A Model of Head-Related Transfer Functions based on a State-Space Analysis

by

Norman Herkamp Adams

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2008

Doctoral Committee:

Associate Professor Gregory H. Wakefield, Chairperson
Professor David R. Dowling
Professor David L. Neuhoff
Professor Mary H. Simoni

For Elena, and my parents

ACKNOWLEDGEMENTS

During my time as a Ph.D. student at the University of Michigan I have had the privilege of being taught and mentored by many remarkable and talented people. I would like to acknowledge several individuals in particular.

The four faculty that served on my committee have influenced me and my research greatly. Each member has brought their own expertise, and the dissertation would have been weaker without their guidance. Prof. David Dowling taught me linear acoustics with great clarity and thoroughness, and gave me the opportunity to review papers for JASA. Prof. David Neuhoff is also a remarkably clear and complete instructor, whom I learned source coding and information theory from. I also had the pleasure of GSI'ing and grading for Prof. Neuhoff.

Prof. Mary Simoni has advised several projects that I have participated in, from a composition included in the first 60×60 concert in NYC, to time-frequency visualization and analysis of musical signals. She also helped me win a CARAT/Rackham fellowship. Under Prof. Simoni's guidance I had the pleasure of authoring a chapter for an edited volume.

And, of course, my dissertation would not have been possible without Prof. Greg Wakefield, who has provided excellent advice and encouragement throughout. Together we have published several papers in MIR and binaural display. Prof. Wakefield has also helped me win two fellowships, a Rackham Pre-Doctoral fellowship and an AFCEA fellowship.

I would also like to recognize Prof. Jason Corey for his consultations on real-time room auralization, and Dr. Tom Santoro for his assistance with HRTF measurement.

There are many other people and places that I am indebted to for enabling this dissertation. None more than my wife, Dr. Elena Adams, who has supported and encouraged me in spite of my relative taking-of-the-scenic-routeness. And, finally, I never would have made it to U.Mich. without my parents, Norman and Mary-Ellen, as well as my brother, Dan. To all of you, thank you.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

GLOSSARY

Symbol	Description
Δ	Path length difference between the two ears
$\mathbf{\Sigma}$	Full-order (N_0) state-space system, $\mathbf{\Sigma} = \big(\mathbf{A}, \mathbf{B}, \mathbf{C}\big)$
$\widehat{\mathbf{\Sigma}}$	Low-order $(N \ll N_0)$ state-space system, $\widehat{\mathbf{\Sigma}} = \big(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}\big)$
α	Radius of the head
θ	Azimuth angle: $-180° \leq \theta < 180°$ ($\theta < 0$: left, $\theta > 0$: right)
σ_k	The k^{th} largest singular value of a matrix
τ	Interaural time difference (ITD)
ϕ	Elevation angle: $-90° \leq \phi \leq 90°$ ($\phi < 0$: below, $\phi > 0$: above)
ω	Radian frequency, $\omega = 2\pi f$

Symbol	Description
$\mathbf{A}$	Full-order state-update matrix ($N_0 \times N_0$)
$\widehat{\mathbf{A}}$	Low-order state-update matrix ($N \times N$)
$\mathbf{B}$	Full-order input matrix ($N_0 \times M$)
$\widehat{\mathbf{B}}$	Low-order input matrix ($N \times M$)
$B_L(\omega), B_R(\omega)$	Acoustic signals received at the ears (freq.-domain)
$\mathbf{C}$	Full-order output matrix ($P \times N_0$)
$\widehat{\mathbf{C}}$	Low-order output matrix ($P \times N$)
C	Computational cost, # of multiplies per sample period
$\mathcal{H}$	Hankel operator/matrix
$H_L(\omega \,\|\, \theta, \phi)$	Left HRTF for direction (θ, ϕ)
$H_R(\omega \,\|\, \theta, \phi)$	Right HRTF for direction (θ, ϕ)
$\mathcal{L}$	Convolution operator/matrix
M	Number of input channels of MIMO system
N	Order of a state-space approximant, $N \ll N_0$
N_0	Order of a full-order state-space model of HRTFs
P	Number of output channels of MIMO system
$S(\omega)$	Acoustic signal of incoming plane wave (freq.-domain)
$\mathbf{X}$	Matrix ($p \times m$) with rank N_0
$\widehat{\mathbf{X}}$	Matrix ($p \times m$) with rank $N < N_0$

Symbol	Description
$b_L(t), b_R(t)$	Acoustic signals received at the ears (time-domain)
c	Speed of sound, 343 m/s
$\mathbf{h}[n]$	Impulse response matrix ($P \times M$) at time n
$h_L(t \,\|\, \theta, \phi)$	Left HRIR for direction (θ, ϕ), (continuous-time)
$h_d^L[n]$	Left HRIR for direction $1 \leq d \leq D$, (discrete-time)
$h_R(t \,\|\, \theta, \phi)$	Right HRIR for direction (θ, ϕ), (continuous-time)
$h_d^R[n]$	Right HRIR for direction $1 \leq d \leq D$, (discrete-time)
n	Discrete time (sample number)
r	Distance of the source from the listener
$s(t)$	Acoustic signal of far-field point source (time-domain)
t	Continuous time (seconds)
$\mathbf{u}[n]$	Input vector (length M) at time n
$\mathbf{x}[n]$	State vector (length N) at time n
$\mathbf{y}[n]$	State vector (length P) at time n

Acronym	Description
BMT	Balanced Model Truncation (order-reduction method)
HOA	Hankel-norm Optimal Approximation (order-reduction)
HRIR	Head-Related Impulse Response (time-domain HRTF)
HRTF	Head-Related Transfer Function
ILD	Interaural Level Difference
IPD	Interaural Phase Difference (integral of ITD w.r.t. freq.)
ITD	Interaural Time Difference
LTI	Linear Time-Invariant
MIMO	Multiple-Input Multiple-Output
MIPS	Millions of Instructions per Second
MISO	Multiple-Input Single-Output
PCA	Principal Component Analysis
SIMO	Single-Input Multiple-Output
SISO	Single-Input Single-Output
SVD	Singular Value Decomposition
VBAP	Vector-Based Amplitude Panning

ABSTRACT

This dissertation develops and validates a novel state-space method for binaural auditory display. Binaural displays seek to immerse a listener in a 3D virtual auditory scene with a pair of headphones. The challenge for any binaural display is to compute the two signals to supply to the headphones. The present work considers a general framework capable of synthesizing a wide variety of auditory scenes. The framework models collections of head-related transfer functions (HRTFs) simultaneously. This framework improves the flexibility of contemporary displays, but it also compounds the steep computational cost of the display. The cost is reduced dramatically by formulating the collection of HRTFs in the state-space and employing order-reduction techniques to design efficient approximants. Order-reduction techniques based on the Hankel-operator are found to yield accurate low-cost approximants. However, the inter-aural time difference (ITD) of the HRTFs degrades the time-domain response of the approximants. Fortunately, this problem can be circumvented by employing a state-space architecture that allows the ITD to be modeled outside of the state-space. Accordingly, three state-space architectures are considered. Overall, a multiple-input, single-output (MISO) architecture yields the best compromise between performance and flexibility. The state-space approximants are evaluated both empirically and psychoacoustically. An array of truncated FIR filters is used as a pragmatic reference system for comparison. For a fixed cost bound, the state-space systems yield lower approximation error than FIR arrays for $D > 10$, where D is the number of directions in the HRTF collection. A series of headphone listening tests are also performed to validate the state-space approach, and to estimate the minimum order N of indiscriminable approximants. For $D = 50$, the state-space systems yield order thresholds less than half those of the FIR arrays. Depending upon the stimulus uncertainty, a minimum state-space order of $7 \leq N \leq 23$ appears to be adequate. In conclusion, the proposed state-space method enables a more flexible and immersive binaural display with low computational cost.

CHAPTER I

Introduction

The human sense of hearing, audition, provides a remarkable degree of spatial sensitivity. Using only two acoustic sensors, our left and right ears, we perceive a rich three-dimensional auditory environment. We intuitively localize simultaneous sound sources, even in noisy and chaotic listening conditions. Furthermore, we can often perceive not only the location of a sound source, but also the spatial extent of the source and characteristics of the surrounding environment, or enclosure. That so much spatial information is perceived without the aid of a large array of acoustic sensors raises many questions about the human auditory system, as well as presents intriguing possibilities for virtual auditory display.

The two acoustic signals that we observe are sufficient input for us to construct the spatially rich auditory scene that we perceive. Therefore, the percept of an arbitrary auditory scene can be created by synthesizing two acoustic signals and presenting those signals over headphones. The difficulty, of course, is in creating the two signals so that the correct spatial cues are perceived by the listener. Researchers from across psychoacoustics, signal processing, and music composition have explored this question. Burgeoning applications, from video games to sonar to sonification,

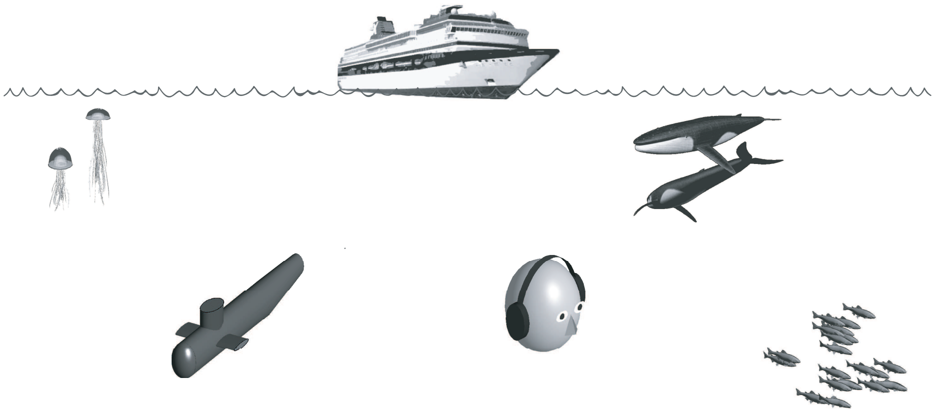


Figure 1.1: A sample virtual auditory scene for a sonar operator using a binaural display. The sonar operator is able to simultaneously hear and localize the innocuous schools of fish and cruise ship, as well as the nefarious submarine passing behind.

have further motivated research on how to display a *virtual auditory scene*[1] (VAS) to a listener. For example, submarine sonar operators currently must steer a hydrophone array and listen, over headphones, to individual sources. A sonar system that displays all sources simultaneously, with appropriate spatial cues, would allow the sonar operator to listen to the entire underwater scene in realtime. A sample sonar auditory scene is portrayed in Figure 1.1. A headphone system for the synthesis and display of VAS is known as a *binaural auditory display*, or simply a binaural display.

This dissertation considers a general framework for a flexible binaural display system. We describe a system that can display complex and dynamic virtual auditory scenes, scenes that contain multiple moving sources, spatially-extended sources, acoustic reflections, and listener motion. The complete binaural display has several

[1] An *auditory scene* is the aggregate percept experienced by a listener due to the acoustic waves at the listener's ears. A virtual auditory scene is a scene that is not the result of a physical acoustic scene, but rather is displayed to the listener using headphones.

inputs and two outputs. The inputs to the system are a collection of (monaural) source signals, the time varying positions of the sources, the time-varying position of the listener, and information about the acoustics of the listener and the environment. The system outputs the binaural signal to display over headphones. The framework is based on a simple 'ray-acoustics' model of wave propagation. This model is often used to predict the collection of wavefronts that impinge upon a listener due to a sound source in an enclosure. The focus of this dissertation is the subsystem after the ray-acoustics model has been applied: a collection of wavefronts due to multiple moving sources in an enclosure are impinging upon a listener, and the final binaural signal at the listener's two ears is computed from this collection.

