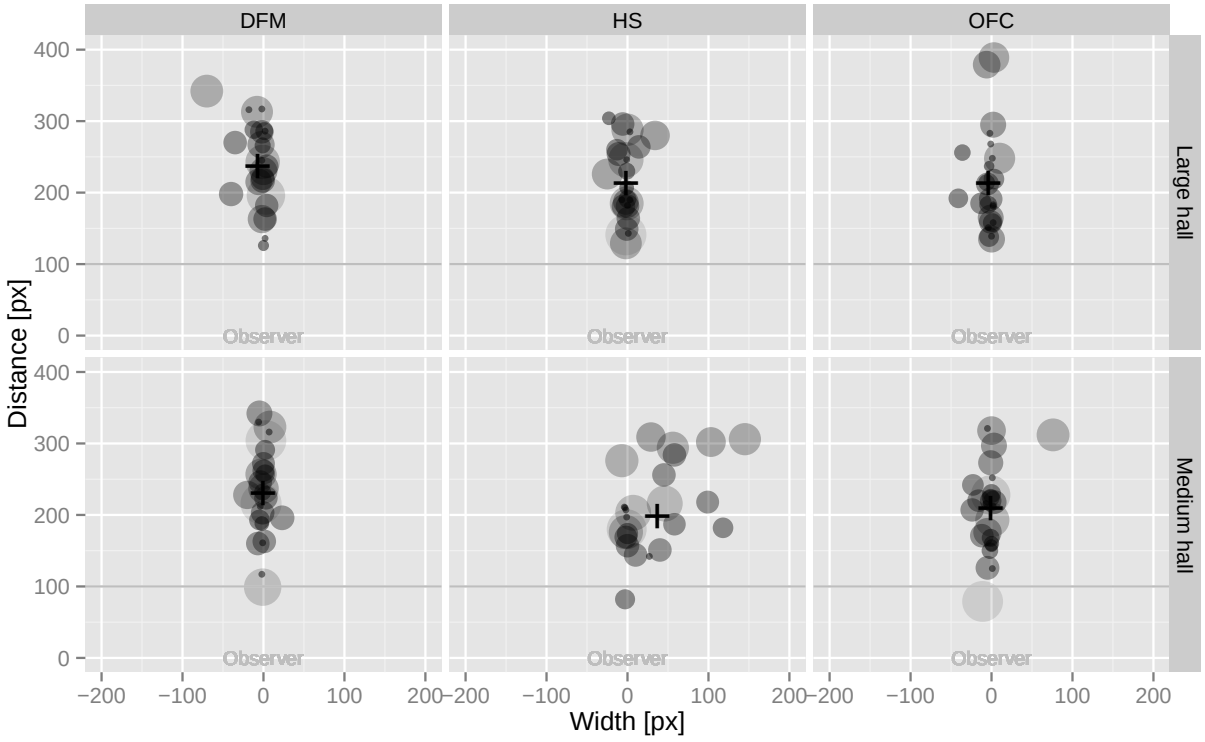

Figure 5.9: Perceived position of the cello in two halls with three methods. Size and transparency of the circles are determined by the width assigned to the cello auditory event; larger radius and more transparency imply more perceived width. The cello sample was always reproduced by the centre loudspeaker. The participant is positioned at [0, 0] pixels and the curtain is at distance of 100 px. The crosses denote the average distance and horizontal offset.

and HS resemble each other, but OFC localizations are clustered more narrowly in distance. Moreover, the bass is localized closer to the curtain than the two other instruments or the bass in the other methods. In the medium hall DFM differs from HS in the localizations of the synthesizer and bass by producing less distant auditory events. Here OFC and DFM resemble each other. This observation is in line with the depth of stage results in Figure 5.8.

5.4.6 Discussion

DFM introduces frequency-dependent amplitude and phase variation similar to that found in an acoustic diffuse field, these variations are similar to what is found if an array of spaced microphones is set up in a recording space. Timbre and sound quality are a substantial focus of the professional sound recording community, and, in this sense, we were pleased

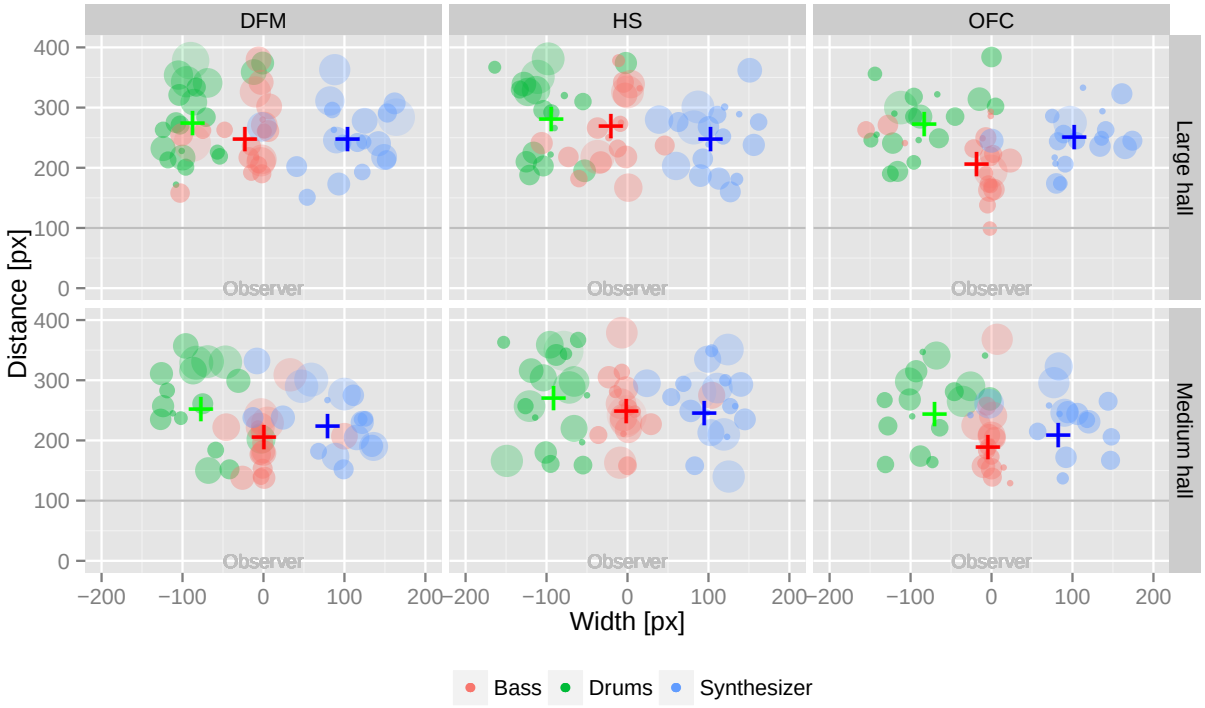

Figure 5.10: Perceived positions of the band instruments in two halls with three methods. Size and transparency of the circles are determined by the width assigned to the individual auditory events; larger radius and more transparency imply more perceived width. The drumset sample was panned between the left and centre loudspeakers, the bass sample was panned to the centre loudspeaker, and the synthesizer sample was panned between the centre and right loudspeakers. The participant is positioned at [0, 0] pixels and the curtain is at distance of 100 px. The crosses denote the average distance and horizontal offset.

to see only small differences in the overall quality between the methods. Further, there is no significant difference compared to HS and a non-significant tendency in favor of DFM compared to OFC. Similar to Schroeder's comments on his artificial reverberator [68], the spectral variation introduced by DFM is too fine to be discerned by the auditory bandwidth of the ear and instead serves to properly decorrelate the signals.

DFM and HS were both rated as having a deep sound stage. It is surprising that this does not directly translate to higher ratings of overall quality based on the regression analysis. However, a listening environment that feels clear and understandable is perceived as having high quality regardless of the depth of stage rating. This implies there were participants who perceived the deep sound stage as a clear listening environment and participants who perceived the opposite: deep sound stage makes the listening environment unclear.

It is extremely surprising that OFC was not rated substantially lower given our expectation of an overly-correlated late field. Listened to in isolation, the reverberation produced with OFC had a narrow and unpleasant image. As a component of a virtual acoustic model, many listeners found it acceptable. We note that OFC's narrow image may have confounded any results with respect to instrument distance. DFM and HS present a physically-plausible approximation of an acoustic space and thus preserved the fragile psychoacoustic cues needed for distance perception. On the other hand, OFC has a narrow image and more distant objects subtend a lesser angle. It is possible that these two different distance cues led to similar results. The lateral facing bi-directional microphones used in HS are likely responsible for the widely scattered source locations in Figure 5.9.

The lack of significant differences in overall quality in Experiment 2 may also result from the heterogeneous nature of our stimulus set. A single mean opinion score is not sufficient to reflect the variation in our participants' reasoning, and qualitative data describing the quality evaluation decisions would have been needed. Informal discussions with our participants revealed that some preferred the closeness of the auditory image and the lack of room impression by OFC, whereas others preferred being immersed in the room and perceiving a deep stage in front of them in the cases of DFM or HS. In the end, it is an aesthetic decision and depends on the application. In this context, immersion in the performance space was not always desired and may even confuse the listener. The situation may be different in games and movies or even for different musical genres or recording traditions. Furthermore, adding visual stimulus may give rise to a different set of quality preferences, as speculated by [67], saying wider sound sources are maybe preferred in conjunction with visual input. Examining DFM in audiovisual settings is subject to future studies.

5.5 Conclusions

Diffuse Field Modeling approach used in this study is a physically-plausible virtual acoustic model for sources that were captured with close microphone placement. This model replaces the panning, delay, and reverberation that would be typically used. DFM is presented in detail in the companion paper (Part I; [60]) and evaluated here (Part II). When comparing different treatments of DFM to create the room effect for a 20-loudspeaker array, we found that it is necessary to model the earliest reflections and to use all four channels of the B-Format room impulse response. Furthermore, we compared DFM to common microphone techniques used to create the room effect for 5.1 content and DFM was found to be perceptu-

ally comparable with the Hamasaki Square technique. Especially, DFM was found to be able to elicit a perception of depth of stage in 5.1 listening which requires new listening habits but also creates new opportunities for aesthetic decisions. In sum, Diffuse Field Modeling is a perceptually viable method to create room impression allowing free placement of anechoic point sources in arbitrary multichannel loudspeaker setups. We see direct application possibilities in object-based audio systems and in game audio.

Chapter 6

Conclusions

6.1 Summary

In Chapter 2, the Hamasaki Square was rated as having greater distance and envelopment compared to an alternative technique. Chapter 3 demonstrated that the auditory system is sensitive to directional energetic differences in diffuse field reverberation. These findings are discussed in the Interim Summary and interpreted in light of known physical acoustics and psychoacoustics.

Chapter 4 describes a systematic method for the reproduction of diffuse sound fields using decorrelation filters based on the statistical description of reverberation. This algorithm is referred to as Diffuse Field Modeling (DFM) and is based on physical acoustics and validated using numerical simulations based on the Kirchhoff / Helmholtz Integral. It was demonstrated that the resulting fields have the expected spatial autocorrelation and that the filters can be tuned to introduce random variation that has physically-plausible frequency autocorrelation. The latter heavily influences the spatial impression. DFM allows the creation of arbitrary observation points on a virtualized array in the recording space from a 4-channel, B-Format microphone array. The channels of the microphone array must be treated such that they can be assumed to be diffuse. The virtual array of outward-facing cardioid microphones used in DFM is able to discriminate between directional energy in the diffuse RIR. The angular differences in directional energy are limited by the angular resolution of the cardioid microphone.

Chapter 5 presents a perceptual evaluation of DFM as a component of a physically-plausible virtual acoustic model that can be systematically adapted to various loudspeaker arrays. The first experiment used a 20-loudspeaker array with a 16-channel lower ring and

a 4-channel height ring. The direct path and reflections were modeled geometrically and positioned using VBAP. The findings indicate that it is necessary to model reflections in conjunction with DFM, and the majority of the participating sound recording professionals found the initial implementation to be useable in practice. A second experiment used the 5.1 loudspeaker configuration and no significant differences were observed between DFM and the Hamasaki Square. This indicates DFM is a viable and flexible solution for practice.

It is notable that the direct path and reflections used in Chapter 5 were modeled geometrically and positioned manually using VBAP. There is a particularly elegant synergy with the virtual microphone (ViMiC) concept introduced by Braasch [10] when the virtual microphones are made to be coincident with the loudspeakers - in fact the cardioid-to-cardioid approximation derived in Chapter 4 could be seen as a physical substantiation of the ViMiC technique. Further, the Max/MSP software package published by Peters [52] could be used as the necessary direct and early reflection "engine" for DFM, and the conventional feedback-based reverberation algorithm used in this software package could be substantially improved upon by incorporating DFM.