The relationship between a plane wave impinging upon a listener, and the two waves received at the ears is described by the *Head-Related Transfer Function* (HRTF) for that listener and direction (1). HRTFs form the cornerstone of contemporary binaural displays, and efficient modeling and implementation of HRTFs has been the subject of much research throughout the last two decades. Traditional binaural displays model individual HRTFs, and hence can only display a single stationary far field source in free space. Such primitive auditory scenes are rarely experienced in everyday life, and binaural displays designed in this way often yield virtual auditory scenes that lack *presence*[2]. The more flexible binaural display described in this dissertation can display a wide variety of auditory scenes with an improved sense of presence.

In the present work we develop and evaluate a novel model for a *collection* of HRTFs. That is, rather than filter a source signal with a single HRTF for display,

[2] Presence is the sense that a percept originates from physical space. Percepts that result from physical acoustic scenes usually exhibit presence, whereas percepts that result from virtual scenes often exhibit little presence.

filter the signal with a multiple of HRTFs. This approach has recently been explored in several studies on the binaural display of complex auditory scenes that include reflective environments (2; 3; 4), source or listener motion (5; 6), or spatially-extended sources (7). Remarkably, while this approach has been adopted for the display of a variety of complex auditory scenes, few authors have noted that this approach provides a simple framework for modeling all of the phenomena described above.

The primary contribution of this dissertation is an efficient state-space model of a collection of HRTFs. The similarity exhibited between HRTFs at different directions allows us to accurately model multiple HRTFs with a single low-order state-space system. There is growing interest in state-space models of collections of HRTFs. Three recent studies (8; 9; 10) describe state-space systems that accurately model a collection of HRTFs. These studies are reviewed in Section 2.1.2. None of these studies develop state-space models that are simultaneously accurate and low-order in any specific sense, however. In the present work, we propose methods for constructing state-space models of collections of HRTFs that are both accurate and low-order. We find that order-reduction techniques designed for applications in robust control are well-suited to the HRTF approximation problem. Through a series of empirical and psychoacoustic studies, we demonstrate that the state-space approach yields significantly lower net computational cost than an array of independent filters. Overall, we find that a perceptually indiscriminable state-space model yields a computational savings of nearly two orders of magnitude relative to a full-order HRTF implementation.

The dissertation is organized as follows. The next chapter introduces the state-space formulation of HRTFs, and describes the order-reduction techniques used to find efficient approximants. The order-reduction methods are based on the Hankel operator. This operator is developed and compared to the convolution operator. One aspect of HRTFs is found to be poorly modeled by low-order state-space systems; the *interaural time delay* (ITD). However, this problem can be circumvented by considering alternative state-space architectures. Section 2.3 considers three architectures, and also describes other (unsuccessful) attempts at modeling ITD in state-space. In Chapter III the performance of the state-space approximants is characterized empirically. Both Hankel and $\mathcal{L}^\infty$ approximation errors are reported, as well as a common perceptual audio error measure. In particular, we demonstrate that state-space systems outperform arrays of FIR filters of equal net cost. Chapter IV then reports on a series of headphone listening experiments using state-space approximants of HRTFs. These experiments estimate the minimum approximant order such that a listener cannot tell the difference between an approximate rendering and an ideal rendering. For comparison, an array of truncated FIR filters is included in both the empirical and the psychoacoustic experiments.

The remainder of this chapter develops the fundamentals of spatial hearing and binaural display, as well as motivates the proposed framework. The next section describes the classical duplex theory of spatial hearing. Section 1.2 introduces the HRTF and describes the measured HRTF dataset used throughout the dissertation, as well as discusses the individuality of measured HRTFs. Section 1.3 then describes three common models for HRTFs, and Section 1.4 describes the architecture of conventional binaural displays. The fifth section motivates the framework that is considered in the subsequent chapters. In particular, Section 1.5.2 addresses acoustic reflections, and how such reflections can be modeled by a binaural display. The final section summarizes the chapter and distills the myriad of motivations into a single hypothesis for the dissertation.

For completeness, a preliminary study on the binaural display of ‘clouds’ of point sources is also included. In this study, a two-stage IIR filter implementation is proposed to efficiently model small collections of nearby point sources. This study is described in Appendix A.

1.1 Duplex Theory

Throughout most of the last century, research in spatial hearing focused on what is known as the “duplex theory” of sound localization, a theory derived from the pioneering work of Lord Rayleigh (11). The duplex theory of sound localization is predicated on the assumption that the principal cues of a sound’s location are identified as the *differences* between the sound field at each ear (12). This assumption is based on the fact that “the main difference between the two ears is that they are not in the same place”(13). However, as will be described below, there are significant limitations to only modelling the difference between the acoustic signals received at each ear.

The classical formulation of duplex theory further simplifies the model of human sound localization by assuming the acoustic signals received at each ear are identical outside of a constant time shift and amplitude difference. The time shift is referred to as the interaural time difference (ITD) and the amplitude difference is referred to as the interaural level difference (ILD). Despite its simplicity, this model has demonstrated great explanatory power for sounds on the lateral axis. Binaural signals created in this way[3] yield sound objects located inside or near the listener’s head (14). Specifically, the sound objects are perceived to be located on the line connecting the two ears. While there is no general rule relating localization to lateralization,

[3] A monaural signal that is displayed binaurally with non-zero ITD and ILD

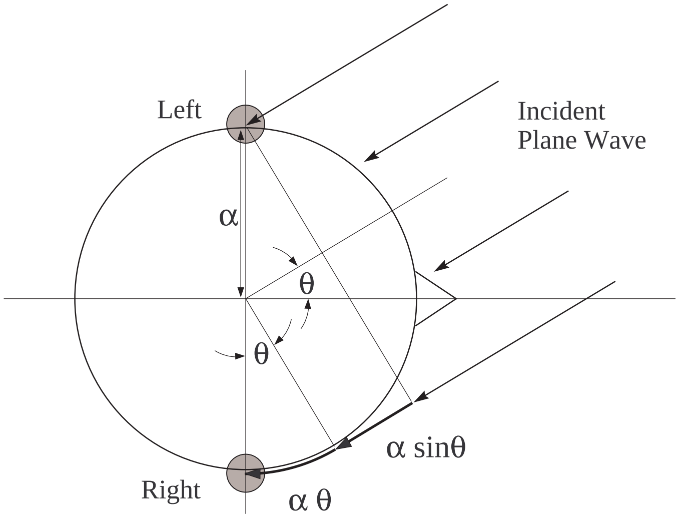


Figure 1.2: A plane wave impinging a spherical approximation of the human head.

many of the conclusions drawn from duplex theory lateralization experiments can be generalized to spatial hearing in three dimensions (1).

Because the two ears are separated by an acoustically significant distance, a sound source that is located off the median plane[4] yields different path lengths to each ear (15). The resulting time difference between when the wave front arrives at each ear is the ITD. For periodic signals, the ITD gives rise to a phase difference between the two ears, known as the interaural phase difference (IPD). For a monaural signal presented binaurally with non-zero ITD, the sound object is perceived on the side of the median plane containing the ear that receives the first wavefront. By modeling the head as a rigid sphere with two point receivers on opposite sides of the sphere, we find the following far field expression for the difference in path lengths, as show in figure 1.2 (12),

$$\Delta = \alpha\,(\,\theta + sin\,\theta\,) \tag{1.1}$$

[4] The median plane is the vertical plane that bisects the line connecting the two ears.

where α is the radius of the head in meters, and θ is the angle of the sound source from the median plane in radians. Note that this expression only holds for sound sources in the far field; at a distance of at least 10α from the head. In this case the sound field incident upon the head is approximately planar. Assuming the speed of sound is constant across frequency, the ITD is given by

$$\tau = \Delta\, c \tag{1.2}$$

where c is the speed of sound (343 m/s), and the IPD is given by $\tau\omega$, where ω is radian frequency. For an average head radius of $8.75cm$, the maximum possible ITD is about $630\mu s$. Psychoacoustic experiments have shown that for click stimuli, the lateral displacement of a sound object is approximately linearly proportional to the ITD of the binaural sound, with maximum lateral displacement achieved for ITD's of about $1ms$ (1). If the ITD is much larger than $1ms$ the stimuli is perceived as two sound objects (16).

For 'continuous' sounds[5] there are no sharp transients to indicate which ear is leading. Hence the lateral displacement of the sound object must be determined by the phase difference between the two ears. For sinusoids with wavelength greater than twice the diameter of the head (frequencies less than about 1 kHz), the IPD is restricted to $\pm\pi$ and yields an unambiguous ITD. For frequencies greater than 1 kHz there is a spatial aliasing problem and only a fraction of the $\pm\alpha$ lateral displacement can be achieved by varying the IPD from $-\pi$ to π (17). Furthermore, the auditory system seems to be insensitive to interaural phase differences for frequencies above about 1.5 kHz (12). Hence the ITD is primarily a low-frequency cue.

The ILD, on the other hand, is primarily a high-frequency cue. The different path

[5] Sounds without an onset transient or other abrupt changes in the amplitude envelope.

lengths to the two ears yield a slight level difference. More significantly, the acoustic opacity of the head has a pronounced affect on the amplitude for wavelengths smaller than the diameter of the head. Above 3 kHz, the head causes two acoustic effects that vary the pressure at each ear for sources located away from the median plane. On the ipsilateral side[6], the head tends to reflect pressure waves that impinge upon it, increasing the net pressure at the surface. For normal incidence this yields up to a 6 dB increase relative to the free field. Simultaneously, the contralateral side[7] of the head is acoustically shadowed, yielding up to a 20 dB decrease for high frequencies relative to the free field (18). For a monaural signal presented binaurally with non-zero ILD, the sound object is perceived on the side of the median plane containing the ear that receives the stronger wavefront. Pyschoacoustic headphone studies have shown that the lateral displacement varies linearly with ILD for ILD's up to about 15-20 dB, and for ILD's greater than 20 dB the sound object is perceived "all on one side" (1).

While the duplex theory has proven enormously useful in binaural research, and ITD and ILD are firmly established as primary cues used in sound localization, the incompleteness of the theory is easily demonstrated. ITD and ILD cues allow virtual acoustic displays to place a sound object anywhere on the line connecting the two ears. Ideally, we seek a virtual acoustic display that allows us to place virtual sound sources anywhere in 3D space.

Consider the set of points in 3D space that yield a common ITD and ILD pair. If the influence of the head is neglected, and only the separation of the two ears is considered, it is straightforward to show that the set of points that yield a single

[6] The *ipsilateral* ear is the ear that is closer to the sound source.

[7] The *contralateral* ear is the ear that is farther from the sound source.

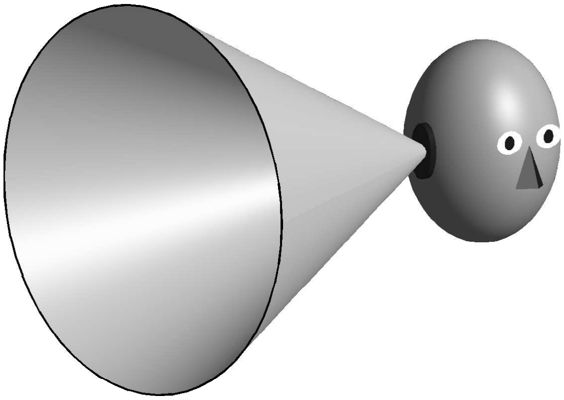


Figure 1.3: Sample “cone of confusion”

ITD/ILD pair define a cone centered on the interaural axis[8] A sample cone is shown in Figure 1.3. This cone is referred to as the “cone of confusion” (13) due to that fact that listeners often make localization errors in which the listener localizes the source within the correct cone, but located at an incorrect position within the cone. The rate at which such errors occur varies considerably between individuals, but is generally small in physical space and relatively large in virtual auditory space (15). A special case of the cone of confusion is the median plane, in which case the acoustic waves received at the two ears are nearly identical. Clearly, only monaural cues can be used to estimate the elevation and distance of a source in the median plane.

Finally, it should be noted that listeners that are profoundly deaf in one ear can still localize sound sources somewhat. In this case, monaural cues alone must be used to localize the sound. Hence it is insufficient to only consider the differences between the two ears signals.

[8] Strictly speaking, the set of points form a cone only in the far field. In the near field the set of points define a 3D hyperbola.

1.2 Head-Related Transfer Functions

Consider a listener in free space and a far field point source, as shown in Figure 1.4. In this case, the acoustic wave is approximately planar by the time is reaches the listener. Due to the acoustic interactions with the listener's head, torso and pinnae, the waves received at the listener's ears are different from the incident plane-wave. The ratio of the amplitude of the received and incident waves are referred to as the head-related transfer function (HRTF) for that angle of incidence relative to the listener. Assuming that air is a uniform isothermal gas, and the magnitude of the pressure fluctuations is not large, then the HRTF can reasonably be modeled as a linear time-invariant (LTI) system that is independent of source distance beyond a scale factor. Accordingly, all source signals for a given direction are related to the signals received at the listener's ears by a pair of LTI filters. Let the source signal be $X(\omega)$, and $Y_L(\omega)$ and $Y_R(\omega)$ are the waves received at the listener's two ears. Then

$$\begin{aligned} Y_L(\omega) &= H_L(\omega \,|\, \theta, \phi)\, X(\omega) \\ Y_R(\omega) &= H_R(\omega \,|\, \theta, \phi)\, X(\omega) \end{aligned} \tag{1.3}$$

where $H_L(\cdot)$ and $H_R(\cdot)$ are the HRTFs for the left and right ears for direction (θ, ϕ), and θ is the source 'azimuth' and ϕ is the source 'elevation' relative to the listener[9]. The time-domain equivalent of the HRTF is the head-related impulse response (HRIR). Eq. (1.3) may be expressed in the time-domain using convolution,

$$\begin{aligned} y_L(t) &= h_L(t \,|\, \theta, \phi) * x(t) \\ y_R(t) &= h_R(t \,|\, \theta, \phi) * x(t) \end{aligned} \tag{1.4}$$

[9] The precise definitions of θ and ϕ depend on the specific coordinate system. Three distinct coordinate systems are commonly found in the spatial hearing literature, the vertical-polar system (19), the lateral-polar system (20), and the double-pole system (21). The vertical-polar coordinate system is the most common, and arguably the most intuitive. This is the coordinate system used throughout the present work. However, the lateral-polar and double-pole coordinate systems have the notable benefit that each azimuth θ defines an interaural cone.

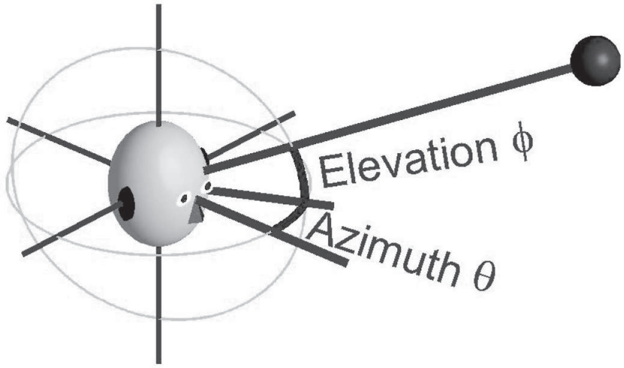


Figure 1.4: A far-field point source at azimuth θ and elevation ϕ.

Note that in (1.3) and (1.4) the source signal is the incident plane wave that impinges on the listener. Alternatively, the source signal can be viewed as the pressure wave emanating from an omnidirectional point source. In this case, the right side of (1.3) and (1.4) must be divided by r, the distance of the source from the listener.

In principle, measuring HRTFs is straightforward. By presenting a listener with signal $X(\omega)$ at direction (θ, ϕ), and recording $Y_{L,R}(\omega)$ at the listener's ears, the HRTFs are given by $H_{L,R}(\omega) = Y_{L,R}(\omega)/X(\omega)$. In practice, accurate measurement of a listener's HRTFs is a painstaking process. Great care must be taken in system identification to avoid loudspeaker and microphone bias. Furthermore, the listener must remain motionless in an anechoic chamber in order to present individual plane wave stimuli.

1.2.1 HRTF Measurements

The HRTFs used throughout the present work were measured using a custom built apparatus at the Naval Submarine Medical Research Laboratory (NSMRL) in Groton, CT (17). The HRTFs of eight individuals, including the author, are used in the present work.

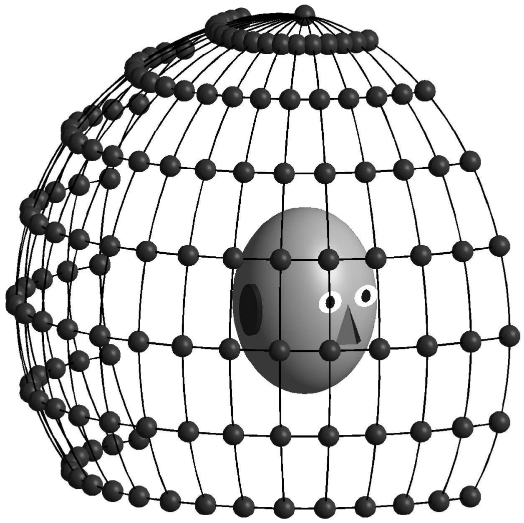


Figure 1.5: Spatial angles for which the HRTF is measured.

The NSMRL facility consists of a cubic anechoic chamber, 10 m on edge, with a desk chair mounted at the center of the chamber. A vertical circular arc, on which 15 small 2-way Realistic speakers are mounted, is centered around the head of the listener. The arc has radius 1.9 m, and subtends 252°. The speakers are uniformly distributed around the arc in 18° elevation increments; from $\phi = -36°$ up to $\phi = 90°$, and back down to $\phi = -36°$ on the far side. The chair is rotated in 10° increments to measure the HRTF at different azimuths around the head. For a given chair angle, seven HRTFs are measured for azimuth θ_0 (from elevation $-36°$ to $72°$), seven HRTFs are measured for azimuth $\theta_0 - 180°$ and one HRTF is measured for $\phi = 90°$. The chair is rotated 18 times, yielding HRTFs measured 360° around the head in azimuth. For each listener 253 unique HRTFs are measured. Figure 1.5 shows the HRTF measurement points for the right-side of the listener. Note that the spatial density of HRTF sample locations is not uniform.

Measurements are made using a pair of in-ear blocked-meatus microphones (22), Knowles FG3329 microphones mounted in children-size earplugs (17). For each chair angle, each speaker is individually driven by a sequence of ten pairs of 512-point Golay codes, which have several desirable properties for acoustic system identification tasks (17). Tucker-Davis Technologies System II AD/DA converters with a 50 kHz sampling rate and 16-bit resolution are used throughout. A total of 512 samples are recorded, although the later half of the measured responses was found to be stationary background noise. The impulse responses are equalized for the speaker and microphone transfer functions. For compatibility with common audio hardware, the HRIRs are resampled at 44.1 kHz and truncated to 256 samples. The complete measurement procedure for an individual takes about 45 minutes.

1.2.2 Individuality of HRTFs

HRTFs are unique to the individual listener. It is often argued that binaural displays should be designed using individualized HRTFs. However, HRTF measurement is a time-, equipment- and subject-intensive process (17; 19), and as such there is considerable interest in using non-individualized HRTFs. The most common non-individualized HRTFs are measured either from simple geometric models (23), or from anatomically average mannequins (24). Such HRTFs may yield reduced externalization and more frequent localization errors, but are common in analytic and numerical studies nonetheless.

In fact, it is unclear if individualized HRTFs are necessary for the display of VAS. Localization of broadband noise is clearly degraded when non-individualized HRTFs are used (25; 26). In particular, front-back reversals and other errors along the cone of confusion are problematic. However, the localization of low-pass signals, such as speech, does not degrade substantially when non-individualized HRTFs are used (2; 27). This is perhaps due to the fact that most of the energy in speech signals is below 4 kHz, and as such the perceived location is dominated by simple ITD and ILD cues, and not by fine spectral detail due to pinna-interations with the acoustic field that are expected at higher frequencies. Finally, other studies have found that the use of non-individualized HRTFs does not 'degrade localization' so much as it simply biases localization judgements (28; 29). For example, a listener might perceive a virtual source at a higher elevation than that of the measured HRTF, but the percept is still compact and well externalized.

Additionally, it has been proposed that localization performance can be improved in some cases by deliberately using non-individualized HRTFs. For example, the use of a 'magnified-head' model has been proposed to improve distance perception in binaural auditory displays (30; 31). Furthermore, listener's that localize sound sources accurately typically have measured HRTFs that are strongly dependent on direction, whereas listeners whose HRTFs show less variation with direction often demonstrate less accurate localization abilities (28). Hence with training, even a listener with relatively poor localization skills in physical space may demonstrate improved localization performance in virtual space constructed with someone else's HRTFs. This idea has motivated numerous binaural displays. One such implementation uses HRTFs measured from several individuals, and then constructs a single representative set by clustering the HRTFs into groups for each direction, and finally selecting the most representative token from each cluster (32). Other binaural displays allow a listener to efficiently search a database of numerous HRTFs to find a good HRTF for each direction (28; 33), although other authors have found such techniques to be unsuccessful (27).

There is currently a great deal of interest in methods for adapting HRTFs measured for one listener to another listeners. A recent study has found that the difference in HRTFs measured from different individuals is largely accounted for by a simple scaling of the frequency axis, and that this scaling is proportional to pinna size (28; 29; 34). This implies a simple procedure for customizing non-individualized HRTFs to new listeners (35). Other techniques involve the decomposition of measured HRTFs into features that vary with individual, such as fine spectral detail, and features that do not, such as ITD and ILD (36; 37). Finally, there is growing interest in coupling HRTF measurements with physical models so as to facilitate direct anthropometric calibration (3; 28; 38; 39).

In the present work, we do not address the issue of individualized HRTFs. In Chapter III the state-space model that we develop is evaluated empirically with HRTFs measured from eight individuals. For each set of listener-dependent HRTFs, we measure the approximation error between the measured HRTFs and the low-order approximant. Chapter IV reports on a listening experiment in which five participants are asked to discriminate between measured HRTFs and low-order approximants. Individualized HRTFs are available for only one of the five participants. For this participant, we show that discrimination performance is not affected by the use of individualized versus non-individualized HRTFs. For the main component of the listening experiment, the discrimination task is performed with non-individualized HRTFs.

1.3 Modeling HRTFs

HRTF datasets are typically large, with hundreds of transfer functions measured for an individual. Furthermore, measured HRTFs typically yield high-order impulse responses, with $\sim 200 - 500$ filter taps. Within this abundance of data, numerous patterns are apparent. Accordingly, there is considerable interest in modeling

HRTFs, both analytically and numerically, so as to identify important perceptual features of the HRTF, as well as reduce the computational cost of HRTF filtering for immersive binaural display applications.

1.3.1 Physical Models

The earliest model of an HRTF approximates the head as a rigid sphere and the source as an omnidirectional point source at distance r from the center of the head (11). The acoustic pressure at any point on the surface of the head, relative to the pressure at that position in free space, is given by (18; 31)

$$\hat{H}(\omega \,|\, \theta, r) = -\frac{rc}{\omega \alpha^2}\, e^{-j\frac{\omega r}{c}} \sum_{m=0}^{\infty} (2m+1) P_m(\cos\theta) \frac{f_m(\frac{\omega r}{c})}{f_m'(\frac{\omega \alpha}{c})} \tag{1.5}$$

where ω is the frequency, θ is the angle of incidence[10], and α is the radius of the head. P_m is the Legendre polynomial of degree m, f_m is the m^{th}-order spherical Hankel function, and f_m' is the derivative of f_m with respect to its argument.