A number of results in the evaluation of DFM can be interpreted in terms of known psychoacoustics. The first of these is the lack of coloration introduced by the algorithm for the case where reflections are modeled (the reflections modeled treatment in experiment 1 and the DFM treatment of experiment 2.) DFM introduces substantial amplitude and phase variation into the generated channels of diffuse reverberation analogous to what would be found if an array of room microphones were set up in an actual acoustic diffuse field. It appears that this introduced variation is finer than the frequency resolution of the inner ear (reviewed in Appendix B) and tends to "average out" over a critical bandwidth. A similar point is made by Schroeder in his seminal paper on artificial reverberation [68]. It is also the case that the introduced spectral magnitude variation tends to disappear as the number of loudspeaker channels increases because the mean of the exponential distribution specifying the magnitude squared is unity. This said, the "clipped" treatment containing some early reflections was rated as having a lesser timbral quality than the "reflections modeled" treatment and reinforces the need to insure that specular components not be given diffuse field properties. It is notable that while these early reflections are subject to localization dominance of the direct path, they did influence the timbral quality. It is also notable that the verbal label "too distant" was associated with the "reflections off" treatment. In the experiment presented in Chapter 2 it appears that the reflections may have contributed to high distance ratings (discussed in the Interim Summary). Here, however,

their contribution seems to have been additional source width, which may have contributed to an overall appropriateness of the auditory image.

6.2 Research Answers

The Hamasaki Square is a "direct" microphone technique that has physical interpretation as a sparse sampling of a bounding surface. It provides plausible channel correlation, a facsimile of early reflections, and independence from the direct sound capture. Given the physical interpretation and the acceptance in practice, the first research question was to understand the perceptual qualities of this technique by way of a comparative perceptual evaluation. When used in conjunction with an acoustic curtain array and a delay plan, this approach is rated as having greater distance than all alternative treatments. With either the acoustic curtain or with spot microphones, the Hamasaki Square is rated as having greater envelopment.

Both the physical and perceptual literature typically assume that the reverberant diffuse field has equal energy incident from all directions. The second research question was to determine if the human auditory system is sensitive to directional energetic differences in late field reverberation and if these differences can be observed in existing acoustic spaces. The thresholds for perceptible energetic differences were estimated for a lateral condition (-2.5 dB), for a height condition (-6.8 dB), and for a frontal condition (-3.2 dB). These energetic differences were demonstrated to exist in a shoebox style concert hall by measuring both the experimental stimuli and the concert hall with a B-Format microphone and by forming opposing cardioid microphones. Further, it was shown with dummy head recordings that the spectral differences at the ear drum are surprisingly subtle.

The third objective of the dissertation was to develop an algorithm that allows the systematic creation of physically-plausible diffuse energy. This was done by specially treating a B-Format room impulse response such that structure of the signal could be assumed to be diffuse. This allowed the creation of decorrelation filters based on the statistical description of reverberation that introduce variation similar to what would be found on a real array of cardioid microphones in an acoustic diffuse field. Most notably, the frequency-dependent channel correlation corresponds to known relationships for microphone arrays, the reconstructed fields have the correct spatial autocorrelation, and the frequency spectra at different channel locations change smoothly or coarsely depending on the RT60.

The fourth objective was to validate the developed algorithm in practice. It was demon-

strated that reverberation for a 20-loudspeaker array could be created without coloration or temporal smearing and that the majority of sound recording professionals who participated in the experiment found the algorithm usable in practice. Analysis of both semantic scales and free-format verbal descriptions demonstrated that it is necessary to model early reflections in tandem with the algorithm and that this treatment was described as "being there," "realistic," and "natural." A subsequent experiment used the 5.1 loudspeaker configuration and found no significant differences were found with respect to the Hamasaki Square, which indicates DFM is a viable and flexible solution for practice.

6.3 Contributions

The contributions of this dissertation are:

- The practical knowledge of the perceptual properties of the Hamasaki Square as rated by professional sound recording engineers; particularly notable are the high ratings of distance and envelopment.
- The theoretical knowledge that the auditory system is sensitive to directional energetic differences in diffuse field reverberation and that these differences can be found in existing acoustic spaces.
- The theoretical perspective that the physical structure of an acoustic diffuse field can be efficiently modeled using decorrelation filters.
- The practical ability to generate a large number of channels of reverberation from a single B-Format room impulse response and the theoretical demonstration that these channels correspond to simulated measurements on a virtual microphone array in a diffuse field.
- The practical validation of this algorithm in a large 20-loudspeaker array and in a smaller 5.1 (3/2) array.

6.4 Limitations and Future Work

It was noted in the Introduction (Section 1.1) that a common technique for capturing the diffuse field reverberation is a pair of spaced microphones near the back of the hall. The

strategy compared to the Hamasaki Square in Chapter 2 is a pair of spaced cardioid microphones mounted on top of the acoustic curtain array and directed away from the musical instruments. This was adapted from William's MMA and is called Electronic Time Offset (ETO). Very little direct sound is captured by this technique and it is electronically time aligned in a manner consistent with the delay plan used in conjunction with the Hamasaki Square. While this technique is qualitatively similar to spaced room microphones, it does not enjoy the same familiarity in the sound recording community and has been met with some skepticism. It is believed that the ETO strategy was a suitable comparison, however the perceptual evaluation of the Hamasaki Square would be more palatable to a broader community if spaced room microphones had been used.

In Chapter 3 it was decided to study the reverberant diffuse field in isolation due to a possible interaction with the direct sound and early reflections. The perceptual thresholds established in this work can be viewed as a baseline, however it is clear that future studies are needed to understand the perception of non-ideal diffuse field reverberation in a more plausible context having both direct and early reflections. It would be useful to determine the physical properties (likely the absorption coefficients α) that correspond to the perceptual thresholds and to measure the directional differences in a number of existing acoustic spaces. In the context of concert halls (Appendix B), Bradley and Soulodre have argued that auditory source width (ASW) and listener envelopment (LEV) are distinct perceptual quantities. It would be particularly interesting to substantiate this independence by comparing the perceptual thresholds of non-ideal diffuse field reverberation with and without the early response.

From a technical standpoint, a number of improvements to the Diffuse Field Modeling algorithm are necessary for broad adoption. In the current state of development, Diffuse Field Modeling can generate physically-inspired decorrelation filters for rings of regularly spaced loudspeakers. The two-dimensional restriction allows the introduced randomness ($A_{channel,frequency}$) to be filtered in one dimension "along the array." A three-dimensional array would require a considerably more sophisticated two dimensional filtering scheme "over the surface." A superior solution would be to initially structure the introduced randomness such that any arbitrary loudspeaker array had the proper channel correlation in all three cartesian dimensions. This can be achieved by sampling a numerical simulation of a pure tone diffuse field at different points in space (and frequency.) By sampling a numerical simulation of a diffuse field interacting with a head model (or empirically measured HRTF database), a binaural adaption amenable to head tracking for use in virtual reality scenarios

can be created.

Establishing the correspondence between the room impulse response's RT60 and the frequency autocorrelation of the ensemble of filters Γ is cumbersome in the current implementation. The RT60 of the room impulse response must be estimated for a number of frequency bands and the decay length ν of the ensemble of filters is manually tuned until Γ sufficiently corresponds to the RIR by way of Equation 4.18. The ability to automatically tune the decorrelation filters from an impulse response based on an automatic analysis of the RIR would be a substantial economization of workflow. Finally, the spectral coloration of the diffuse field at the point that the B-Format RIR was measured will influence the entire acoustic display; an equalization scheme would minimize this controllable impact to sound quality.

The channel correlation introduced by Diffuse Field Modeling is physically substantiated, however more research is needed to understand the perceptual ramifications. The known perceptual research on this topic is limited to Blauert's discussion of Damanske's work, Hamasaki's paper regarding his microphone technique, Theile's work citing both Blauert and Hamasaki, and, indeed, Chapter 2 of this dissertation. All of these studies are for four loudspeakers with nearly-regular spacing and suggest that the effect of frequency-dependent correlation is to fuse the loudspeakers into single large auditory image of the diffuse field. Experiments by Hiyama and Hamasaki as well as Santala and Pulkki compared diffuse fields presented on sparse loudspeaker arrays to a dense reference loudspeaker array. Both experiments used the assumption of broadband decorrelated noise for each channel, it would be particularly interesting to revisit this work using the frequency-dependent channel correlation introduced by DFM.

Generating B-Format RIR from offline, high-resolution models appears promising. This could be done, for instance, using the Computer Aided Drafting (CAD) tools employed in architectural acoustics or the geometric meshes used in virtual reality and gaming. This would allow a more computationally efficient auralization of a virtual acoustic space, would serve as an authoring format that does not require sophisticate software tools, and could be adapted to various loudspeaker and (eventually) binaural configurations. For highly varied acoustic spaces, it may be required to use multiple B-Format RIR measurements depending on the position and a scheme for interpolation would be necessary for interactive acoustic displays. Well-regarded concert and recital halls are often suited to particular styles of music; broad adaption will require that DFM be able to translate this aesthetic suitability.

The tuning of the ensemble frequency autocorrelation in DFM leads to substantial

changes in spatial impression. In the pilot testing of Chapter 5, it was observed that the virtual acoustic space could be made "too big" by introducing decorrelation with low frequency autocorrelation. Physically, this would lead to lower values of interaural cross correlation within a given auditory bandwidth. To the best of the author's knowledge, no study exists that links the perceived size of an acoustic space to frequency bands of interaural cross correlation.

The ambiguity regarding the role of early reflections in distance perception is a gap in the literature and should be resolved. A perceptual evaluation could be done using virtual acoustic spaces and might vary the level and timing of the reflections as well as the geometry and material properties of the space. Because DFM is both systematic and physically-substantiated, it is well-suited to being used to model the diffuse field reverberation in perceptual research and, in this particular case, manipulations to the decorrelation filters could be used to vary the interaural cross correlation at the participants ears.

The specific investigations suggested above have emerged directly from the contributions of this dissertation. More generally, the goal of natural, immersive, and aesthetically malleable virtual auditory experiences requires the constant refinement of the physical approximations that enable a perceptually-plausible presentation. Reverberation is a particularly complex physical process and is unlikely to ever be modeled in its exact detail; for this reason the perceptual attributes and interactions of the approximating model must always be evaluated. DFM was developed in the context of music recording and reproduction and offers a potentially unifying approach to the systematic reproduction of reverberation. It is hoped that it will also find application in gaming, virtual reality, telecommunications, and perceptual and hearing research.

Appendix A

Acoustics

A.1 Basic Acoustics

This appendix reviews the acoustic fundamentals that underly this dissertation, in particular the physically-based algorithm presented in Chapter 4. Section A.1.1 discusses the acoustic equations, their linear approximation, and the acoustic wave equation. Section A.1.2 discussed constant frequency solutions to the wave equation including spherical basis functions and multipoles. Section A.2 reviews the physical aspects of room acoustics including the Sabine model as well as the modal, temporal, and statistical descriptions of reverberation.

A.1.1 Acoustic Equations and the Wave Equation

Derivation of the acoustic wave equation relies on the conservation of mass, Newton's Law, and the relationship between pressure and density. Simplifying assumptions are made to yield the linear acoustic equations and are used to derive the wave equation. The wave equation relates spatial and temporal features of an acoustic variable (typically pressure) and solutions to this equation predict observable acoustic phenomenon. Commonly encountered solutions include plane waves and spherical waves.

The Conservation of Mass

The conservation of mass requires that changes in the mass of a fluid particle be equal to the mass flux through the surface bounding the same particle:

$$\frac{d}{dt}\iiint_V \rho \, dV = -\iint_S \rho \mathbf{v} \cdot \mathbf{n} \, dS \tag{A.1}$$

where ρ is the density, V is the volume defining the particle, S is the surface bounding the particle, $\mathbf{v}$ is the (continuum) fluid velocity, and $\mathbf{n}$ is the surface normal vector of S. Gauss's Law is used to relate changes in flux through the surface to the divergence of the vector field within the bounded volume:

$$\iint_S \mathbf{A} \cdot \mathbf{n} \, dS = \iiint_V \nabla \cdot \mathbf{A} \, dV \tag{A.2}$$

Taking $\mathbf{A}$ to be equal to $\rho\mathbf{v}$ and substituting divergence for the surface integral yields:

$$\iiint_V \left[\frac{d\rho}{dt} + \nabla \cdot \rho\mathbf{v} \right] dV = 0 \tag{A.3}$$

Because the volume V is arbitrary, this relationship must be true on a point by point basis as well.

$$\frac{d\rho}{dt} + \nabla \cdot \rho\mathbf{v} = 0 \tag{A.4}$$

If mass is flowing out of an infinitesimal region, then the density of this region is decreasing. If mass is flowing into an infinitesimal region, then the density is increasing.