For sources positions beyond arm's reach of the listener, $\hat{H}$ is approximately independent of r (18; 40), and $\hat{H}$ simplifies to

$$\hat{H}(\omega \,|\, \theta, \infty) = -(\frac{c}{\omega \alpha})^2 \sum_{m=0}^{\infty} (2m+1) P_m(\cos\theta) \frac{(-j)^{m-1}}{f_m'(\frac{\omega \alpha}{c})}. \tag{1.6}$$

Figure 1.6 gives $|\hat{H}(\omega \,|\, \theta, \infty)|$ for $0^\circ \leq \theta \leq 180^\circ$ in 20° increments for radius $\alpha = 8.75 cm$. For low frequencies, it is evident that $\hat{H}$ is independent of θ. For frequencies above 200 Hz however, $\hat{H}$ is increasingly direction dependent. The high frequency response generally decreases as θ increases, although there is a pronounced 'bright spot' at $\theta = 180^\circ$ due to the intersecting pressure waves being in phase at that point.

While the spherical model is a reasonable first-order physical approximation, it does not necessarily yield low-order filter implementations. Binaural displays that

[10] The angle between the line connecting the center of the head and the point source, and the line connecting the center of the head and the point of interest on the surface of the sphere

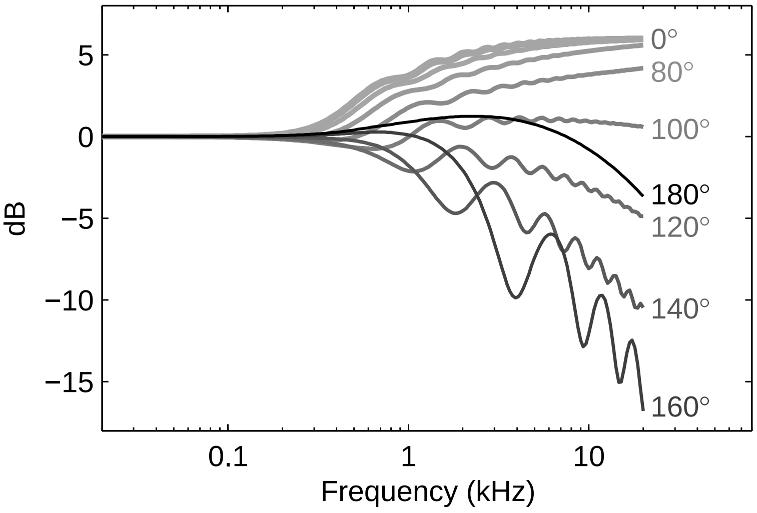


Figure 1.6: Magnitude response for a far-field source impinging on a rigid sphere of radius 8.75cm.

are based on this model typically result in frequent cone of confusion localization errors. In particular, virtual sources located outside the horizontal plane are routinely collapsed to zero elevation.

The simple spherical model has recently been expanded to include the torso and pinnae (23; 41). Furthermore, the complex acoustic characteristics of the pinna have been the focus of much research (42; 43; 44). Numerical acoustics techniques such as the boundary element method (BEM) have also been applied for predicting the HRTF from a detailed computer model of a listener (45).

1.3.2 Statistical Models

Linear subspace projection, such as principal component analysis (PCA), is often used to reduce the dimensionality of HRTF data sets. These methods model the data set as a linear combination of underlying functions (46). PCA has been applied to several HRTF data sets and found that most of the variance in the data set can be accounted for with only a small number of principal components (47; 48; 49; 50; 51).

Kistler and Wightman (48) found that only the first 5 principal components are required for accurate localization performance, implying that higher-order components contribute little to perceived location (48). Recently, other methods of dimensionality reduction have been applied to HRTFs, such as nonlinear manifolds (52).

Often statistical data reduction techniques, both linear and nonlinear, provide little, if any, physical insight. The PCA studies cited above do present a curious physical analogy, however. The physical analogy depends on the specific HRTF representation used in the PCA. Martens (47) and Wu et al. (51) applied PCA directly to measured HRTFs (or equivalently the HRIRs), whereas Kistler and Wightman (48) and Middlebrooks (49) applied PCA to the log magnitude spectrum. The former is equivalent to modeling the transfer function as a collection of parallel LTI systems that are summed, whereas the latter is equivalent to a series cascade of a collection of LTI systems. A parallel LTI model is appropriate for multipath modeling, such as individual reflections from the pinna ridges. In contrast, the series arrangement is appropriate for modeling the transfer functions as a sequence of separate acoustic effects, such as head interactions, followed by pinna interactions, followed by ear-canal interactions (53). Of course, this is only an informal analogy, as the principal components do not have any specific physical meaning. Finally, Blommer (50) applied PCA to the HRTF, treating space rather than frequency as the observation variable. In this case, observations are made at different frequencies, and each each observation is a two-dimension *spatial frequency response surface* (SFRS) (17).

While such techniques reduce the dimensionality of the HRTF data set, they do not reduce the computational complexity of individual HRTF filters. Nonetheless, PCA representations provide a natural space for HRTF interpolation (48), and may yield computationally efficient structures for implementing multiple HRTFs simultaneously. Such implementations would be all-zero (FIR), however. A multi-HRTF pole-zero system would seem to be a more powerful technique for efficiently implementing the spectral detail of the HRTFs. In subsequent chapters we explore one such pole-zero structure, state-space systems, for modeling multiple HRTFs simultaneously.

1.3.3 Minimum-Phase Model

The phase responses of measured HRTFs are nearly minimum-phase[11] if the ITD is neglected (19; 44). This property greatly simplifies the modeling of HRTFs. There are two benefits that the minimum-phase approximation provides for binaural auditory displays. First, it implies that the complex HRTF spectrum is completely specified by its magnitude response; for the magnitude and phase response of a causal minimum-phase system form a Hilbert transform pair (55). Second, a causal minimum-phase system has the minimum energy delay of all causal systems with the same magnitude response $|H(\omega)|$ (55). These two properties allow us to model a left-right pair of measured HRTFs as two minimum-phase systems (specified only by their magnitude response) and a frequency-independent ITD.

The minimum-phase model of the HRTF is used throughout this dissertation, and has become ubiquitous during the last two decades[12]. Several psychoacoustic studies have shown that the human auditory system is not sensitive to monaural phase spectra, and only modestly sensitive to binaural phase spectra (57; 58). Furthermore, it has been found that use of the minimum-phase model of HRTFs is usually indistinguishable from measured HRTFs (48; 59).

[11] Equivalently, nearly all zeros of the measured HRTFs lie inside the unit circle. One recent study found a single prominent non-minimum-phase zero in measured HRTFs for frontal locations, and hypothesized that this zero is due to a strong reflection off the pinna flange (54).

[12] Many binaural displays, including those described in this thesis, neglect the energy delay of minimum-phase HRTFs, which while small compared to the ITD, is not strictly zero (56).

1.3.4 Reduced-Order All-Zero and Pole-Zero Models

Direct implementation of measured HRTFs using convolution is computationally expensive. Consider a left-right pair of measured HRIRs of length 256 at a sampling rate of 44.1 kHz. Implementing a single pair of HRIRs through convolution requires 22.5 million additions and multiplications per second. While this computational load is within the reach of modern computers, once 'real-world' complications such as multiple sources, moving sources or non-anechoic environments are considered, a 22 MIPS starting point is problematic.

A simple and common method for reducing the order of the HRIR is simply to truncate the minimum-phase impulse response. For an HRIR truncated to length $N + 1$, the resulting approximant yields the minimal $\mathcal{L}^2$ error[13] (55). However, it is well established that such an error metric is perceptually ill suited to approximating HRTFs, for it assigns relatively too much weight to high-frequencies and large-amplitude portions of the HRTF (60). Generally speaking, humans are more sensitive to lower frequencies, and for spatial hearing in particular, spectral notches are known to be perceptually important (1). A modified method that uses a perceptual spectral distance measure was proposed in (61). Recently, an efficient FIR technique was proposed that couples wavelet smoothing with sparse FIR filter implementations (62).

IIR filters are often preferable to FIR filters, as freeing the poles from the origin gives the filter design additional degrees of freedom. Numerous IIR filter design techniques have been applied to HRTFs (56; 60; 63; 64). In designing the pole-zero filter, as for the all-zero case, optimizing the conventional $\mathcal{L}^2$ spectral error yields a solution that over-emphasizes high-frequencies and large-amplitudes. To address

[13] The lowest RMSE in both the frequency and time domains.

this problem Blommer and Wakefield (63) employ an $\mathcal{L}^{\infty}$ spectral error measure applied to the log-amplitude spectrum. Alternatively, the use of warped IIR filters, so as to emphasize low-frequency accuracy, have also been employed for constructing perceptually adequate low-order IIR filters (60).

Another IIR filter design technique that is finding increased popularity for HRTF implementations is *balanced model truncation* (BMT) (7; 65; 66; 67). BMT operates by first transforming the measured HRTF into an equivalent state-space representation. The system is then *balanced* and truncated to order N and converted back into transfer function form. BMT yields a solution that is not optimal in any specific sense, but generally exhibits features that appear favorable for HRTF approximation. In the following chapters we explore BMT, as well as another state-space order reduction technique. Unlike the studies cited above, however, we leave the HRTF model in state-space form and consider the net computational cost of a state-space implementation.

1.4 Binaural Display Systems

The binaural display of an auditory scene is predicated on the assumption that "identical stimuli at a listener's eardrum will be perceived identically independent of their physical mode of delivery" (68). Therefore, to immerse a listener in a virtual auditory scene, it is sufficient to compute only the two signals that arrive at the listener's ears. That is, it is not necessary to compute the entire sound field of the acoustic scene. Furthermore, the criteria for the display of a virtual auditory scene is looser than the principle above; it is sufficient to display a binaural signal that is perceptually equivalent, rather than physically identical, to the binaural signal that results from the physical acoustic scene. For example, minimum-phase HRTFs are

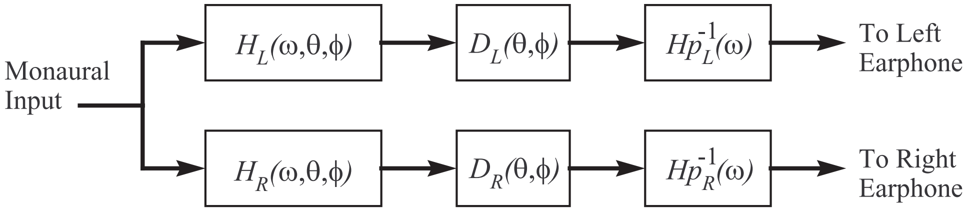


Figure 1.7: Block diagram of a simple binaural auditory display.

often used instead of the measured HRTFs in generating binaural auditory displays.

A primitive auditory scene that consists a single stationary far field point source in free space can be displayed to a listener by filtering the monaural source signal with an appropriate HRTF and presenting the result on a pair of headphones. The block diagram for a typical binaural display is shown in figure 1.7. In this system, $H_{L,R}$ is implemented as a minimum-phase filter, $D_{L,R}$ provides the frequency-independent ITD, and ${Hp_{L,R}}^{-1}$ is the inverse of the direction-independent *headphone-to-ear-canal transfer function* (HpTF), which is discussed below.