Equations of Motion

The equations of motion are based on Newton's Second Law which is commonly stated as "The external force acting on a particle is equal to the particles mass times acceleration." The acoustic derivation is based on a moving particle of fluid and instead expressed in terms of momentum as "The external force acting on a particle is equal to the time rate of change of the particle's momentum." Mathematically this is expressed as:

$$\frac{d}{dt}\iiint_{V^*} \rho\mathbf{v} \, dV = \iint_{S^*} \mathbf{f}_S \, dS + \iiint_{V^*} \mathbf{f}_B \, dV \tag{A.5}$$

The first term is the derivative of momentum, with momentum expressed as the volume integral of density (kg/m^3) times velocity. The second term is the force $\mathbf{f}_S$ due to external pressure; the net force will be in the direction of lowest pressure. The third term is the force $\mathbf{f}_B$ due gravity and is typically neglected. The use of the asterisk denotes a moving fluid particle (Lagrangian frame) as opposed to a stationary particle (Eulerian frame). A number of manipulations gives the point by point expression of Newton's Second Law for acoustic particles:

$$\rho \frac{D\mathbf{v}}{Dt} = -\nabla p \tag{A.6}$$

where $\frac{D\mathbf{v}}{Dt}$ is the total derivative accounting for convective flow:

$$\frac{D\mathbf{v}}{Dt} = \frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial x}\frac{dx}{dt} + \frac{\partial \mathbf{v}}{\partial y}\frac{dy}{dt} + \frac{\partial \mathbf{v}}{\partial z}\frac{dz}{dt} \tag{A.7}$$

The spatial terms are neglected in the linear acoustics equations and reduces to $d\mathbf{v}/dt$.

Pressure-Density Relations

Early attempts to characterize the relationship between pressure and density were linear but predicted a speed of sound that disagreed with measurements. The accepted relationship between pressure p and density ρ is:

$$p = K\rho^{\gamma} \tag{A.8}$$

where the exponent γ is determined by thermodynamic processes within the particle of fluid. The specific heat capacity is the energy in Joules needed to raise a unit of mass one degree Celsius. c_p is the specific heat capacity for a gas at constant pressure and c_v is the specific heat capacity for a gas at constant volume. The exponent γ is the ratio $\gamma = c_p/c_v$.

Two processes happen within a particle of fluid. In the first, heat is added at a constant volume (and density) causing pressure and temperature to increase. In the second, heat is removed at a constant pressure causing temperature and volume to decrease while density increases. The two processes are happening simultaneously - the first process accounts for increased pressure, the second process accounts for increased density. They exchange an equal and opposite amount of heat and there is negligible heat exchange with neighboring particles (adiabatic).

Linear Acoustics

A homogeneous medium is one in which all of the ambient variables are independent of position and a quiescent medium is one in which they do not depend on time and the initial velocity is 0. Acoustic phenomena are small amplitude disturbances to the ambient state $(p_0, \rho_0, \mathbf{v}_0)$ of a homogeneous and quiescent medium and are denoted with a prime:

$$p = p_0 + p' \qquad \rho = \rho_0 + \rho' \tag{A.9}$$

These relationships are substituted into the equations for the conservation of mass, Newton's Law, and the pressure density relationship. Temporal or spatial derivatives of constants become zero, and terms involving the product of two primed values $(\rho'\mathbf{v}')$ are neglected. The pressure density relationship is expanded into a Taylor Series and only the first term is kept. These approximations yield the linear acoustic equations:

$$\frac{\partial \rho'}{\partial t} + \rho_0 \nabla \cdot \mathbf{v}' = 0 \tag{A.10}$$

$$\rho_0 \frac{\partial \mathbf{v}'}{\partial t} = -\nabla p' \tag{A.11}$$

$$p' = c^2 \rho' \qquad c^2 = (\frac{\partial p}{\partial \rho})_0 \tag{A.12}$$

The conservation of mass relates the temporal rate of change in density to the divergence (source or sink) of particle velocity - when particles are diverging from a point the density is decreasing. Newton's Law $f = ma$ shows that the force is the negative of the pressure gradient and is equal to the density (infinitesimal mass) multiplied by the acceleration. Particles are accelerated away from the direction of maximum pressure increase. The pressure and density are approximately related by the constant c^2, which is the derivative of pressure with respect to density at ambient conditions.

Wave Equation

The acoustic wave equation states a relationship between the spatial and temporal derivatives of acoustic quantities. It is derived by manipulating the linear acoustic equations and is:

$$\nabla^2 p - \frac{1}{c^2} \frac{\partial^2 p}{\partial^2 t} = 0 \tag{A.13}$$

where ∇^2 is the Laplacian operator $(\nabla \cdot \nabla)$ and represents curvature in space; $\partial^2 p/\partial^2 t$ can be interpreted as curvature in time. The primes have been dropped from the acoustic variables and the ambient state variables have been eliminated from the equation. The Helmholtz equation is the constant frequency version of the acoustic wave equation derived under the assumption that all spatial points vary with the same temporal frequency. Constant frequency solutions are typically expressed in terms of complex numbers, in this case the complex pressure $\hat{p}$.

$$(\nabla^2 + k^2)\hat{p} = 0 \tag{A.14}$$

where k is the wave number describing the angular frequency with respect to space (radians / meter), ω is the angular frequency with respect to time (radians / second), and $k = \omega/c$. For point sources and room modes, the Helmholtz equation can be expressed as a function operator $(\nabla^2 + k^2)$.

A.1.2 Constant Frequency Solutions

The plane wave is an idealized solution to the wave equation where acoustic quantities depend only on time and a directional vector. Acoustic quantities are constant on the plane perpendicular this vector and referred to as surfaces of constant phase. Considering a plane wave directed along the x-axis, the solution to the wave equation is the sum of two arbitrary functions $f(t - x/c)$ and $g(t + x/c)$:

$$p = f(t - x/c) + g(t + x/c) = f(\xi) + g(\eta) \tag{A.15}$$

where the variables ξ and η show the link between time and space for acoustic disturbances:

$$\xi = t - x/c, \quad \eta = t + x/c \tag{A.16}$$

The function $f(t - x/c)$ propagates in the direction of the positive x axis, the function $g(t + x/c)$ propagates in the direction of the negative x axis. Acoustic velocity is related to pressure by the characteristic impedance ρc:

$$v = \frac{1}{\rho c}[f(t - \frac{x}{c}) - g(t + \frac{x}{c})] \tag{A.17}$$

For a plane traveling wave of constant frequency,

$$p = p_{pk} cos(\omega t - kx - \phi_0) = \mathbf{Re}\left[\ p_{pk}\ e^{-j\omega t} e^{jkx} e^{j\phi_0}\right] \tag{A.18}$$

where $k = \omega/c$ (discussed above) and p_{pk} is peak amplitude of the pressure. Plane traveling waves of constant frequency are cyclical in time and in space. The wavelength λ describes the length of a cycle in space, $\lambda = 2\pi/k$, or $\lambda f = c$. A spherically symmetric wave is another idealized solution to the wave equation and, similar to plane traveling waves, the solution is the combination of two arbitrary functions:

$$p(r,t) = \frac{1}{r} f(t - r/c) + \frac{1}{r} g(t + r/c) \tag{A.19}$$

If there is only an outward traveling wave, we consider only one of these functions:

$$p = \frac{1}{r} f(t - r/c) \tag{A.20}$$

At a large distance from a spherical source the outward traveling wave locally resembles a plane wave and the $1/\rho c$ relationship to velocity is approximately true.

Green's Functions

A point source is a solution to the inhomogeneous (driven) Helmholtz operator:

$$(\nabla^2 + k^2) G_k(\mathbf{x}|\mathbf{x_s}) = -4\pi\delta(\mathbf{x} - \mathbf{x_s}) \tag{A.21}$$

where $\mathbf{x}$ indicates the location of observation, $\mathbf{x_s}$ indicates the location of the source, and $\delta(\mathbf{x} - \mathbf{x_s})$ is the Dirac delta distribution which indicates that the equation is driven only at $\mathbf{x_s}$. $G_k(\mathbf{x}|\mathbf{x_s})$ is the free-space Green's function e^{jkR}/R. When combined with the complex constant $\hat{S}$, the free-space Green's function describes an acoustic monopole:

$$\hat{p} = \hat{S}\frac{e^{jkR}}{R} \tag{A.22}$$

This principle of acoustic reciprocity is a property of the Green's function that states that locations of the source and observation can be interchanged:

$$G_k(\mathbf{x}|\mathbf{x_s}) = G_k(\mathbf{x_s}|\mathbf{x}) \tag{A.23}$$

Green's functions can be superimposed to describe more complicated solutions to the Helmholtz equation:

$$\hat{p} = \sum_{n=1}^{N} \hat{S}_n G_k(\mathbf{x}|\mathbf{x_n}) \tag{A.24}$$

Note that each element of the summation may have arbitrary amplitude and phase and may additionally be extended to considered continuous distributions of sources. The dipole is created by spacing two monopoles at a short distance $\mathbf{d}$ with opposite phase:

$$\hat{p} = \hat{\mathbf{D}} \cdot \nabla_S G_k(\mathbf{x}|\mathbf{x_s}) \tag{A.25}$$

where the dipole-moment amplitude vector $\hat{\mathbf{D}}$ is defined by the complex source strength $\hat{S}$ and $\mathbf{d}$ is the vector between the two monopoles. The operator ∇_S is the gradient in local coordinates and the dot product between $\hat{\mathbf{D}}$ and ∇_S gives rise to the common "figure-8" radiation (and reception) patterns. The response is ± 1 on the dipole-moment, and 0 when perpendicular. The approximation holds when the wavelength is much, much longer than the spacing in between two monopoles, that is $kd << 1$.

A.1.3 Spherical Basis Functions and Multipoles

Spherical Basis Functions (SBF) are factored solutions to the Helmholtz equation in spherical coordinates:

$$\Psi(r, \theta, \phi) = \Pi(r)\Theta(\theta)\Phi(\phi) \tag{A.26}$$

where θ is the elevation angle, ϕ is the azimuthal angle, k is the wavenumber, and r is radial distance. In this discussion, order refers to $n = 0, 1, 2, 3.....$ and modes $m = -n, ..., n$. Radial solutions are the Spherical Bessel Functions of the 1st kind $l_n(kr)$ and 2nd kind $y_n(kr)$, as well as Hankel function, which are a combination of the two. Note that Bessel functions are normally denoted $j_n(kr)$ but have been changed to $l_n(kr)$ to avoid confusion with $j = \sqrt{-1}$. For an outgoing spherical wave, $h_n^1(kr) = l_n(kr) + jy_n(kr)$. The superscript 2 on $h_n^2(kr)$ denotes an incoming spherical wave, however these are not typically considered and the notation $h_n(kr)$ is assumed to denote an outgoing spherical wave. The 0^{th}, 1^{st}, and 2^{nd} orders of these functions can be described in terms of elementary transcendental functions, and Hankel functions can be expressed as complex exponentials using Euler's relationship:

$$h_n(kr) = l_n(kr) + jy_n(kr) = \frac{e^j kr}{jkr} \tag{A.27}$$

The angular solutions $Y_n^m(\theta,\phi)$ are typically expressed as the product of an elevation solution $\Theta(\theta)$ and an azimuthal solution $\Phi(\phi)$ and are referred to as Spherical Harmonic Functions (SHF). The elevation solution is a Legendre Function $\Theta(\theta) = P_n^m(\cos(\theta))$ and is a solution to a Legendre differential equation that arises from the factorization and subsequent manipulations. For positive m and n, Legendre Functions are often expressed in terms of simpler Legendre Polynomials and negative orders and modes can be described using symmetry relationships. The azimuthal solution $\Phi(\phi)$ can be thought of as a Fourier Series with angle ϕ:

$$\Phi(\phi) = B_1 \sin(m\phi) + B_2 \cos(m\phi) \tag{A.28}$$

Functions of negative order can be expressed as the complex conjugate of functions of a positive order $Y(\theta,\phi)_n^{-m} = \overline{Y(\theta,\phi)_n^m}$. SHF with $m < n$ are tesseral (checker board pattern), $m = n$ are sectoral (bi-directional patterns on the x/y plane), and $m = 0$ are zonal (similar to global trade winds). The product of radial and angular functions yields solutions to the Helmhotlz equation:

$$R_n^m(\mathbf{r}) = Y(\theta,\phi) l_n(kr) \qquad S_n^m(\mathbf{r}) = Y(\theta,\phi) h_n(kr) \tag{A.29}$$

where $R_n^m(\mathbf{r})$ is regular at $\mathbf{r} = 0$ and $S_n^m(\mathbf{r})$ is singular at $\mathbf{r} = 0$. Given the weights of of each SBF, arbitrary regular fields can be expressed as an infinite series of R_n^m. The fields generated by acoustic sources can be described as infinite series of S_n^m. In practice, these series must be truncated; higher-order series have greater radial extent, greater angular variation, and are convergent with the original field at greater distances from the origin. The weights of these series can be determined in a number of ways. The first, most common means is the surface scalar product of the pressure and pressure gradient with the SBF on a unit sphere. The second involves derivatives of an arbitrary, regular sound field at the origin. Each derivative can be linked to various SBF using appropriate coefficients, the 0^{th} and 1^{st} orders being respectively proportional to the pressure and pressure gradient at the origin. The derivatives of the radial functions of a given radial order can be expressed in terms of radial functions of the two adjacent orders. That is, $f_n' \propto f_{n+1} - f_{n-1}$.

Multipoles are directional derivatives of the Green's function $G(r)$ and are closely related to Spherical Basis Functions [53][28]. Green's Functions $G(\mathbf{r})$ are related to SBF by:

$$G(\mathbf{r}) = \frac{e^{jk|\mathbf{r}|}}{4\pi|\mathbf{r}|} = jk\sqrt{\frac{1}{4\pi}} S_0^0(\mathbf{r}) \tag{A.30}$$

$$S_0^0(\mathbf{r}) = Y_0^0(\theta,\phi)h_0(kr) = \sqrt{\frac{1}{4\pi}}\frac{e^{jk|\mathbf{r}|}}{jk|\mathbf{r}|} \tag{A.31}$$

where the notation $G(\mathbf{x}|\mathbf{x_0})$ has been changed to $G(\mathbf{r})$, $\mathbf{r} = \mathbf{x} - \mathbf{x_0}$. A dipole can be expressed as a configuration of out-of-phase monopoles or as a directional derivative. The relationships between the two horizontal plane dipoles and SBF are:

$$\frac{d}{dx}G(\mathbf{r}) = \frac{jk^2b}{2\sqrt{4\pi}}[S_1^1(\mathbf{r}) + S_1^{-1}(\mathbf{r})] \tag{A.32}$$

$$\frac{d}{dy}G(\mathbf{r}) = \frac{k^2b}{2\sqrt{4\pi}}[S_1^1(\mathbf{r}) - S_1^{-1}(\mathbf{r})] \tag{A.33}$$

where b is a constant based on order and mode. Note that the real MP dipole requires the sum of the complex SBF. The equation for the z-axis and the general relationship between multipoles and Spherical Basis Functions is given in Gumerov and Duraiswami [28].

A.2 Physical Aspects of Room Acoustics

The mathematical model of diffuse sound field is first reviewed followed by a discussion of Sabine's empirical reverberation model. Modal acoustics are reviewed and then tied back to the Sabine model through Schroeder's criterion for modal density. The time domain view of reverberation is then presented, followed by the statistical / stochastic view.

A.2.1 Diffuse Sound Fields

The diffuse field model is a mathematical idealization that approximates the field resulting from an acoustic source that has been reflected many, many times [53][22]. The field is regarded as a superposition of freely propagating plane waves, each of which is traveling in a different direction. For constant frequency fields, the complex pressure field $\hat{p}$ is described as a summation of plane waves with individual amplitude $\hat{p}_q$ and direction $\mathbf{n}_q$:

$$\hat{p} = \sum_q \hat{p}_q e^{\mathbf{n}_q \cdot \mathbf{x}} \tag{A.34}$$

The acoustic energy density w is used in the diffuse field model, the Sabine reverberation

model, and to determine absorption coefficients:

$$w = (\frac{1}{2}\rho_0 v^2) + (\frac{1}{2}\frac{p^2}{\rho_0 c^2}) \tag{A.35}$$

The quantity $(\frac{1}{2}\rho_0 v^2)$ is commonly referred to as Acoustic Kinetic Energy Density, and $(\frac{1}{2}\frac{p^2}{\rho_0 c^2})$ is referred to as Acoustic Potential Energy Density. For a traveling plane wave, energy density can also be described in terms of pressure p:

$$(\frac{1}{2}\rho_0 v^2) = (\frac{1}{2}\frac{p^2}{\rho_0 c^2}) = \frac{w}{2} \tag{A.36}$$

Because only plane waves traveling in the same direction contribute to the energy density the summed average value $\bar{w}$ can be computed as the sum of the constituent plane waves and is approximately equal to the time average of the pressure squared:

$$\bar{w} = \frac{1}{2\rho_0 c^2}\sum_q |\hat{p}_q|^2 \quad \approx \quad \frac{1}{2\rho_0 c^2}\bar{p^2} \tag{A.37}$$

A portion $w_{\delta\Omega}$ flows through the cone defined by the solid angle $\delta\Omega$ centered at a single point. If this cone is centered around the direction $\mathbf{e}$, the directional energy density $\mathbf{D}(\mathbf{e})$ is the limit as $\delta\Omega$ becomes small. For an ideal diffuse field, $\mathbf{D}(\mathbf{e})$ is independent of direction $\mathbf{e}$. For average energy density $\bar{w}$,

$$\mathbf{D}(\mathbf{e}) = \bar{w}/4\pi. \tag{A.38}$$

Near an open window or highly absorptive material a diffuse field will departs from this ideal but is a reasonable approximation in the center of a room. For rooms with pronounced variation in absorptive materials $\mathbf{D}(\mathbf{e})$ will depend upon direction $\mathbf{e}$ in a frequency-dependent manner.

A.2.2 Sabine-Franklin-Jaeger Theory of Reverberant Rooms

The Sabine reverberation model is an approximation [53]. It is based upon the idea that reverberant acoustic energy is approximately the same in all spatial regions of the room, and that the absorption due to the bounding walls can be treated "on average." It is derived in terms of the conservation of acoustic energy. P is the acoustic power that is inputted to the room. P_d is the acoustic power dissipated by the walls bounding the room. The acoustic energy density w (defined above) is the sum of the potential acoustic energy due to pressure,

and the kinetic acoustic energy due to particle velocity. The integration of w over the volume of the room will be influenced by the difference between the acoustic power inputted, and the acoustic power dissipated.

$$\frac{d}{dt}\iiint w dV = P - P_d \tag{A.39}$$

The local spatial average of w is approximately independent of position after a large number of reflections have occurred, one common estimation is the hundredth reflected wave. Following Pierce's derivation, the quantities above are replaced by their running time averages, indicated by $\bar{w}, \bar{P}, \bar{P}_d$. Given the assumption of spatial uniformity, the absorbed power $\bar{P}_d$ depends only upon $\bar{w}$.

$$\bar{P}_d = \frac{c}{4} A_s \bar{w} \tag{A.40}$$

where c is the speed of sound, and A_s is a frequency-dependent property of the room with units of area, discussed below. The assumption of spatial uniformity simplifies the triple integration of acoustic energy density to a product of $V\frac{d\bar{w}}{dt}$. Substitution of the power dissipation quantity P_d into the above equation yields a differential equation:

$$V\frac{d\bar{w}}{dt} + \frac{c}{4} A_s \bar{w} = \bar{P} \tag{A.41}$$

The solution to this differential equation is the decaying exponential:

$$\bar{w}(t) = \bar{w}_{initial} e^{-t/\tau}, \tau = \frac{4V}{cA_s} \tag{A.42}$$

where the quantity τ is the characteristic decay time. It is more common, however, to discuss reverberation in terms of $RT60$. On a logarithmic scale, exponential processes have a linear scale, $RT60$ is defined as the amount of time needed for $\bar{w}(t)$ to decay 60dB from an initial amplitude. Sabine's major contribution is the fact that he established the empirical relationship:

$$RT60 = \frac{0.161V}{\sum_i \alpha_i A_i} \tag{A.43}$$

Here α_i is the *absorption coefficient*, determined empirically from the measurement of $RT60$ in rooms. α_i is 1 for an open window, which will transmit all energy incident upon it. Other materials are defined as the ratio of the $\delta\bar{P}_d/\delta A$ (power dissipated per area) to this open window. The summation of α_i in the denominator indicates that the various surfaces of a room can be considered to contribute to an overall absorption constant A_s, determined by

their relative area A_i. α_i is typically frequency dependent. Brick has values of approximately 0.05, draperies range from 0.03 for low frequencies, to 0.35 for high frequencies. Suspended acoustic tiles have values as high as 0.99 over 1kHz, with comparative values below this (0.93 at 250 Hz).

The frequency-dependent reverberant decay of an acoustic space can be graphed using the energy decay relief. This is obtained by bandpass filtering the room impulse response to the desired bandwidth and reverse integrating the signal squared from $t = \infty$ to $t = 0$. High frequencies typically decay faster than low frequencies. In Pierce, the authors Franklin and Jaeger are appended to the Sabine model to credit them with the substantial work done after Sabine to refine various aspects of the model.

A.2.3 Modal Theory of Room Acoustics

The Sabine-Franklin-Jaeger model was developed using meaningful approximations to an extremely complicated physical scenario and was substantiated empirically. The modal approach to room acoustics is derived by finding solutions to the Helmholtz equation in a rigidly bounded room. By making considerations of the bandwidth and density of these solutions, a "considerably less approximate" verification of the Sabine-Franklin-Jaeger model is found [53]. We seek a solution to the Helmholtz equation (constant frequency wave equation) given the boundary conditions of perfectly rigid walls and a rectangular geometric configuration. Walls are the planes $x = 0$, $x = L_x$,$y = 0$, $y = L_y$,$z = 0$, $z = L_z$. Individual solutions will be of the form:

$$p = \mathbf{\Psi}(\mathbf{x}, n) \tag{A.44}$$

where n is an index and $\mathbf{\Psi}(\mathbf{x}, n)$ is a constant frequency eigenfunction that describes the complex amplitude (magnitude and phase) with respect to space. The Helmholtz equation must be satisfied within the volume, and that the spatial derivate normal to the bounding surfaces must be zero:

$$[\nabla^2 + k^2]\mathbf{\Psi}(\mathbf{x}, n) = 0 \quad within\ V \tag{A.45}$$

$$\nabla\mathbf{\Psi}(\mathbf{x}, n) \cdot \mathbf{n}_{out} = 0 \quad on\ S \tag{A.46}$$

Algebraic manipulation leads to equations for x, y, z separately. The equation for the x dimension is:

$$\frac{d^2X(x)}{dx} + k_x^2 X(x) = 0 \tag{A.47}$$

where k_x is the projection of k onto the x axis and $k_x^2 + k_y^2 + k_z^2 = k^2$. The normal gradient of $\mathbf{\Psi}(\mathbf{x}, n)$ must be zero at the boundaries $\mathbf{x} = 0$ or $\mathbf{x} = L_x$ and will be satisfied when:

$$\frac{d\mathbf{\Psi}(\mathbf{x}, n)}{dx} = \frac{d\cos k_x x}{dx} = \sin k_x x \tag{A.48}$$

where $k_x = \pi n_x / L_x$. The three dimensional eigenfunction is:

$$\mathbf{\Psi}(\mathbf{x}, n_x, n_y, n_z) = A\cos\frac{n_x \pi x}{L_x}\cos\frac{n_y \pi y}{L_y}\cos\frac{n_z \pi z}{L_z} \tag{A.49}$$

and the wavenumber is:

$$k^2(n_x, n_y, n_z) = \pi^2\left[(\frac{n_x}{L_x})^2 + (\frac{n_y}{L_y})^2 + (\frac{n_z}{L_z})^2\right] \tag{A.50}$$

Following the definition of orthogonality, the volume integral of the point-by-point product of any two eigenfunctions (volume dot product) will be zero when the indexes n_x, n_y, n_z are not equal. The arbitrary constant A can be chosen such that the mean square volume average is equal to 1, forming an orthonormal spatial basis set. An arbitrary pressure field can then be decomposed onto this basis set using the volume dot product.

Axial, tangential and oblique are the names given to different index combinations. The time domain solution describes a mode as two traveling waves superimposing over space to create points of constructive and destructive interference. Axial modes are analogous to strings or acoustic pipes bouncing back and forth between two parallel walls, this corresponds to a single dimension having a non-zero index. The visualization of tangential and oblique modes is less straight-forward.

A scenario of particular interest is the rectangular, rigid-walled room driven by the Green's function $\delta(\mathbf{x} - \mathbf{x_0})$. The derivation is based upon the decomposition onto the modal orthonormal basis set, an involves a number of approximations. It yields:

$$\hat{p} \approx \frac{\hat{S}}{V}\sum_n \frac{\mathbf{\Psi}(\mathbf{x}, n)\mathbf{\Psi}(\mathbf{x_0}, n)}{k^2 - k(n)^2 + ik/c\tau} \tag{A.51}$$

The complex pressure in the room is proportional to the source power $\hat{S}$, and inversely proportional to the volume of the room. The Green's function excites the various eigenfunctions differently depending upon their response at the excitation point x_0 and the spatial variation will be determined by both the excitation and response at various observation points.