1.4.1 Headphone Equalization

The HpTF has been the subject of some debate within the binaural research community. Numerous binaural displays have been designed that account for the HpTF (19; 56; 69; 70), while others make no attempt to equalize for the HpTF at all (3; 71; 72; 73). Because the HpTF is independent of source direction, the contribution of the HpTF to the final signal received at the eardrums is equivalent to filtering the source signal with a fixed transfer function[14], without influencing localization cues. While monaural cues can influence the perception of source location, experimental evidence suggests that the influence of the HpTF on perceived spatial location is negligible (74; 75; 76).

[14] Assuming that the HpTF is identical for both ears.

In practice, accounting for the influence of the HpTF exactly is difficult. The HpTF is dependent on both the headphones and the listener (77; 78); individual measurements for a listener with a specific pair of headphones are required (74; 79). Furthermore, the HpTF has been found to vary when the listener repositions the headphones slightly (80). Overall, accurate compensation for the HpTF is prohibitive, and may not even be beneficial. Accordingly, we have chosen to neglect headphone compensation in the present work.

1.4.2 Interpolation and Moving Sources

HRTFs are measured for a finite number of directions surrounding the listener. As such, it is necessary to perform interpolation if a source is to be rendered at a position for which an HRTF has not been measured. Interpolation is especially important for the display of moving sources, where the sound object must move smoothly through auditory space.

One simple method of interpolation is to estimate the HRTF for the desired location as a linear combination of HRTFs for nearby locations that surround the desired point (81). In spite of its simplicity, the perceptual artifacts of this interpolation method are small, so long as the measured HRTFs are near the desired location (82). Several variants of this method have been proposed (17; 83; 84), but without perceptual validation.

For FIR filter implementations, linearly combining the impulse responses is a stable method for interpolating the HRTFs. For IIR filters however, interpolating and dynamically updating the filter coefficients can lead to instabilities and audible artifacts (56). This is a significant practical drawback to IIR implementations and may large explain why IIR filters are not found in any binaural auditory displays that allow for dynamic updating (3; 4; 56). This problem can be solved by updating the pole/zero locations rather than the filter coefficients, although this method is computationally burdensome (85).

Rather than compute an interpolated HRTF (and use the interpolated HRTF to filter the source), an alternative method is to filter the source with several nearby HRTFs and then take a linear combination of the binaural outputs. This is the binaural analog of *vector based amplitude panning* (VBAP) for multichannel audio (86), and has recently been implemented in several binaural auditory displays (5; 7; 67). This implementation is attractive because it removes the need to dynamically update filter coefficients; only the weights in the linear combination need to change in realtime. However, it also requires that the source signal be filtered with multiple HRTFs, thus compounding the computational load.

1.5 Current Challenges for Binaural Displays

From the first generation of binaural display technology, several critical challenges that limit the technology were evident (1; 19; 87). While progress has been made to address these challenges, none have been solved entirely. This section describes the critical limitations of contemporary binaural displays and motivates the framework we apply to address this problem.

Users of binaural displays often report the following types of localization errors[15]:

1. Loss of presence, in which the sound object is poorly externalized or 'inside the head.'

2. The sound object is not well-focused, in which case it yields ambiguous or

[15] We note the ambiguity in the interpretation of 'error' in this case, as there is no one-to-one mapping between the location of a sound source in physical space and the location of a sound image in perceptual space. See, for example (1; 28).

diffuse directional cues.

3. Front-back reversals, in which the sound object location is mirrored across the vertical binaural plane.

4. Compressed elevation, in which the perceived location of the sound object is pulled toward the horizontal plane. This often occurs with virtual sources located below the listener.

All of these problems are exacerbated when non-individualized HRTFs are used, but persist for many listeners even with the use of individualized HRTFs and proper headphone equalization.

Of the problems listed above, the first is perhaps the most pervasive and the least understood. Many authors attribute the loss of presence in binaural displays to the reliance on primitive auditory scenes, scenes that are stationary and in free space (88). We rarely experience such primitive auditory scenes. Listeners regularly move their heads, if only slightly, to localize sound sources. Furthermore, listeners rarely experience free-space conditions, as there is almost always at least a single reflecting surface: the ground (89).

To address these problems, we propose a binaural display framework that renders each source signal not at a single location, but at a collection of locations. The remainder of this chapter motivates this framework, and describes methods for modeling reflective environments using this framework.

1.5.1 Motivation for Modeling Collections of HRTFs

Many of the limitations of contemporary binaural displays stem from their reliance on individual HRTF filters. A single HRTF pair models a primitive auditory scene. Given that we rarely experience such primitive auditory scenes in everyday listening, it is not surprising that binaural displays based on this model yield 'less-than-convincing' virtual experiences. In contrast, a collection of HRTF pairs may be used to model more 'realistic' scenes. We propose a framework for modeling auditory scenes that include multiple moving sources in a reflective environment with a collection of stationary sources in free-space. In practice, this framework requires that each monaural source signal be filtered with multiple HRTFs simultaneously.

Rendering a monaural signal at many directions simultaneously provides several advantages for binaural auditory displays. It can be used to display sources with spatial-extent. For example, consider a passing flock of geese. The spatial extent of the flock is an important aspect of the perceived sound object. Filtering a monaural recording of squawking geese with a single HRTF pair collapses the sound object into a point source. In order to maintain the spatial extent of the flock we should render the squawks of each goose with a separate HRTF[16].

Modeling multiple source directions simultaneously may alleviate problems associated with the individuality of measured HRTFs as well. A monaural signal that is displayed at a collection of nearby directions instead of a single direction may give the listener additional spatial cues. If the HRTFs are not individualized to the listener, the spatial cues of a single HRTF may be distorted or unfocused, whereas the cues from a collection of HRTFs may reinforce each other and improve the spatial percept. Furthermore, the relative change in the responses between nearby directions are less dependent on the listener than the absolute response (17; 32).

Dynamic binaural displays that employ a head-tracking device coupled with a time-varying HRTF filter, are becoming prevalent. Head movement is an important

[16] Here we are assuming that the squawks of each goose are essentially identical. In this case, a single squawk, or collection of squawks, is filtered with multiple HRTFs, and then the binaural squawks are combined according to a poisson process to synthesize the flock (7).

aid for resolving front-back confusions and other acoustic localization errors (90). Recent studies have found that dynamic cues strengthen the sense of presence and source externalization, and greatly improve localization accuracy of binaural displays (2; 91; 92; 93). In particular, head motion has been found to improve localization accuracy in the median plane. In the static case, only relatively unreliable monaural cues can be used for localization, whereas in the dynamic case, the rate of change in ITD/ILD and other binaural cues can be used for localization (94).

Real-time updating of HRTF filters for a time-varying system presents a formidable challenge. Transient clicks and other spurious phenomena common with adaptive filters are easily perceived by the auditory system. Rather than dynamically updating the HRTF filters themselves, a popular alternative is to render a monaural signal for several spatial locations, and then take a time-varying linear combination of the binaural signals to accomplish source motion (5; 7; 67).

While the aforementioned advantages are significant, perhaps the most intriguing advantage of modeling multiple HRTFs is that acoustic reflections can then be easily included in the virtual auditory scene. In this case, the binaural display is no longer limited to modeling anechoic environments. This idea is discussed in greater detail below.

1.5.2 Reflections and Reverberation

We rarely experience anechoic environments. Indeed, anechoic chambers are surprisingly uncomfortable for human subjects largely because they lack the acoustic reflections that the auditory system expects. The lack of acoustic reflections in virtual auditory space is often pointed to as the primary reason for the poor externalization of many binaural displays (2; 95; 88). This has motivated recent interest in *binaural* *environment modeling*, and a new generation of binaural displays that incorporate acoustic reflections and reverberation (3; 4; 56; 96).

Classically, the study of *auralization* is a branch of architectural acoustics concerned with modeling acoustic fields in an enclosure, such as a concert hall (97). The end result of such modeling is usually a monaural impulse response at the fixed listening location. For an ideal omnidirectional point source and receiver in a rectangular enclosure with infinitely rigid walls, the impulse response between the source and receiver can be solved analytically (98). Unfortunately, loosening any of the restrictions concerning the source, receiver or enclosure, yields a problem that is analytically intractable. However, with the advent of numerical acoustics, numerous room simulators have been proposed. Until recently, room simulators were restricted to monaural auralization, although binaural auralization is now an active area of research as well.

Acoustic room models can be divided into two categories, 'wave'-based models and 'particle'-based models (4). Wave-based models, such as the waveguide mesh and the finite element method (FEM), construct a dense network of nodes through which acoustic waves pass (99; 100). Such methods require that the density of the mesh be small compared to the wavelengths under consideration, hence wave-based results are only valid for low frequencies[17]. Furthermore, wave-based methods are computationally very expensive, and are not appropriate for real-time applications.

Particle-based models, such as the image-source method (101; 102) and the ray-tracing method (98; 103), are based on the assumption of specular reflection. This assumption is only valid for wavelengths that are small compared to the dimensions

[17] This is especially troublesome for binaural displays, where frequencies up to 15 kHz are critical to the perception of elevation. Accurate simulation up to 15 kHz would require a very dense network of nodes, with separation on the order of ~ 1 mm.

of the enclosure ($\gtrsim$ 200 Hz for a medium sized room), and large compared to the roughness of the enclosure walls. Another drawback of these methods is that they are not well-suited to the incorporation of diffusion (4; 104). Despite the limitations of assuming specular reflection, it is used in many contemporary artificial reverberators (104; 105; 106; 107; 108). For binaural displays, the image-source method is attractive because the direction of each reflection relative to the listener is easily determined from the image source position.

Consider a point source in a rectangular enclosure with rigid walls. If each wall is considered independently, and is assumed to be infinite in extent and hardness, then the wall can be modeled by mirroring the source across the wall and then removing the wall. Repeating this process for all six walls yields the six first-order image sources, as shown in Fig. 1.8. These image sources must themselves be mirrored across the other walls, and this recursive process is repeated ad infinitum. The temporal density of reflections increases quadratically with time, while the strength of each reflection decreases quadratically with time (98; 108). The modeled impulse response, when plotted versus time, appears as a sequence of pulses that are initially strong and sparse (separated by several ms), and gradually become weaker and more densely packed, and eventually overlap. For more complex enclosures, modeling becomes difficult; it is necessary to perform a 'visibility' check, which is dependent on listener location, for each image source[18] (102; 109).

The individual pulses in the early part of the impulse response ($\lesssim$ 100 ms for a medium-sized room) are termed *early reflections* whereas the later portion of the response is termed *diffuse reverberation.* The wavefronts of the early reflections impinge upon the listener from a specific direction. In contrast, the later portion

[18] Strictly speaking, even for rectangular enclosures, not all image sources are 'visible'. However, in this case it can be shown that exactly one image source in each mirrored enclosure is 'visible.'

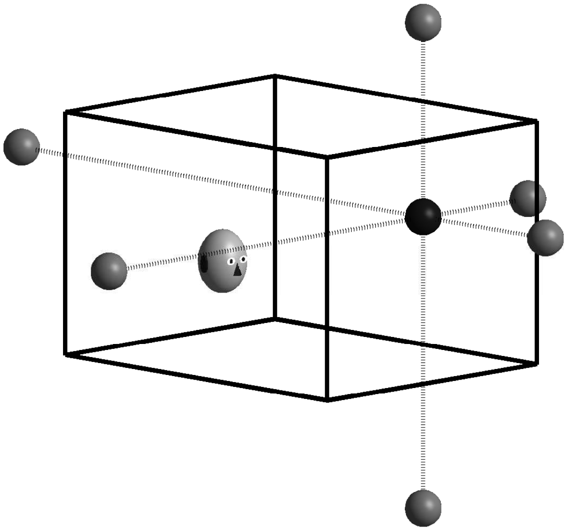


Figure 1.8: The first-order image sources of a rectangular enclosure.

of the reverberation is diffuse, with a high density of weak reflections impinging on the listener from all directions (98). Hence, the early reflections are dependent upon the listener position and orientation, whereas the diffuse reverberation is not. Accordingly, many artificial reverberators employ two subsystems for auralization, treating early reflections and diffuse reverberation separately (3; 4; 56; 104; 105; 106).