The denominator depends upon the excitation's spatial frequency k, the individual eigenfunction's spatial frequency $k(n)$, and the reverberation decay constant τ. Various modes are exited depending upon the driving frequency, the degree to which a mode is excited depends upon the difference between k and $k(n)$ and the bandwidth of the mode. When the modes decay quickly, their bandwidth is large, and more distant excitation frequencies will influence a mode. When the modes ring for some length of time, their bandwidth is more narrow, and distant excitation frequencies have less influence over a given mode. The modal density is the number of modes within a frequency bandwidth, and is defined by:

$$\frac{dN}{d\omega} = \frac{1}{2}\frac{V}{\pi^2}\frac{\omega^2}{c^3} \tag{A.52}$$

When the modal resonances are closely spaced relative to their bandwidth, multiple modes are excited for a given input, and the associated resonance effect is not as clear. The Schroeder cutoff frequency is defined as the frequency where the modal spacing is less that 1/3 the modal bandwidth. Above this, the distribution of modes can be considered a continuous distribution. The primary utility of this is that it allows the modal summations to be considered as integrals, which allows the development of equations that bear striking resemblance to Sabine's. There is a particular elegance to the establishment of this empirical theory and its later mathematical validation by modal theory. Sufficiently high modal density has implications for spatial uniformity - a large density of simultaneously ringing modes will insure that there are not pronounced spatial variations in pressure.

A number of "real world" considerations are relevant to the discussion of modal acoustics [21]. Below the lowest mode of a given space, there are no resonance due to modes. This is referred to as the pressure chamber region. While there is no "gain" due to resonance, the acoustic pressure is bounded to the volume, and has a greater amplitude than if the excitation source were unbounded. Above the lowest mode, and below the Schroeder cutoff frequency, the modal density of the room is sparse. There is resonant boosting of certain frequencies, but insufficient modal density to satisfy the criterion of 3 modes per bandwidth. This leads to uneven frequency response and spatial standing wave patterns, a problem endemic to small listening spaces [21]. Lastly, given the highly directive nature of high frequency sources, and the typically quick absorption of those frequencies, it is unclear if high frequency modes are truly excited. It is common to speak of high frequencies as being in the "ray region" of the room [21]. Ray (geometric) acoustics is discussed in Section A.2.4.

A.2.4 Time Domain View of Reverberation

The time domain view of reverberation is typically discussed in terms of geometric acoustics, a view that considers acoustic energy to travel in narrow, nearly planar beams [21][7]. Typical acoustic sources radiate in all directions and some of this energy will be reflected off of the walls. In idealized scenarios the image-source method can be used to determine the direction, time of arrival, and spherical spreading attenuation for the 2^{nd}, 3^{rd} ..... N^{th}-order reflections. Echo density is the is the infinitesimal limit of the number of reflections per unit time, and for simple cases it increases as time squared. With no absorption, spherical spreading predicts that a given high-order reflection will have energy inversely proportional to t^2 while the echo density conversely increases as t^2. Reverberation is the gradual spreading of temporally compact energy into temporally spread energy. Spectral power density is conserved, however phase and timing information is not. The coherent phase relationship of spectrally complicated sources is disturbed and instead replaced with random phase.

More realistically, the acoustic processes of absorption, scattering, diffraction, shadowing, and refraction influence every individual reflection based on the local properties of the wall and the frequency of the beam being considered. The direction of propagation is always normal to surfaces of constant phase and the width of the beam must be large compared to the wavelength of the sound. Absorptive materials will attenuate the level and typically have low-pass frequency characteristics. Scattering materials "break up" the reflection: instead of a specular reflection with a clear angle of incidence and reflection, the reflected wave has a broad radiation pattern. Qualitatively, a scattering material disrupts the surfaces of constant phase in the quasi-planar ray of sound. For a given wavelength, a larger scattering object disrupts a surface more than a smaller scattering object. For a given object size, a smaller wavelength disrupted more than a longer wavelength. Diffraction allows sound to "bend around a corner" when there is no geometric ray path and depends heavily on frequency. Conversely, shadowing is the loss of higher frequencies when an object obstructs a geometric path. In mediums with constant temperature and density the speed of sound c is constant and the rays travel in straight lines. Rays can be refracted when thermal currents cause varying c.

For simple, regular shapes such as the rectangular room the Laplace transform can convert an impulse response into an modal pattern. Closed form solutions only exist for very simple shapes [7]. Computer aided drafting (CAD) tools can be used for complicated, irregular geometrics and include models of absorption and scattering.

A.2.5 Statistical Aspects of Room Acoustics

The physical aspects of room acoustics are frequently described statistically; Blesser comments that statistical interpretations are appropriate when that the underlying process has elements of physical indeterminacy such as thermal noise, or that the process so complex that it is implausible to describe minute detail [7]. An infinite sum of random oscillators will produce a random time waveform; the same is true a near-infinite sum of random, exponentially decaying room modes (discussed above with respect to the Schroeder cutoff frequency). Blesser, citing Pollack, notes that the time domain view of the reverberation process "is provably statistical when the number of simultaneous echo arrivals reaches a limit of about 10." Neither the modal nor temporal views are valid prior to the mixing time of the room.

Two statistical characterizations of physical reverberation are particularly important in this dissertation. Spatial autocorrelation tells us how the pressure p at various locations in space change relative to each together. For a set direction and distance, a high correlation value tells us that pressure is changing in phase and with a similar amplitude. Low correlation values tell us that the pressure is in a reactive phase and/or has differing amplitudes. High negative correlation values tell us that the pressure is changing with similar amplitude but in opposite phase. Frequency autocorrelation tells us the degree to which adjacent frequencies ω and $\omega + \Delta\omega$ change coherently. It is a description of how jagged or smooth a spectrum is, note that the spectrum is different at different locations. In tandem, the two quantities tell us how much an acoustic field changes in space with respect to frequency. Detailed treatments are found in Pierce [53].

Spatial Autocorrelation

The diffuse field is a mathematical idealization that was described in Section A.2.1 and which considers the field to be a superposition of plane waves of random direction, phase, and amplitude. It is physically accurate after the mixing time of the room, that is, after an acoustic source has been reflected many, many times. For a constant frequency pressure field, the autocorrelation is given by expected value with respect to space of the product of $p(\mathbf{r})$ and $p(\mathbf{r} + \mathbf{\Delta r})$ where r is the observation point and Δr is a fixed offset. This offset will cause the arguments of each constituent plane wave q to proceed differently, in the limit of infinite plane waves this allows us to treat the pressure as a random variable. Mathematical

manipulation leads to the familiar description of spatial autocorrelation [53][15]:

$$\langle p(\mathbf{r})p(\mathbf{r}+\Delta\mathbf{r})\rangle = \langle\overline{p^2}\rangle\frac{\sin k|\Delta r|}{k|\Delta r|} \tag{A.53}$$

where k is the wave number and $\langle\overline{p^2}\rangle$ is the mean square pressure averaged over space. Lower frequencies will be more correlated for greater amounts of space, higher frequencies less so. For a given frequency, autocorrelation is periodic with space, though greater distances are considerably less correlated. The same equation arises if we instead fix our observation to a single point in space and consider the sphere defined by the magnitude vector $|\Delta r|$. This formulation is used in the simulation in Section 4.5, a similar approach was used by Elko in [19]. The derivation of Equation A.53 is based on the value of pressure and corresponds to measurement by an omni-directional microphone. Additional considerations regarding the microphone directivity pattern and orientation are shown in [19]. Parallel bi-directional microphones at various spacings are correlated to higher frequencies than are omni-directional microphones. Cardioid microphones have correlation determined by the pattern orientation as well as the spacing. For identical orientation, the spatial autocorrelation for bi-directional, cardioid, and omni-directional microphones differ by their frequency shape. For coincident microphones, Faller [23] shows that correlation ranges from 1/3 for opposing cardioids, 2/3 for a 90 degree orientation, and 1 for the identical orientation.

The diffuse field approximation yields an important constraint that can be used to inform the introduction of randomness in DFM. Consider the pressure observed along the vector $\mathbf{r}_{1,2}$ connecting two arbitrary observation points $\mathbf{r}_1$ and $\mathbf{r}_2$. For a given spatial frequency $k = \omega/c$, the maximum phase progression corresponds to the plane wave propagating along $\mathbf{n_q}$ exactly parallel to $\mathbf{r}_{1,2}$. Due to their projection onto $\mathbf{r}_{1,2}$ the arguments of all other diffuse plane waves proceed more slowly and will contribute a lesser value of k. If a collinear array of equally spaced points is considered, the observed spatial frequencies on $\mathbf{r}_{1,2}$ are all k up to an ideal brick-wall cutoff of $k = \omega/c$. If we consider the observations at each point to be a random process, the autocorrelation is a transform pair with the power spectral density [30]. That is, that the observed spatial frequencies on a collinear array (or nearly collinear array) in a diffuse field are linked to the spatial autocorrelation of the observations through the spatial Fourier Transform. The use of dipole or cardioid capsules changes the weight of the observed spatial frequencies and thus the spatial autocorrelation of the observations.

Frequency Autocorrelation

The modal approach to room acoustics is obtained by finding solutions to the constant frequency wave equation (Helmholtz equation) in a rigidly bounded room, the structure of the modes is determined by the room's geometry [53][7]. Qualitatively, each mode is a spatial description of the pressure field that will be excited to varying degrees depending on the location of a source and the difference between the modal frequency and the source's driving frequency. There are many discrete modes spaced over frequency, and the modes typically overlap depending upon their frequency bandwidth.

Rooms with short RT60 values have modes that decay quickly and which have large bandwidths. A given driving frequency will influence a relatively large range of modes. Rooms with long RT60 values have modes which ring for a greater duration of time and which have more narrow bandwidth. As a result, distant driving frequencies have less influence. When the modal resonances are closely spaced relative to their bandwidth, multiple modes are excited for a given input. The Schroeder cutoff frequency is defined as the frequency beyond which the modal spacing is less that 1/3 the modal bandwidth, and, in this case, the summation of discrete modes can be considered a continuous distribution of frequency components.

For a constant frequency the complex pressure field is the superposition of all excited modes and is calculable for a given source and observation point. Similar to the plane waves of the diffuse field idealization, however, the contribution of each individual mode changes if either the source or observation move. Above the Schroeder frequency the field can be described statistically in terms of the frequency autocorrelation and the mean square pressure distribution. The frequency autocorrelation is the expected value of the product of the mean square pressure at two adjacent frequencies averaged over all spatial observation points. For the time average of the squared acoustic pressure $\overline{p^2}(\omega)$ and source frequency ω, this is given by:

$$\langle \overline{p^2}(\omega)\overline{p^2}(\omega+\Delta\omega)\rangle = \langle \overline{p^2}(\omega)\rangle^2 \left[1+\frac{1}{1+(\tau\Delta\omega)^2}\right] \tag{A.54}$$

The fact that the frequency autocorrelation depends on the characteristic decay time τ is extremely important in DFM. When the term $(\tau\Delta\omega)^2$ is large, the right hand term is near 1.0. This indicates that the mean squared pressure at adjacent frequencies differs, as would be expected for large frequency spacings $\Delta\omega$ or for long modes with large values of τ and narrow bandwidths. Conversely, when $(\tau\Delta\omega)^2$ is vanishingly small, the right hand term is near 2.0, which indicates that the mean square pressure at adjacent frequencies is largely

identical, as would be expected for small frequency spacings $\Delta\omega$ or for short modes with small values of τ and wide bandwidths. The smoothness of the spectrum appears to play a large role in the spatial impression.

The mean squared pressure in the modal approximation of a nearly-rigid walled room follows an exponential distribution with probability distribution function $\mu(\overline{p^2})$ of mean squared pressure $\overline{p^2}$.

$$\mu(\overline{p^2}) = \lambda e^{-\lambda \overline{p^2}} \text{ for } \overline{p^2} \geq 0 \tag{A.55}$$

with $\lambda = \frac{1}{\langle \overline{p^2} \rangle}$. The value of $\overline{p^2}$ is always positive, however it is approximately twice as likely for a given value of $\overline{p^2}$ to be below the mean $\langle \overline{p^2} \rangle$ than above the mean. Because of this, the average sound pressure level (SPL) is 2.5 dB lower than the SPL level corresponding to $\langle \overline{p^2} \rangle$. The exponential distribution is used in DFM to define the magnitude squared values of the decorrelation filters, and the 2.5 dB SPL is used to compensate the gain of the filters.

The real and imaginary parts of a modal sound field are uncorrelated, zero-mean Gaussian random variables. As such, all phase values between $-\pi$ and π are equally likely.

A.3 Summary

The wave equation is a mathematical expression that is derived from the conservation of mass, the laws of motion, and a relationship between pressure and density. Planar and spherical solutions to the wave equation are used throughout the dissertation, and spherical basis functions and modal solutions are conceptually related. Reverberation is the process that describes an acoustic source bound within a volume. The diffuse field model considers and infinite number of plane waves incident from all directions and approximates the acoustic field when a source has been reflected many, many times. This can be related to the time domain view of reverberation but does not explicitly describe the processes of absorption and scattering related to a particular room geometry. The modal model considers solutions to the wave equation that describe variation in space when considering rigid or semi-rigid boundary conditions. Qualitatively modes are sharp resonances formed by traveling waves being reflected within the enclosure and depend on the geometry of that enclosure. The resonances are sharp when the RT60 is long and wide when the RT60 is short. The diffuse sound field gives rise to the statistical view of spatial autocorrelation, and the modal model gives rise to the statistical view of frequency autocorrelation.