For binaural displays, there is no consensus on what features of early reflections are perceptually necessary. Diffuse reverberation is less controversial, and often modeled as an all-pass filter (105). The handling of early reflections varies considerably. For real-time systems, complete physical modeling of early reflections is beyond the reach of modern computing technology. As such, it is common to employ perceptual techniques that model only the perceptually salient aspects of the reflections.

Modeling early reflections is complicated by a collection of perceptual phenomena known as the *precedence effect* (1; 16). Briefly stated, when a listener experiences a direct wave-front followed by multiple reflections (i.e. delayed wave-fronts similar to the direct wave-front), the direct wave-front dominates many aspects of perception.

In particular, the precedence effect accounts for the fact the we perceive the location of a sound source largely from cues derived from the first wave-front. For click stimuli, reflections that follow within 5 ms of the direct wave-front are perceptually fused with the direct wave-front, and later arriving reflections are 'grouped' with the direct wave-front and inhibited[19] (16). However, while reflections do not corrupt the perception of source direction, they do contribute to the overall percept of the auditory scene.

Early reflections are known to influence our perception of timbre; concert halls are judged, in part, on their pattern of early reflections (110). Although our perception of source direction is usually dominated by the direction of the first wave-front, this does not mean that the reflected wave-fronts play no role in spatial hearing. In particular, early reflections and reverberation are known to be important to the sense of presence and source distance (1; 95). For example, many binaural displays control source distance by the ratio of direct to reverberant energy (3; 4).

Several studies have addressed human perception of early reflections (56; 111; 112; 113; 114; 115). These studies focused on the audibility of individual reflections. The studies estimated the minimum intensity of a reflection with fixed delay such that a listener can reliably detect whether or not the reflection is present. Few psychophysical studies have explored how early reflections contribute to the spatial hearing. For binaural display applications, there is evidence that lateral reflections aid localization, especially for sources located in the median plane (95; 111).

Numerous binaural displays have been proposed that incorporate early reflections (3; 4; 56; 72; 95), although no studies exist comparing the efficacy of the different methods. Jot (56) performed an informal listening experiment and concluded

[19] Assuming the time delay is not so large that the reflection is perceived as a distinct echo

that only the lateral component of the direction of each reflection is perceived, and proposed a binaural display that renders early reflections by giving them a direction-dependent ITD and ILD, and filtering them with a direction-independent average HRTF. In contrast, Begault proposed a binaural display in which early reflections are computed using a ray-tracing algorithm and individually filtered with the appropriate HRTF (95; 2). Recently, Hacihabiboglu proposed a binaural display in which the direct wave-front is rendered using a high-order HRTF filter, but the length of the filter is shortened as reflections arrive at the listener. Low-order filters are used for the reflections, yielding a total computational complexity that is constant as more reflections are received (72). All of these methods share the same framework: a monaural signal is rendered by filtering the source with multiple HRTFs, and then scaling and delaying each binaural signal to model the appropriate enclosure.

This method of synthesizing directional early reflections is implemented as a part of the dissertation. In particular, Chapter IV includes stimuli that model a listener in a rectangular enclosure. Only the early reflections are included however, as diffuse reverberation is difficult to incorporate into a system based upon ray-acoustics.

1.5.3 General-Purpose Binaural Display

The binaural display shown in Figure 1.7 can only display a rudimentary auditory scene. We are now ready to consider a general-purpose binaural display, a display that can accommodate multiple sources, spatially-extended sources, acoustic reflections, and source and listener motion. A simple block diagram showing the inputs and outputs for this binaural display is shown in Figure 1.9. In the next chapter we divide this system into two parts, one component that incorporates the dynamic updates and reflections, and another component that filters the final collection of

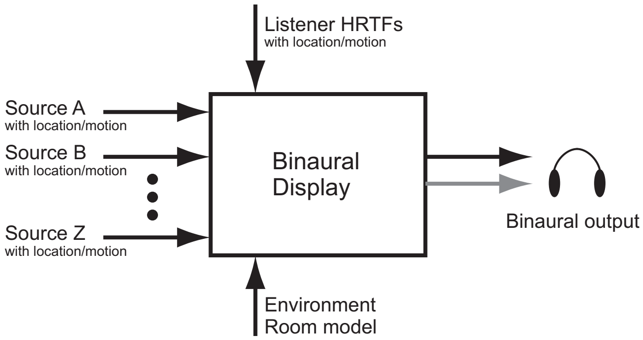


Figure 1.9: A diagram showing the inputs and outputs for the general-purpose binaural display considered in this dissertation.

acoustic rays with direction-appropriate HRTFs. The focus of the dissertation is the latter component.

1.6 Summary

This chapter reviewed the relevant literature on binaural display technology, and described the challenges that motivate the methods presented in the remainder of the dissertation. Classical duplex theory was reviewed and the HRTF was defined. Section 1.3 described several approaches to modeling HRTFs. In particular, section 1.3.4 summarizes research efforts to construct low-order HRTF approximants.

Section 1.5 enumerated some of the key challenges that must be overcome for binaural display technology to be broadly adopted. We have proposed a general framework for addressing many of these challenges: modeling collections of HRTFs surrounding the listener, rather than individual HRTFs. This idea has been previously proposed for specific applications, such as modeling acoustic reflections. However, no studies have observed that this framework can model reflective environments, source and listener motion, and spatially-extended sources, simultaneously. Hence the proposed framework is quite flexible, albeit computationally expensive.

The following chapters describe efficient implementations of this framework based on multiple-input multiple-output (MIMO) state-space systems. In the next chapter we formulate the HRTF in the state-space, describe order reduction methods based on the Hankel operator, and consider the affect of the ITD on the order reduced approximants.

CHAPTER II

Modeling Collections of HRTFs in State-Space

Emerging applications in binaural display present a unique opportunity for the design of efficient state-space systems. For traditional single-input single-output (SISO) filter design applications, state-space implementations do not provide a computational savings over tapped-delay IIR implementations. However, the previous chapter proposed a method for binaural display in which multiple HRTFs are implemented simultaneously. In this case, a naive filter array is computationally burdensome, as the computational cost scales linearly with the number of transfer functions modeled. Implementing the collection of HRTFs with a single multiple-input multiple-output (MIMO) state-space system may be advantageous.

This chapter describes techniques for constructing efficient and accurate state-space approximants of collections of HRTFs. The remainder of this section describes the synthesis of virtual auditory space using the proposed binaural display framework, and reviews related HRTF models, including three recent studies that consider state-space models of HRTFs.

Section 2.2 formulates the HRTF filter array in the state-space, and describes two order reduction techniques based on the *Hankel operator*. As the Hankel-operator is not traditionally used to analyze audio filters, we relate this operator to the more common convolution operator. The two order reduction methods are closely related, although one is *ad hoc* and relatively simple to implement whereas the other is optimal in the Hankel-norm sense but is somewhat more complicated to implement. The two methods have been compared extensively for SISO filters, but relatively few comparisons have been made for the MIMO case. The author is not aware of any studies that compare the two methods for systems as large as those required for binaural display. Both methods are adapted to HRTF modeling and complete algorithms are given in appendix B.

Section 2.3 explores a perceptually important property of HRTFs, the *interaural time delay* (ITD). The ITD between left-right HRTFs pairs is found to limit the potential for finding state-space approximants that are both low-cost and accurate. This limitation is addressed with a variety of techniques, including time-delay state-space systems and hybrid state-space/FIR systems. However, the most effective means of addressing this limitation is found to be the use of alternative state-space architectures, in which case the ITD does not need to be modeled by the state-space system itself. Three state-space architectures are considered. The performance of the state-space systems, using both order reduction techniques and all three architectures is them characterized with an empirical experiment in the next chapter, and then a headphone listening experiment in chapter IV. The next section provides background and related research.

2.1 Background

Before formulating the HRTF filter array in the state-space, we first describe how such an array might fit into a complete binaural display system. Few studies consider efficient implementations of such HRTF arrays. However, numerous stud-

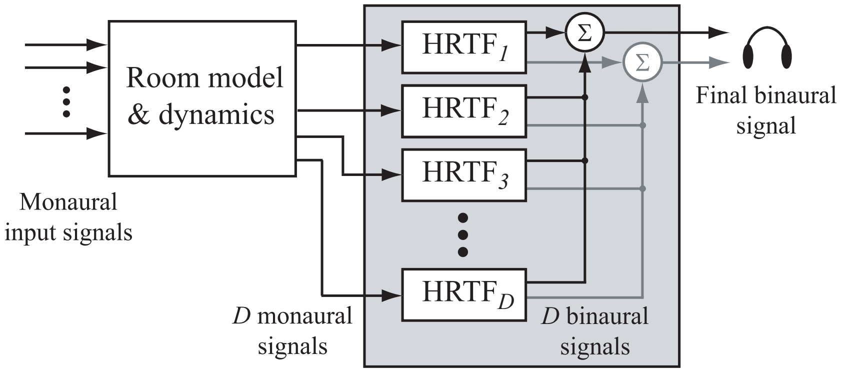


Figure 2.1: A naive implementation of the proposed binaural display framework.

ies have considered computationally efficient implementations of individual HRTFs, while other studies describe models for collections of HRTFs without considering computational cost. The relevant research is reviewed in this section.

2.1.1 Virtual Auditory Scene Synthesis

The block diagram of a conventional implementation of the proposed binaural display framework is shown in Figure 2.1. With this framework, a complex auditory scene, as heard by the listener, is first modeled as a collection of stationary point sources in free space. This step is performed by the block labeled “Room Model and Dynamics”. After the input signal for each of these ‘intermediate’ sources is computed, the final binaural signal is computed by filtering each intermediate source with the appropriate HRTF and summing the result. This later stage, the HRTF filter array, is the focus of the dissertation. Clearly, the computational cost of the implementation shown in Figure 2.1 scales linearly with D, the number of directions included in the HRTF filter array.

We seek to replace the array of filters shown in the gray box with a single D-input, 2-output state-space system. However, in Section 2.3 it is shown that this arrangement of inputs and outputs is problematic, and two alternative state-space architectures are considered. Accordingly, the methods described in Section 2.2 are valid for state-space systems with any number of inputs and outputs.

2.1.2 HRTF Models

HRTFs measured for different directions are similar. A system that models HRTFs at many directions simultaneously may be able to utilize this similarity to reduce the net cost of the system. Numerous studies have found that collections of HRTFs can be reasonably represented in low dimensional spaces. In (48) it is shown that most of the variance of an HRTF dataset can be accounted for with the first five principal components. However, principal component analysis yields a high-order FIR structure for filter implementation. Filters that do not restrict the system poles to the origin, such as IIR filters, are known to yield substantial cost savings for individual HRTFs (60; 63; 64). We seek an analogous structure that models multiple HRTFs. Furthermore, it has recently been shown that HRTFs can be accurately modeled using pole-zero filters with common pole locations (54). This implies that a collection of HRTFs can be reasonably approximated using a single MIMO state-space system, as the rational transfer functions between each input and output of the system share the same poles.

At first glance, modeling HRTFs with a state-space system may not appear computationally efficient. Any order N state-space system can be converted to an equivalent array of order N IIR filters. For SISO systems an IIR implementation is guaranteed to be lower cost than a equivalent state-space system. Nonetheless, state-space techniques have been used to design low-order IIR filters from high-order FIR filters (7; 65; 66; 67). These studies do not consider MIMO state-space systems however, the computational cost still scales linearly with the number of directions D.