Appendix B

Psychoacoustics

This appendix reviews the fundamentals of psychoacoustics. Section B.1.1 reviews the binaural localization of coherent sources, Section B.1.2 reviews the structure and function of the inner ear, critical bandwidth, auditory filters, and associated masking effects. The psychoacoustics of complex environments begins with the Precedence Effect in Section B.2.1, diffuse and incoherent sources in Section B.2.2, and distance hearing in Section B.2.3. The perceptual impression of concert halls is reviewed in Sections B.3.1 and B.3.2.

B.1 Binaural Localization of Coherent Sources

B.1.1 Interaural Time and Level Differences

The ability of the auditory system to localize auditory events is based largely upon differences in the acoustic signals received at the two ears [6][48]. In the horizontal plane, a sound event not lying in the median plane will have differing times of arrival at the two ears. This is referred to as the interaural time difference (ITD.) Differences in the sound pressure level at the two ears arise due to a number of acoustic processes and are typically frequency-dependent. Diffraction, which allows sound to bend around obstructions, yields little level difference at low frequencies, but substantial head-shadowing at moderate to high frequencies. When the wavelength of the sound is on the order of the pinna, the acoustic processes of diffraction, dispersion, and reflection impose fine spectral detail upon the signals at each ear. The head related transfer function (HRTF) is an empirical measurement of the signal at each ear for a given direction of incident sound and may be used to determine the frequency-dependent interaural level difference (ILD.)

The acoustic properties of sound give some indication of the frequency regions where each cue provides the most information. The ITD is roughly equal for all frequencies, however, the maximum path difference between the ears is between 600 μs and 700 μs for most normal subjects. This imposes a limit on the signals that can be effectively compared phase differences become ambiguous for half periods that are shorter than this. Said differently, when the period is less than 1.2 ms to 1.4 ms, the auditory system cannot figure out which ear is the leading phase, and which ear is the lagging phase. As such, differences in the fine structure of a signal are only useful cues when the wavelength is longer than this criterion, that is, somewhere below 800 Hz.

Between 800 Hz and 1.5 - 1.6 kHz, head movements and spectral cues help us resolve leading and lagging phase. However, above 1.5 - 1.6 kHz, experiments have shown that we are no longer sensitive to changes in phase. We can, however, pick out interaural differences in the energy envelope of such signals. Localization experiments involving high frequency carriers and differing lower frequency modulation signals give similar results to experiments using signals composed only of the differing low frequency modulators. The maximum firing rate of the auditory nerve roughly corresponds to these frequency limits.

Interaural level differences are used as a cue throughout the frequency range. ILDs are minimal at low frequencies where the wavelengths are much, much greater than the size of the head. In these cases, the incident waves are able to wrap around the head without much loss. ITDs are the dominant cue in this frequency range. As frequency increases, the head's shadowing effect becomes more pronounced. At still higher frequencies, when the wavelength is on the order of the pinna's features, fine spectral detail varies as a function of the direction of the sound event. These fine spectral features are particularly important for tracking elevation and resolving front-back cues.

B.1.2 Inner Ear, Critical Bands, and Masking Effects

The mechanics of the cochlea and basilar membrane are very important aspects of the auditory system [48]. When the ear drum is set into motion by external pressure changes, the vibration is transferred to the cochlea though a mechanical impedance matching network of tiny bones and muscles. The cochlea itself is a conical, fluid-filled tube wound into a nautilus-like shell, and is divided in half by the basilar membrane (BM.) The opening of the cochlea is called the oval window, the far end is referred to as the apex. Sinusoidal ear input signals cause traveling waves in the fluid along the length of the BM. The maximum displacement of the spatial envelope depends on frequency - high frequency sinusoids will

cause maxima closer to the oval window, while lower frequencies will cause maxima closer to the apex.

Along the BM are sensing mechanisms commonly referred to as hair cells, each cell is connected to about 20 nerve fibers of the auditory nerve. Attached to each hair cell are approximately 40 very fine stereocilia (sometimes called "hairs.") The motion of the BM causes a deflection of the stereocilia, and a resulting chemical reaction within the hair cell causes the nerve fibers to fire. The resulting system can be viewed as approximating a short-time Fourier analysis - incident sinusoidal components are mapped to specific places along the BM and then to specific fibers within the auditory nerve. However, even a pure steady-state sinusoidal signal will excite a region of the BM around its center frequency. This non-zero excitation width influences the frequency resolution of the auditory system, and gives rise to the critical bandwidth, the auditory filter, and simultaneous masking (discussed below.)

The critical band is a concept that is defined behaviorally in terms of a detection experiment [48]. A sinusoidal test signal is presented in tandem with a noise signal that has constant power spectral density. Increasing the bandwidth of the noise increases the signal's overall power. The detectability threshold of the sinusoid increases linearly with logarithmic bandwidth until a plateau is reached. Beyond this, increases in the noise bandwidth have no influence upon the detection threshold of the test signal. The interpretation of this well-known result is that when the ear is attempting to detect a signal, it is taking information from select regions of the BM. Only noise within a certain region of the BM will mask the test signal, and frequencies outside of this region do not contribute to the masking effect. The power spectrum model assumes that a signal is detectible when the signal-to-noise ratio within this critical bandwidth exceeds a constant.

The region of the BM that a particular nerve of the auditory nerve is "paying attention to" is described as an auditory filter. It is typically approximated as being rectangular and is quantified using the Equivalent Rectangular Bandwidth (ERB.) This bandwidth is proportional to center frequency and this proportion increases from 11 to 17 percent of the center frequency. More realistically, the auditory filter has sloping edges and is narrower at lower SPL compared to higher SPL. This is due to the outer hair cells which mechanically amplify the motion of the BM; their response can be signal-driven or under some control from higher processing centers of the auditory system [48].

Because a tone of a given SPL and frequency will excite a region of the BM, another tone of lesser SPL and similar frequency might not be readily detected by the auditory

system. This phenomenon is known as simultaneous masking and is also discussed as the frequency resolution of the auditory system. Masking thresholds are described by a family of non-linear curves surrounding a center frequency, their slope and width varies with the SPL and frequency of the masking tone. They are qualitatively similar but not identical to approximations of auditory filters [48].

Temporal masking, also referred to as non-simultaneous masking, is a phenomenon where a probe sound is inaudible when it is in close temporal proximity to a masking sound [48]. The sounds do not overlap in time. Backward masking refers to the case where the probe precedes the masking signal, and can be understood as revising the past. The phenomenon is not well understood. On the other hand, forward masking refers to the situation where the probe signals follows the masking noise. A relatively louder masking sound makes it more difficult to detect a relatively quieter probe, and more masking occurs for shorter time delays. The time scale and signal levels are on the order of the precedence effect (up to approximately 40ms) discussed in Section B.2.1. The major difference between these phenomenon is that forward masking is presented monophonically over headphones and the precedence effect and acoustic reflections pertain to acoustic sources separated spatially. See Moore [48] Figure 3.17 for more information.

B.2 Fundamental Psychoacoustics of Complex Environments

B.2.1 The Precedence Effect and Early Reflections

The Precedence Effect, Haas Effect, and Law of the First Wavefront all refer to the fact the direct sound determines the primary direction of the auditory event in the presence of secondary wavefronts [6], [48]. Three different phenomenon occur depending upon the time delay between spatially separated coherent sources. The auditory event will be localized between the two sources when the wavefronts are separated by a delay of approximately 630 μs to 1000 μs. This is referred to as time-dependent summing localization and is the phenomenon responsible for phantom images in stereophonic media (in addition to level-dependent summing localization) [6]. As the delay approaches the limit of 1000 μs, the auditory event will eventually "snap" to the leading source and will continue to appear here for delays up to approximately 40 ms. The secondary wavefront is perceptible in the fact that it influences the width and tone color of the auditory event. Beyond 40 ms, the second

wavefront becomes a distinctly perceptible echo. More generally, the relative timing and amplitude of the secondary wavefront determines if it is detectible (masked threshold) or if it will be heard as an independent auditory event (echo threshold). See Blauert [6] Figure 3.13 for more information. Time values are somewhat less for impulsive sounds [6],[48].

The direction from which the secondary wavefront comes will also influence the perception of the auditory event, in general an echo is more likely to be audible when the azimuth of the second wavefront differs greatly from the direct sound. The presence of more than one secondary wavefront extends the duration of masking depending on the configuration of the various wavefronts. When audible, the additional wavefronts influence the spatial extent and timbre in a similar manner [6]. Considering early reflections in a reverberant environment, the precedence effect is what allows the auditory system to locate a source in the presence of the conflicting directional information of the reflections. Given the myriad possibilities of reflection geometries and frequency-dependent wall properties, it is difficult to define any general rule other than to say that the auditory event associated with primary wavefront will be influenced to some degree [6]. One example of this influence is Auditory Source Width, discussed below in Section B.3.1.

B.2.2 Coherence

The diffuse field is a mathematical model comprised of many plane waves in many directions and approximates what happens when a source is reflected many times in an enclosure (A.2.1)[22][53] and is perceptually related to an auditory event heard "everywhere" [6]. Considering the interaction of a diffuse field with two spaced microphone channels, it can be seen that each constituent plane wave will project differently depending on the spacing, directivity pattern, the angle of incidence, and frequency. The measure of cross-correlation is a means of estimating the similarity of two given signals. The equation is defined as:

$$\Phi_{xy}(\tau) = \frac{\int_{-\infty}^{\infty} x(t)y(t-\tau)dt}{\sqrt{\int_{-\infty}^{\infty} x^2(t)dt \int_{-\infty}^{\infty} y^2(t)dt}} \tag{B.1}$$

The numerator integrates the product of two time sequences over their duration, this is done for any specified time offset. The denominator is the product of their root-mean-squares and normalizes the numerator to the energy in the signals. The maximum value of this equation is unity, and this can be achieved by signals differing only by their amplitude, signals differing only by a time delay, or signals differing only by a 180° phase shift [6]. If the signals are

the two microphones discussed above, a high cross-correlation value tells us that pressure is changing in phase and with a similar amplitude. Low cross-correlation values tell us that the pressure is in a reactive phase and/or has differing amplitudes. High negative correlation values tell us that the pressure is changing with similar amplitude but in opposite phase. The structure of the diffuse field model leads to frequency and space dependent cross-correlation; this is further complicated when the two "microphones" are outer ears, in which case the statistical relationship will be determined by each plane wave's interaction with the torso, head, and pinna.

Signals with values of cross-correlation near unity are referred to as coherent. Signals with cross-correlation values near zero are incoherent. In the perceptual literature, coherence is defined as:

$$k_\Phi = \max_\tau |\Phi_{xy}(\tau)| \tag{B.2}$$

Where the subscript Φ is used to distinguish between coherence k_Φ and wavenumber k. In headphone presentation, coherence influences the lateral width of an auditory event heard on the interaural axis [6]. If the same signal is sent to both ears (perfect coherence), the auditory event will be heard in the center of the head with relatively concise lateralization. As coherence nears zero, that is, perfect incoherence, two separate auditory events will be heard: one located at each ear. In between these two extremes, the width of the auditory event has width related to the value of k_Φ with lower values being relatively wider. This scenario is related to the perception of a diffuse field, the particular values of k_Φ depend on the listener's anthropomorphic features and properties of the room, most notably the RT60 and frequency autocorrelation discussed in Section A.2.5. See Blauert [6] Figure 3.24 for more information.

Channel coherence in loudspeaker presentation [6] is fundamental to the work presented in this dissertation. If a number of coherent sources are presented, the auditory event is heard near or above the head and can be unpleasant. As it approaches zero, distinct auditory events are heard at the individual loudspeakers. Intermediate to this is an auditory event of large spatial extent filling the listening area, the width of which is related to the channel coherence k_Φ. See Blauert [6] Figure 3.43 for more information. This wide and "fused" image corresponds to the reproduced image of the diffuse field that is desired in Chapters 4 and 5.

When multiple sources of varying coherence are presented over headphones, the auditory system is able to resolve them into separate auditory events depending upon their coherence. When monophonic speech is presented in conjunction with monophonic noise, both will be heard at the center of the head. If the noise is instead made to be incoherent, the auditory

system will resolve three events the speech at the center of the head and a noise event at each ear [6].

It was noted by von Bekesy in 1931 that an acoustic event seems more reverberant when the listener plugs one ear; it is perceived as less reverberant when listened to binaurally. An experiment by Danilenko used amplitude modulated white noise in a reverberant room, the effect of the reverberation was to smooth out the effects of the amplitude modulation. Subjects could listen to the sound binaurally within the room or to a monophonic headphone presentation outside of the room where a single microphone signal was sent to both ears. Subjects were able to detect significantly smaller amounts of amplitude modulation in the binaural presentation. This result can be interpreted by considering the fact that reverberation is uncorrelated between the ears whereas the direct path is correlated. This demonstrates that the auditory system is able to recognize the coherent, amplitude-modulated direct path as distinct from the diffuse energy. This experiment is one substantiation of the auditory system's ability to localize a sound source within a reverberant environment [6].