Furthermore, converting filters from state-space form to transfer function form may yield IIR filters that are sensitive to coefficient quantization errors (8). In the present work we avoid this problem by leaving the reduced-order system in state-space form and considering its net computational cost.

Three recent studies (8; 9; 10) propose state-space systems that model HRTFs at multiple directions simultaneously. In (8) MISO systems are designed that model multiple HRTFs for each ear. HRTF redundancy is not fully exploited in this work however, as separate SISO systems are designed for each HRTF individually, and then combined into one large MISO system. In contrast, (10) constructs a MISO system directly from a PCA reconstruction of HRTFs. A MIMO state-space architecture is considered in (9). Low-order systems are designed for a collection of HRTFs in the horizontal plane. It was shown that for sufficiently large system order, the localization performance with a state-space system was similar to the performance with an array of measured HRIRs (9). All three studies employ *balanced model truncation* (BMT) (116) to reduce the order of the state-space model. However, neither study considered the computational advantages of state-space implementations. Below we consider the computational cost of low-order state-space implementations relative to common FIR arrays, and show that system orders well below those found in (9) may be perceptually adequate in many cases.

2.2 State-Space Formulation

Consider a stable, causal, discrete-time MIMO state-space system[1]

$$\begin{aligned}\mathbf{x}[n+1] &= \mathbf{A}\mathbf{x}[n] + \mathbf{B}\mathbf{u}[n] \\ \mathbf{y}[n] &= \mathbf{C}\mathbf{x}[n] \end{aligned} \tag{2.1}$$

where $\mathbf{x}[n]$ is the state vector of size N_0, $\mathbf{u}[n]$ is the input vector of size M, and $\mathbf{y}[n]$ is the output vector of size P. To simplify notation, let $\mathbf{\Sigma} = \big(\mathbf{A}, \mathbf{B}, \mathbf{C}\big)$ represent the state-space system. The matrix impulse response of $\mathbf{\Sigma}$ is

$$\begin{aligned}\mathbf{h}[n] &= \begin{bmatrix} h_{11}[n] & \dots & h_{1M}[n] \\ \vdots & \ddots & \vdots \\ h_{P1}[n] & \dots & h_{PM}[n] \end{bmatrix} \\ &= \begin{cases} \mathbf{C}\mathbf{A}^{n-1}\mathbf{B} & n > 0 \\ \mathbf{0} & n \leq 0 \end{cases}\end{aligned} \tag{2.2}$$

A state-space system can be viewed as a multi-channel filter; the system receives an M-channel input and computes a P-channel output. Block diagrams for a SISO FIR filter, and a MIMO state-space are shown in Figure 2.2. Two equivalent forms of the FIR filter are shown; a traditional tapped delay line form, and a vector multiplication form. The impulse response for the FIR filter is $\mathbf{b} = [b_0, b_1, \cdots b_{N_0}]$. Clearly, the length of the impulse response of the FIR filter can be no longer than $N_0 + 1$ samples. In contrast, the state-space system is a feedback structure. Hence the impulse response can be infinite in length, depending upon state-update matrix $\mathbf{A}$.

[1] For convenience the systems considered here have no feed-through term (the $\mathbf{Du}[n]$ term), similar to (8; 9). The Hankel operator, described below, is not influenced by the $\mathbf{D}$ matrix, hence the choice of $\mathbf{D}$ is somewhat arbitrary for this class of order reduction methods. In the present work we simply set $\mathbf{D} = \mathbf{0}$. The interested reader is referred to (117) for a detailed discussion of this term.

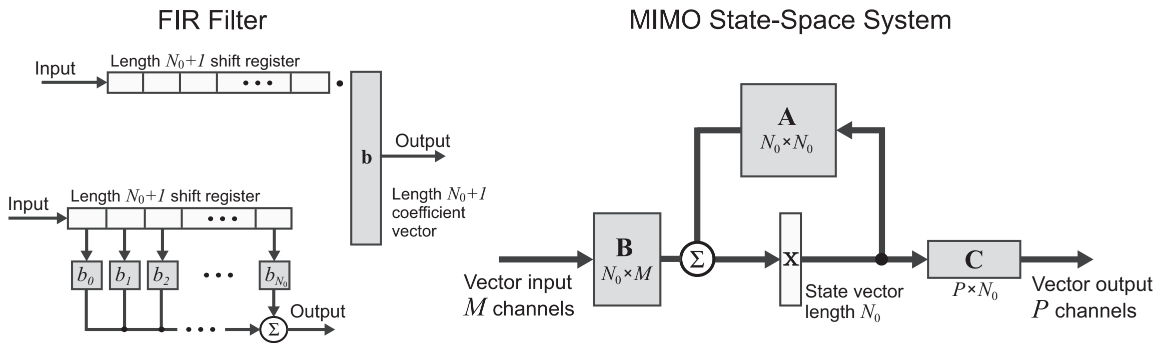


Figure 2.2: Block diagram of an FIR filter (left) and a MIMO state-space system (right). Two equivalent forms of the FIR filter are shown: vector multiplication (top) and tapped delay line (bottom).

Furthermore, for any given state-space system $\mathbf{\Sigma}$ there are an infinite number of distinct systems $\tilde{\mathbf{\Sigma}}$ with the same input-output response. All such state-space systems can be related to each other via a *similarity transform* (118).

The FIR filter in Figure 2.2 requires that the current input sample, as well as N_0 previous input samples, be available for the vector multiplication. In contrast, the state-space system only requires the current input sample for each channel. This length M column vector is multiplied by $\mathbf{B}$ and added to the length N_0 state vector, which is itself updated by $\mathbf{A}$, and also determines the output via $\mathbf{C}$. Note that the length N_0 state vector is the only memory requirement for the entire MIMO state-space system.

It is straightforward to design a state-space system $\mathbf{\Sigma}$ that implements a collection of $2D$ HRIRs exactly[2]. For example, a D-input 2-output system that models the HRTFs in the gray box in Figure 2.1 has the following block impulse response

$$\mathbf{h}[n] = \begin{bmatrix} h_1^L[n] & h_2^L[n] & \dots & h_D^L[n] \\ h_1^R[n] & h_2^R[n] & \dots & h_D^R[n] \end{bmatrix} \tag{2.3}$$

where $h_d^L[n]$ and $h_d^R[n]$ are the HRIRs for the left and right ears for direction d.

[2] For example, the controller canonical form, as described in Appendix B.

However, such a state-space system is high order ($N_0 \approx 500$) and computationally prohibitive. As such, we explore order reduction techniques to design low-cost approximants $\widehat{\mathbf{\Sigma}} = \left(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}\right)$ with order $N \ll N_0$. The two reduction methods that we explore are based on the Hankel operator, which is described next.

2.2.1 Operators and Norms

Relative to SISO systems, there are few methods for reducing the order of MIMO systems such that the resulting low-order approximant is optimal in some sense. One metric for which optimal solutions can be found is the *Hankel norm.* However, interpreting this metric for audio applications, including binaural displays, is somewhat subtle. The Hankel norm is a lower bound to the $\mathcal{L}^\infty$ norm, which has a clear spectral interpretation. Additionally, it is often observed in practice that Hankel-optimal methods yield solutions for which the Hankel error is a fortuitously tight lower bound on the $\mathcal{L}^\infty$ error. As such, we explore Hankel-based order reduction techniques in the next section. To aid in the interpretation of these techniques, the $\mathcal{L}^\infty$ and Hankel norms are described below along with the corresponding convolution and Hankel operators. The development below is based on a review given in (119), and is general for all MIMO systems.

Consider a matrix $\mathbf{X} \in \mathbb{R}^{p\times m}$. The 2-induced norm of this matrix is defined as

$$\|\mathbf{X}\|_{2\text{-ind}} \triangleq \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\|\mathbf{X}\mathbf{u}\|_2}{\|\mathbf{u}\|_2} \tag{2.4}$$

where $\|\cdot\|_2$ is the standard Euclidean, or 2-norm, of a vector. If the matrix is viewed as a linear map, $\mathbf{X} : \mathbb{R}^m \rightarrow \mathbb{R}^p$, then the 2-induced norm is the maximum gain of the map. The *singular value decomposition* (SVD) of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^* \tag{2.5}$$

where $\mathbf{U}$ and $\mathbf{V}$ are square unitary matrices. $\mathbf{S}$ is a diagonal rectangular matrix with the *singular values* of $\mathbf{X}$ arranged along the diagonal, $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{N_0})$, where $N_0 \leq \min(p, m)$ is the rank of $\mathbf{X}$. The SVD can be viewed as a *dyadic decomposition* of $\mathbf{X}$ into a sum of rank one matrices,

$$\mathbf{X} = \sum_{k=1}^{N_0} \sigma_k u_k v_k^* \tag{2.6}$$

where u_k and v_k are the k^{th} columns of $\mathbf{U}$ and $\mathbf{V}$. It is simple to show that $\|\mathbf{X}\|_{2\text{-ind}} = \sigma_1$. Furthermore, for a matrix $\widehat{\mathbf{X}}$ with size $p \times m$ and rank $N < N_0$, then

$$\|\mathbf{X} - \widehat{\mathbf{X}}\|_{2\text{-ind}} = \sigma_1\big(\mathbf{X} - \widehat{\mathbf{X}}\big) \geq \sigma_{N+1}\big(\mathbf{X}\big) \tag{2.7}$$

where $\sigma_1\big(\mathbf{X} - \widehat{\mathbf{X}}\big)$ is the largest singular value of $\mathbf{X} - \widehat{\mathbf{X}}$ and $\sigma_{N+1}\big(\mathbf{X}\big)$ is the $(N+1)^{\text{th}}$ largest singular value of $\mathbf{X}$. This inequality is known as the Schmidt-Mirsky theorem and gives a lower bound on how well any low-rank matrix $\widehat{\mathbf{X}}$ can approximate $\mathbf{X}$ in the 2-induced norm.

Consider the state-space system $\boldsymbol{\Sigma}$ with M inputs and P outputs given in (2.1). Associate the following *convolution operator* $\mathcal{L} : \mathbf{u} \mapsto \mathbf{y}$ with $\boldsymbol{\Sigma}$:

$$\mathbf{y}[n] = \sum_{k=-\infty}^{n-1} \mathbf{h}[n-k]\mathbf{u}[k], \quad n \in \mathbb{Z} \tag{2.8}$$

where $\mathbf{h}[n]$ is the matrix impulse response defined in (2.2). The convolution operator can be expressed in matrix form as

$$\left[\begin{array}{c} \vdots \\ \mathbf{y}[-1] \\ \hline \mathbf{y}[0] \\ \mathbf{y}[1] \\ \vdots \end{array}\right] = \underbrace{\left[\begin{array}{cc|ccc} \ddots & & \vdots & & \iddots \\ & \mathbf{0} & \mathbf{0} & \mathbf{h}[1] & \\ \hline \dots & \mathbf{0} & \mathbf{h}[1] & \mathbf{h}[2] & \dots \\ & \mathbf{h}[1] & \mathbf{h}[2] & \mathbf{h}[3] & \\ \iddots & & \vdots & & \ddots \end{array}\right]}_{\mathcal{L}} \underbrace{\left[\begin{array}{c} \vdots \\ \mathbf{u}[0] \\ \hline \mathbf{u}[-1] \\ \mathbf{u}[-2] \\ \vdots \end{array}\right]}_{\mathcal{U}}$$

The 2-induced norm of the system $\mathbf{\Sigma}$ is defined as

$$\|\mathbf{\Sigma}\|_{2\text{-ind}} \triangleq \|\mathcal{L}\|_{2\text{-ind}} = \sup_{\mathcal{U}\neq\mathbf{0}} \frac{\|\mathcal{L}\mathcal{U}\|_2}{\|\mathcal{U}\|_2} \tag{2.9}$$

The discrete-time Fourier transform can be applied to each element of the impulse response $\mathbf{h}[n]$, yielding the $P \times M$ matrix frequency response $\mathbf{H}(\omega)$. Due to the equivalence between time and frequency domains, it is straightforward to show that

$$\|\mathbf{\Sigma}\|_{2\text{-ind}} = \sup_{\omega} \sigma_1\big(\mathbf{H}(\omega)\big) \triangleq \|\mathbf{\Sigma}\|_{\mathcal{L}} \tag{2.10}$$

where $\sigma_1\big(\mathbf{H}(\omega)\big)$ is the largest singular value of $\mathbf{H}(\omega)$. Hence the 2-induced norm is also known as the $\mathcal{L}^\infty$ norm. For SISO systems, this norm is equal to the maximum spectral magnitude.