B.2.3 Distance Hearing

Hearing in the horizontal plane is aided by time and level differences between the two ears. Elevation hearing relies predominantly upon spectral cues from the torso, head, and pinna. For anechoic conditions, distance hearing relies almost exclusively on the intensity of the acoustic signal received from the source and is considerably less accurate than hearing in azimuth or elevation [6][48][80]. Signals with greater intensity tend to be heard closer than quiet signals regardless of a loudspeaker's actual distance. Establishing the relative levels of a sound source can sharpen distance perception; cues can be taken from multiple sources or from the movement of a single source. Considering spherical spreading's influence on intensity, the difference limen for source distance is comparable to the difference limen for source intensity [80]. Expectations regarding the loudness and timbre influence the perception of distance, examples of this include both whispering and yelling. The perception of loudness causes auditory event to become darker as the SPL of the signal is increased and is one of the cues used in the estimation of near-field auditory events. Modest spectral changes occur with respect to distance - below 3m there are shadowing effects due to the curvature of the wavefront and above 15m there is absorption due to the air attenuation of high frequencies.

The extent of auditory space is not necessarily the same as acoustic space [6][48][80]. When level is the only cue, the distance of the auditory event is less than the distance of the sound event. This is referred to as the Auditory Horizon and, considering a number

of different studies on distance, can be well approximated [80] by the compressive power function:

$$r' = \kappa r^a \tag{B.3}$$

where r' is the distance of the auditory event, r is the physical distance, and κ and a are fit parameters derived from multiple studies with typical values of ≈ 1.1 and 0.45, respectively.

Most real world listening situations involve acoustic sources in enclosed spaces and the auditory system will have the additional cues resulting from early reflections and diffuse field reverberation [6][48][80]. In this case, the intensity of the overall ear signals may change very little with respect to physical distance, however the ratio of the coherent binaural signal will change relative to the incoherent binaural signal. Moore, citing experiments by von Bekesy, discusses the influence of the ratio of direct sound to reverberant energy and the timing of the early reflections. Adjusting either parameter led to the impression of the auditory event moving toward or away from the subject. Further, later experiments demonstrated that subjects could make absolute distance judgements of unfamiliar stimuli based upon the direct to reverberant ratio [48]. Estimates of distance are more accurate in reverberant environments [80]. Bronkhorst [12] shows distance to be related to the direct to reverberant ratio and creates a mathematical auditory model to explain the auditory horizon discussed above. Note, however, that a very low number of reflections were used in this paper (1, 3, 9, 27, and an "extreme" case of 800 reflections having an impulse duration of approximately 110ms) and are better interpreted as early reflections [53],[7]. This said, in subsequent work distance was shown to depend on the ratio of deterministic and stochastic energy at the listener's ear and is consistent with the notion that the direct to reverberant ratio is a primary cue. The multitude of reflections in the late reverberation are not individually audible and it is known that the direct sound suppresses the early reflections (the precedence effect.) Both, however, appear to be used by the auditory system to estimate the distance an auditory event.

B.3 Subjective Measures of Room Acoustics

B.3.1 Auditory Source Width and Listener Envelopment

The subjective measure of "Spaciousness" or "Spatial Impression" is considered to have two major components, the Auditory Source Width (ASW) and Listener Envelopment (LEV) [11], [51], [45]. The two other subjective measures frequently encountered are "Reverberance"

and "Clarity" (C80,) both are discussed below in Section B.3.2. These four measures are considered to be the "principle components" of concert hall acoustics [76].

ASW is defined by Bradley and Soulodre as "the width of a sound image fused temporally and spatially with a direct sound image." [11]. Okano, Bernake, and Hidaka have a similar definition: "ASW is the apparent auditory width of the sound field created by a performing entity as perceived by a listener in the audience area of a concert hall." [51]. Listener Envelopment (LEV) is described as "the fullness of sound images around a listener" by Bradley and Soulodre, and is influenced primarily by the late reverberation of the room impulse response. Discussed below, the objective physical measures associated with LEV are identical to those used for ASW, with the notable exception that they correspond to different temporal regions of a hall's response.

The physical measures typically associated with ASW and LEV are the degree of interaural cross-correlation at the listener's ears and the degree of lateral energy in the room response. Interaural cross-correlation is based upon the cross-correlation between the two ears [49], [76],[6] and can be determined using Equation B.1. The Lateral Fraction (LF) is defined in terms of the response of a bi-directional microphone with its null oriented towards the direct source:

$$LF_0^{\infty} = \frac{\int_0^{\infty} p^2(t) cos^2(t) dt}{\int_0^{\infty} p^2(t) dt} \tag{B.4}$$

ASW is typically associated with the value of IACC computed over the first 80 ms of a binaural impulse response and is denoted IACC_0^{80} to indicated the limits of integration to (0 to 80 ms.) LF is also defined in terms of temporal regions of the impulse response and LF for the first 80 ms of the impulse response is denoted LF_0^{80}. The objective physical measures associated with LEV are identical to those used for ASW, with the notable exception that they correspond to the time after 80 ms. These are denoted $\mathrm{IACC}_{80}^{\infty}$ and $\mathrm{LF}_{80}^{\infty}$, the corresponding limits of integration are adjusted to 80 ms to the end of the impulse response.

Okano, Bernake, and Hikada created controlled, plausible replications of concert halls for loudspeaker display [51] using a variety of time delays, angles of incidence, and number of reflections. ASW was found to increase proportionally with number of early lateral reflections up to a limit of about 10 reflections and was also found to depend upon the angle of incidence of the reflections for octave bands centered at 125Hz, 250Hz, and 500Hz. An experiment by Bradley and Soulodre [11] asked subjects to rate ASW for stimuli of varying C80 and LF_0^{80}, analysis by ANOVA revealed a main effect of C80 upon ASW, as well as a main effect of LF_0^{80} upon ASW. ASW is greater as LF_0^{80} is increased and or as C80 is decreased (greater reverberation.) Notable is the significant interaction between C80 and LF_0^{80}, the presence of

more reverberant energy (low C80) made it more difficult to detect changes in LF_0^{80}.

LEV is influenced by both reverberation time and Clarity (C80.) Greater reverberation times are rated as having greater LEV while lesser C80 is rated as having greater LEV. In loudspeaker simulations of concert halls, diffuse energy incident from wider angles is rated as having greater LEV. ASW and LEV are distinct perceptual entities and depend upon distinct physical measures. In an experiment [11] with two distinct levels of both LF_0^{80} and $\mathrm{LF}_{80}^{\infty}$, subjects' ratings corresponded to the expected ASW and LEV.

B.3.2 Reverberance and Clarity

Soulodre and Bradley defined reverberance as "the degree of perceived reverberation in a temporal sense. The blending of one sound into subsequent following sounds." Reverberance is a perceptual quantity that is usually correlated with the objective physical measurements of reverberation time ($RT60$) or Early Decay Time (EDT) [70]. Reverberation decays linearly on a logarithmic amplitude scale, the typical measurement of this is the amount of time needed for the reverberation to decay 60 dB from its original level. It is sometimes extrapolated from the -5dB and -35dB points [70] to insure that the initial direct sound decay is not included in the exponential decay, and such that the ending measurement point is above the noise floor. Early Decay Time (EDT) is an estimate of $RT60$ that is extrapolated from the 0 dB to -10 dB points of the response. Soulodre and Bradley found subjective ratings of reverberance to correlate very well with $RT60$ ($r = 0.740$). Incrementally higher correlations ($r = 0.799$) could be found by restricting the correlation to the octave bands centered at 500Hz and 1kHz. The notable finding, however, was the superior correlation with EDT ($r = 0.971$).

Clarity is closely related to the degree of reverberance was defined by Soulodre and Bradley as "the clarity or definition of the sound. The ability to perceive musical detail. The degree to which notes are separated in time." Perceptual measures of Clarity are usually related to the physical measure C80, which defined as dB ratio of early to late energy:

$$C80 = 10 \log \frac{\int_0^{80} p^2(t)dt}{\int_{80}^{\infty} p^2(t)dt} \tag{B.5}$$

C80 is expressed in dB. High values are comparatively dry, low values are more reverberant. 80 ms is the time used to distinguish the early response from the late response. Some authors object to the notion of a decisive "cutting point" between the early and late aspects, and instead prefer the temporal moment of intertia, or Center Time (TS):

$$TS = \int_0^\infty tp^2(t)dt \tag{B.6}$$

TS is measured in seconds.

B.4 Summary

The auditory system uses level and time differences between the two ears to determine the location of sound events in physical space. Time differences are most important at low frequencies, while level differences and the temporal envelope are more important at high frequencies. The inner ear performs a pseudo-Fourier Analysis on the ear input signal, dividing the sound into band-passed channels and reporting this information as discrete nerve firings to higher-level processes. The width of these band-passed channels, or auditory filters, is fundamental to the frequency resolution of the ear and to simultaneous masking. Forward temporal masking occurs when a relatively louder sound makes it difficult to detect a relatively quieter subsequent sound.

The precedence effect is a phenomenon where the first wavefront to reach the ears is dominant in determining the location of an auditory event. Coherent reflections delayed between 1 ms and 40 ms are audible in the width and color of the primary auditory event. They do not, however, form distinct auditory events. Incoherent sources sound wide, or have no distinct location. Distance hearing relies predominantly on relative level differences, and is highly dependent upon familiarity with the source material. Distance hearing in enclosed spaces is aided by the direct-to-reverberant ratio.

Auditory Source Width (ASW), Listener Envelopment (LEV), Reverberance, and Clarity are perceptual measures of concert hall acoustics. ASW is correlated with the physical measure of interaural cross-correlation and lateral fraction in the first 80 ms of a room response, LEV is correlated with the same quantities measured from 80 ms to the end of the response. ASW and LEV are distinct perceptual parameters. Reverberance is a perceptual measurement of reverberation, and is most highly correlated with the physical measure of Early Decay Time. Clarity is perceptual measure associated with the balance of early and late energy in a room, and is correlated with the physical measures of C80 and Center Time (TS).

Bradley and Soulodre give the following interpretation of ASW and LEV in terms of the Precedence effect: "The difference between ASW and LEV can be explained in terms of well-known properties of human hearing .. sound arriving shortly after the direct sound is

integrated or temporally and spatially fused with the direct sound. Thus increasing levels of early lateral reflections increase the apparent level of the direct sound and cause a slight ambiguity in its perceived location .. resulting [in an] increase [of] ASW. Later arriving sound is not integrated or temporally and spatial fused with the direct sound and lead to more spatially distributed effects that appear to envelop the listener" [11].

Bibliography

[1] Jens Ahrens. *Analytic Methods of Sound Field Synthesis*. Springer-Verlag Berlin Heidelberg, 2012. Chapter 1.

[2] Sarah Le Bagousse, Catherine Colomes, and Mathieu Paquier. State of the Art on Subjective Assessment of Spatial Sound Quality. In *Audio Engineering Society 38th International Conference: Sound Quality Evaluation*, July 2010.

[3] Søren Bech. Spatial Aspects of Reproduced Sound in Small Rooms. *The Journal of the Acoustical Society of America*, 103(1):434–445, 1998.

[4] Søren Bech and Nick Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. Wiley & Sons Ltd, Chichester, England, 2007.

[5] Jan Berg and Francis Rumsey. Validity of Selected Spatial Attributes in the Evaluation of 5-channel Microphone Techniques. In *Audio Engineering Society 112th Convention*, Munich, Germany, 2002.

[6] Jens Blauert. *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, Cambridge, Massachusetts and London, England, Oct. 1996. Chapters 2 and 3.

[7] Barry Blesser. An Interdisciplinary Synthesis of Reverberation Viewpoints. *The Journal of the Audio Engineering Society*, 49(10):867–903, 2001.

[8] Alan D. Blumlein. British Patent Specification 394,325 - Improvements in and relating to Sound-Transmission, Sound-Recording and Sound-Reproducing Systems. *The Journal of the Audio Engineering Society*, 6(2):91–98, 130, 1931.

[9] Maurice Bouéri and Chris Kyriakakis. Audio Signal Decorrelation Based on a Critical Band Approach. In *Audio Engineering Society 117th Convention*, Oct. 2004.

[10] Jonas Braasch. A Loudspeaker-Based 3D Sound Projection using Virtual Microphone Control (ViMiC). In *Audio Engineering Society Convention 118*, May 2005.

[11] John S. Bradley and Gilbert A. Soulodre. The Influence of Late Arriving Energy on Spatial Impression. *The Journal of the Acoustical Society of America*, 97(4):2263–2271, 1994.

[12] Adelbert W. Bronkhorst and Tammo Houtgast. Auditory Distance Perception in Rooms. *Nature*, 397(6719):517–520, Feb. 1999.

[13] Cédric Camier, Francois-Xavier Féron, Julien Boissinot, and Catherine Guastavino. Tracking Moving Sounds: Perception of Spatial Figures. In *International Conference on Auditory Display (ICAD)*, Graz, Austria, 2015.

[14] Sylvain Choisel and Florian Wickelmaier. Relating Auditory Attributes of Multichannel Sound to Preference and to Physical Parameters. In *Audio Engineering Society 120th Convention*, Paris, France, 2006.

[15] Richard K. Cook. Measurement of Correlation Coefficients in Reverberant Sound Fields. *The Journal of the Acoustical Society of America*, 27(6):1072–1077, 1955.

[16] Wesley L. Dooley and Ronald D. Streicher. M-S Stereo: A Powerful Technique for Working in Stereo. *The Journal of the Audio Engineering Society*, 30(10):707–718, 1982.

[17] John Eargle. *The Microphone Book*. Focal Press, Burlington MA, 2004. Chapters 4 and 5.

[18] Edo Hulsebos, Diemer de Vries, and Emmanuelle Bourdillat. Improved Microphone Array Configurations for Auralization of Sound Fields by Wave-Field Synthesis. *The Journal of the Audio Engineering Society*, 50(10):779–790, 2002.

[19] Gary W. Elko. Spatial coherence functions for differential microphones in isotropic noise fields. In Michael Brandstein and Darren Ward, editors, *Microphone Arrays*, Digital Signal Processing, pages 61–85. Springer-Verlag Berlin Heidelberg, 2001.

[20] Sonic Emotion. Commercial Website of the Sonic Emotion Company. `http://http://www2.sonicemotion.com/`, 2016. (Last viewed December 2016).

[21] F. Alton Everest. *The Master Handbook of Acoustics*. McGraw-Hill, New York, 1994.

[22] Frank J. Fahy. *Foundations of Engineering Acoustics*. Academic Press, London, 2000.

[23] Christof Faller. Modifying the Directional Responses of a Coincident Pair of Microphones by Post-Processing. *The Journal of the Audio Engineering Society*, 56(10):810–822, 2008.

[24] Sunish George, Slawomir Zielinski, Francis Rumsey, Phillip Jackson, Robert Conetta, Martin Dewhirst, David Meares, and Søren Bech. Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings. *The Journal of the Audio Engineering Society*, 58(12):1013–1031, 2010.

[25] Michael A. Gerzon and John M. Woram. Blumlein Stereo Microphone Technique and Author's Reply. *The Journal of the Audio Engineering Society*, 24(1):36–38, 1976.

[26] George A. Gescheider. *Psychophysics: The Fundamentals, 2nd Edition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1985. Chapters 1 and 2.

[27] Catherine Guastavino and Brian F. G. Katz. Perceptual Evaluation of Multi-Dimensional Spatial Audio Reproduction. *The Journal of the Acoustical Society of America*, 116(6):1105–1115, 2004.

[28] Nail A. Gumerov and Ramani Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Elsevier Series in Electromagnetism. Elsevier Science, Amsterdam, 2005.

[29] Kimio Hamasaki. Multichannel Recording Techniques for Reproducing Adequate Spatial Impression. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, July 2003.

[30] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, New York, 1996.

[31] Koichiro Hiyama, Setsu Komiyama, and Kimio Hamasaki. The Minimum Number of Loudspeakers and its Arrangement for Reproducing the Spatial Impression of a Diffuse Sound Field. In *Audio Engineering Society 113th Convention*, Oct. 2002.

[32] Patty Huang and Jonathan Abel. Aspects of Reverberation Echo Density. In *Audio Engineering Society 123rd Convention*, Oct. 2007.

[33] Finn Jacobsen and Thibaut Roisin. The Coherence of Reverberant Sound Fields. *The Journal of the Acoustical Society of America*, 108(1):204–210, 2000.

[34] John C. Steinberg and William B. Snow. Auditory Perspective – Physical Factors. *Transactions of the American Institute of Electrical Engineers*, 1(53):12–17, 1934.

[35] Jean-Marc Jot, Laurent Cerveau, and Olivier Warusfel. Analysis and Synthesis of Room Reverberation based on a Statistical Time-Frequency Model. In *Audio Engineering Society 103rd Convention*, Sept. 1997.

[36] Gary S. Kendall. The Decorrelation of Audio Signals and its Impact on Spatial Imagery. *Computer Music Journal*, 19:71–87, 1995.

[37] Corey Kereliuk and Philippe Depalle. Sparse Atomic Modeling of Audio: A Review. In *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11), Paris, France*, Sept. 2011.

[38] Sarah Le Bagousse, Mathieu Paquier, and Catherine Colomes. Families of Sound Attributes for Assessment of Spatial Audio. In *Audio Engineering Society 129th Convention*, San Francisco, (CA), 2010.

[39] SoundField Ltd. Commercial Website of the SoundField Company. `http://www.soundfield.com`. (Last viewed 07/01/2016).

[40] William L. Martens and Sungyoung Kim. Verbal Elicitation and Scale Construction for Evaluating Perceptual Differences between Four Multichannel Microphone Techniques. In *Audio Engineering Society 122nd Convention*, Vienna, Austria, 2007.

[41] Douglas McKinnie and Francis Rumsey. Coincident Microphone Techniques for Three-Channel Stereophonic Reproduction. In *Audio Engineering Society 102nd Convention*, Mar. 1997.

[42] Fritz Menzer and Christof Faller. Investigations on Modeling BRIR Tails with Filtered and Coherence-Matched Noise. In *Audio Engineering Society 127th Convention*, Oct. 2009.

[43] Fritz Menzer and Christof Faller. Investigations on an Early-Reflection-Free Model for BRIR. *The Journal of the Audio Engineering Society*, 58(9):709–723, 2010.

[44] Fritz Menzer, Christof Faller, and Hervé Lissek. Obtaining Binaural Room Impulse Responses From B-Format Impulse Responses Using Frequency-Dependent Coherence Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):396–405, 2011.

[45] Juha Merimaa. *Analysis, Synthesis, and Perception of Spatial Sound - Binaural Auditory Modeling and Multichannel Loudspeaker Reproduction*. PhD thesis, Laboratory of Acoustics and Audio Signal Processing, Department of Electrical and Communications Engineering, Helsinki University of Technology, Finland, 2006.

[46] Juha Merimaa and Ville Pulkki. Spatial Impulse Response Rendering I: Analysis and Synthesis. *The Journal of the Audio Engineering Society*, 53(12):1115–1127, 2005.

[47] Jens Meyer and Gary Elko. A Highly Scalable Spherical Microphone Array Based on an Orthonormal Decomposition of the Soundfield. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1781–1784, 2002.

[48] Brian C.J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 2003.

[49] Masayuki Morimoto and Kazuhiro Iida. Appropriate Frequency Bandwidth in Measuring Interaural Cross-Correlation as a Physical Measure of Auditory Source Width. *Acoustical Science and Technology*, 26(2):179–184, 2005.

[50] Rozenn Nicol. Sound Spatialization By Higher Order Ambisonics: Encoding And Decoding A Sound Scene In Practice From a Theoretical Point Of View. In *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, 2010.

[51] Toshiyuki Okano, Leo L. Beranek, and Takayuki Hidaka. Relations Among Interaural Cross-Correlation Coefficient, Lateral Fraction, and Apparent Source Width in Concert Halls. *The Journal of the Acoustical Society of America*, 104(1):255–65, July 1998.

[52] Nils Peters, Tristan Matthews, Jonas Braasch, and Stephan McAdams. Spatial Sound Rendering in Max/MSP with ViMiC. In *Proceedings of the 2008 International Computer Music Conference*. Citeseer, 2008.

[53] Allan D. Pierce. *Acoustics: An Introduction to Its Physical Principles and Applications*. The Acoustical Society of America, Melville, New York, 1994. Chapter 6.

[54] Irwin Pollack and WJ Trittipoe. Binaural Listening and Interaural Noise Cross Correlation. *The Journal of the Acoustical Society of America*, 31(9):1250–1252, 1959.

[55] Nicolaas Prins and Frederick Kingdom. Palamedes: Matlab Routines for Analyzing Psychophysical Data. `http://www.palamedestoolbox.org`. (Last viewed 08/05/2015).

[56] John Proakis and Dimitris Manolakis. *Digital Signal Processing*. Prentice Hall, Englewood Cliffs, New Jersey, 1996.

[57] Ville Pulkki. Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *The Journal of the Audio Engineering Society*, 45(6):456–466, June 1997.

[58] Ville Pulkki. Generic panning tools for MAX/MSP. In *International Computer Music Conference (ICMC)*, pages 304–307, Berlin, Germany, 2000.

[59] Ville Pulkki. Spatial Sound Reproduction with Directional Audio Coding. *The Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[60] David Romblom, Philippe Depalle, Catherine Guastavino, and Richard King. Diffuse Field Modeling using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part I Algorithm. *The Journal of the Audio Engineering Society*, 64(4):177–193, April 2016.

[61] David Romblom, Catherine Guastavino, and Philippe Depalle. Perceptual Thresholds for Non-Ideal Diffuse Field Reverberation. *Revised for the Journal of the Acoustical Society of America*, 0.0:0–0, 2016.

[62] David Romblom, Catherine Guastavino, and Richard King. A Comparison of Recording, Rendering, and Reproduction Techniques for Multichannel Spatial Audio. In *Audio Engineering Society 133rd Convention*, Oct. 2012.

[63] David Romblom, Richard King, and Catherine Guastavino. A Perceptual Evaluation of Recording, Rendering, and Reproduction Techniques for Multichannel Spatial Audio. In *Audio Engineering Society 135th Convention*, Oct. 2013.

[64] David Romblom, Richard King, and Catherine Guastavino. A Perceptual Evaluation of Room Effect Methods for Multichannel Spatial Audio. In *Audio Engineering Society 135th Convention*, Oct. 2013.

[65] Olli Rummukainen, David Romblom, and Catherine Guastavino. Diffuse Field Modeling using Physically-Inspired Decorrelation Filters and B-Format Microphones: Part II Evaluation. *The Journal of the Audio Engineering Society*, 64(4):194–207, April 2016.

[66] Francis Rumsey. Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *The Journal of the Audio Engineering Society*, 50:651–666, 2002.

[67] Francis Rumsey. Spatial audio and sensory evaluation techniques-context, history and aims. In *Spatial Audio and Sensory Evaluation Techniques*, pages 1–7, Guildford, UK, 2006.

[68] Manfred Schroeder. Natural Sounding Artificial Reverberation. *The Journal of the Audio Engineering Society*, 10(2):219–223, 1962.

[69] Barbara G. Shinn-Cunningham. The Perceptual Consequences of Creating a Realistic, Reverberant 3-D Audio Display. In *Proceedings of the International Congress on Acoustics (CICA), Kyoto, Japan*, 2004.

[70] Gilbert A. Soulodre and John S. Bradley. Subjective Evaluation of New Room Acoustic Measures. *The Journal of the Acoustical Society of America*, 98(1):294–301, 1995.

[71] Ron Streicher and Wes Dooley. Basic Stereo Microphone Perspectives-A Review. *The Journal of the Audio Engineering Society*, 33(7/8):548–556, 1985.

[72] Günther Theile. The New Sound Format: 3/2-Stereo. In *Audio Engineering Society 94th Convention*, Mar. 1993.

[73] Günther Theile. Multichannel Natural Recording Based on Psychoacoustic Principles. In *Audio Engineering Society 108th Convention*, Feb. 2000.

[74] John Usher and Wieslaw Woszczyk. Interaction of Source and Reverberance Spatial Imagery in Multichannel Loudspeaker Audio. In *Audio Engineering Society 118th Convention*, Barcelona, Spain, 2005.

[75] Vesa Valimaki, Julian D. Parker, Lauri Savioja, Julius O. Smith, and Jonathan S. Abel. Fifty Years of Artificial Reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1421–1448, July 2012.

[76] Jasper van Dorp Schuitman, Diemer de Vries, and Alexander Lindau. Deriving Content-Specific Measures of Room Acoustic Perception Using a Binaural, Nonlinear Auditory Model. *The Journal of the Acoustical Society of America*, 133(3):1572, 2013.

[77] Juha Vilkamo, Tapio Lokki, and Ville Pulkki. Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation. *The Journal of the Audio Engineering Society*, 57(9):709–724, 2009.

[78] Felix A. Wichmann and N. Jeremy Hill. The Psychometric Function: I. Fitting, Sampling, and Goodness of Fit. *Perception & Psychophysics*, 63(8):1293–1313, 2001.

[79] Michael Williams and Guillaume Le Du. Multichannel Microphone Array Design. In *Audio Engineering Society 108th Convention*, Feb. 2000.

[80] Pavel Zahorik, Douglas S. Brungart, and Adelbert W. Bronkhorst. Auditory Distance Perception in Humans: A Summary of Past and Present Research. *Acta Acustica united with Acustica*, 91(3):409–420, 2005.