Optimal causal approximations of MIMO systems in the $\mathcal{L}^\infty$ norm are not currently known[3]. However, if the domain and range of the convolution operator are restricted, then optimal solutions are known. For this reason, we consider the *Hankel operator* of $\mathbf{\Sigma}$, $\mathcal{H} : \mathbf{u}_- \mapsto \mathbf{y}_+$, which maps past inputs to future outputs:

$$\mathbf{y}[n] = \sum_{k=-\infty}^{-1} \mathbf{h}[n-k]\mathbf{u}[k], \quad n \in \mathbb{Z}_+ \tag{2.11}$$

Note that if $\mathbf{\Sigma}$ is causal, stable, and has no feed-through term, then there is a one-to-one relationship between the impulse response and Hankel operator. The Hankel operator can be expressed in matrix form as

$$\begin{bmatrix} \mathbf{y}[0] \\ \mathbf{y}[1] \\ \mathbf{y}[2] \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{h}[1] & \mathbf{h}[2] & \mathbf{h}[3] & \\ \mathbf{h}[2] & \mathbf{h}[3] & \mathbf{h}[4] & \dots \\ \mathbf{h}[3] & \mathbf{h}[4] & \mathbf{h}[5] & \\ & \vdots & & \ddots \end{bmatrix}}_{\mathcal{H}} \begin{bmatrix} \mathbf{u}[-1] \\ \mathbf{u}[-2] \\ \mathbf{u}[-3] \\ \vdots \end{bmatrix}$$

[3] In some instances, such as one-step order reduction, $\mathcal{L}^\infty$ optimal approximations are known (119).

The matrix $\mathcal{H}$, while potentially infinite in size, has rank not greater than N_0. The rank is exactly N_0 if and only if the system $\mathbf{\Sigma}$ is minimal. Let $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{N_0})$ be the singular values of $\mathcal{H}$. The *Hankel norm* of the system $\mathbf{\Sigma}$ is defined as the maximum singular value of $\mathcal{H}$

$$\|\mathbf{\Sigma}\,\|_{\mathcal{H}} \triangleq \sigma_1 \; = \|\mathcal{H}\|_{2\text{-ind}} \tag{2.12}$$

It can be shown that the Hankel norm lower bounds the $\mathcal{L}^\infty$ norm. It can also be shown that twice the sum of the Hankel singular values upper bounds the $\mathcal{L}^\infty$ norm

$$\sigma_1 \leq \|\mathbf{\Sigma}\,\|_{\mathcal{L}} \leq 2(\sigma_1 + \cdots + \sigma_{N_0}) \tag{2.13}$$

Let $\widehat{\mathbf{\Sigma}}$ be another state-space system with M inputs, P outputs, and order $N < N_0$. Then

$$\sigma_{N+1} \leq \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\,\|_{\mathcal{H}} \leq \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\,\|_{\mathcal{L}} \tag{2.14}$$

where the first inequality follows directly from (2.7) and (2.12), and the second inequality follows from (2.13). It can also be shown that the $\mathcal{L}^\infty$ error is upper bounded by both the 1-norm applied to $\mathbf{h}[n] - \widehat{\mathbf{h}}[n]$, as well as twice the sum of the Hankel singular values of the error system $\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}$ (120; 121), although we have found in practice that these bounds are typically loose.

Remarkably, it can be shown that there exists a low-order system that achieves the lower bound on the Hankel error in (2.14). This result was proven for Hankel operators by Adamjan, Arov and Krein and is known as the AAK theorem (122). Later, Glover (117) extended this result to state-space systems and developed a method for computing all optimal $\widehat{\mathbf{\Sigma}}$. In the next section we describe both Glover's method, as well as a simpler suboptimal method.

2.2.2 Hankel Order Reduction

Two order reduction techniques are considered, *balanced model truncation* (BMT) and *Hankel-norm optimal approximation* (HOA). In the SISO case, the two methods have been applied to the design of IIR filters (65; 123), and have been directly compared as well (124; 125). However, few comparisons have been made in the MIMO case. Both methods are briefly reviewed below, and detailed algorithms are given in appendix B.

Balanced state-space systems are a special form that allow for direct order reduction (126). In this form, any state $\mathbf{x}_0$ that results in a 'small' amount of energy output (with the input set to zero) also requires a 'large' amount of energy at the input to move the system from zero to state $\mathbf{x}_0$. Such states contribute little to the input-output behavior of the system, and can be truncated without greatly affecting the transfer function. BMT operates by first applying a balancing similarity transform to $\mathbf{\Sigma}$, and then discarding all but the N largest *Hankel singular values* of the system. For HRTF modeling, the BMT solution can be computed directly from the sample Hankel matrix, $\mathcal{H}$, without constructing and explicitly balancing the high-order $\mathbf{\Sigma}$ (116).

While BMT is convenient, it is not optimal in any specific sense. The HOA method is also based on the theory of balanced systems. However, rather than simply truncate the $N_0 - N$ least significant states, HOA operates in a more sophisticated manner. The $N + 1^{\text{th}}$ state is removed directly by means of an *all pass dilation*, and the remaining $N_0 - N - 1$ states are transformed so as to be both antistable and anticausal, in which case they no longer influence the Hankel operator of the system. The order N stable subsystem is then extracted, yielding a Hankel-optimal approximant $\widehat{\mathbf{\Sigma}}$ (117)

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_{\mathcal{H}} = \sigma_{N+1}(\mathcal{H}) \tag{2.15}$$

The following bounds apply to order $N < N_0$ systems reduced using either BMT or HOA[4]

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\,\|_{\mathcal{H}} \leq \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\,\|_{\mathcal{L}} \leq 2(\sigma_{N+1} + \cdots + \sigma_{N_0}) \tag{2.16}$$

In practice, the upper bound on the $\mathcal{L}^\infty$ error is often loose, whereas the Hankel error is often a relatively tight lower bound on the $\mathcal{L}^\infty$ error.

Both the Hankel error and the $\mathcal{L}^\infty$ error are reported in Chapter III for several state-space and FIR systems that approximate HRTF filter arrays. A perceptual RMS error is also reported in Chapter III. However, a significant problem for state-space approximants of HRTFs is not clearly reflected in any of these error measures. This problem relates to the interaural time delay, and is explored in detail in the next section and possible solutions are proposed.

2.3 ITD Modeling

The monaural phase response of HRTFs may not need to be accurately modeled (57), but the binaural time-delay between ipsilateral and contralateral HRTFs is perceptually critical. Human listener's are sensitive to changes in the *interaural time delay* (ITD) as small as several microseconds (1). This perceptual sensitivity presents a difficulty in the design of low-order MIMO state-space systems.

Hankel methods are known to well approximate transfer functions that are minimum phase, or nearly so (65). However, including time-delay terms in the transfer functions introduces zeros outside the unit circle. Cancelling these zeros with state-space poles during order reduction distorts the phase response of the approximant.

[4] If the state-space system is given a feed-through path, then the upper bound on the $\mathcal{L}^\infty$ error is halved for the HOA method (117).

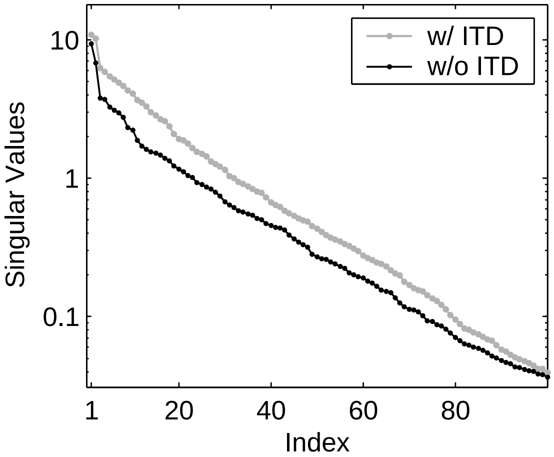


Figure 2.3: The first 100 Hankel singular values for two full-order SIMO systems. The two systems model the same 19 contralateral HRTFs. One system includes the ITD (gray line) and other does not (black line). Both linear and log scales are shown.

For contralateral HRTFs, the time-delay terms introduce up to 40 zeros, at 44.1kHz sampling rate, outside the unit circle. In order to retain the contralateral time-delays in the low-order approximant exactly, 40 state-space poles must be restricted to the origin. The ILD compounds this problem, as pole positions are weighted more by the ipsilateral HRTFs, leaving few 'left-over' poles for time-delay modeling. There are few analytic tools for understanding the time-domain distortion if too few poles are restricted to the origin. In lieu of a simple analytic theory, an example makes clear the practical influence of ITD on state-space order reduction.

Consider a SIMO system that models 19 contralateral HRTFs in the horizontal plane with azimuth angles ranging from 0° to 180° in 10° increments. Two full-order SIMO systems are built from the 19 measured HRIRs. For one system the natural time-delays due to ITD are included in the impulse responses, but are removed for the other system. The Hankel singular values for these two systems are shown in Fig. 2.3. The first 100 singular values of the system with ITD are significantly larger

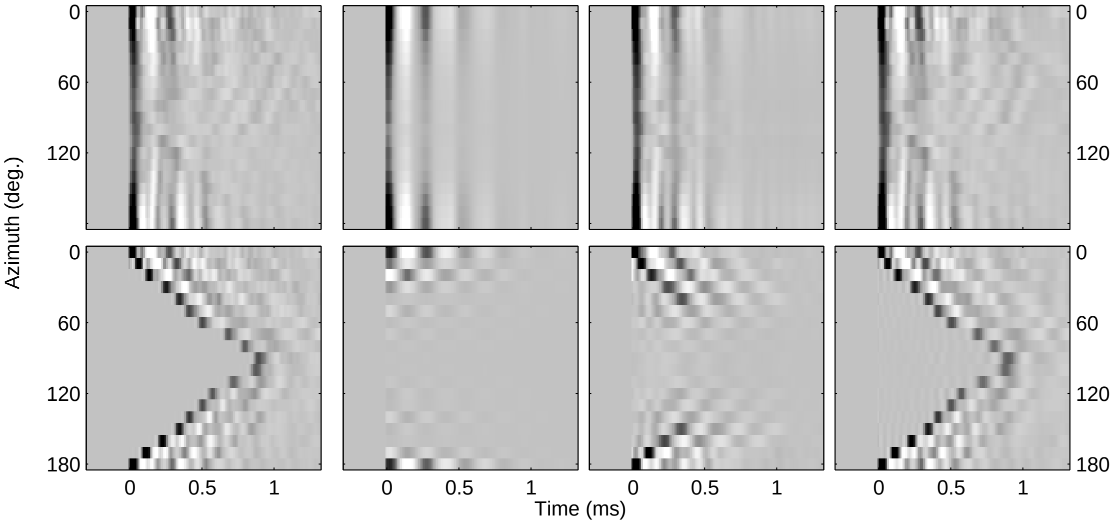


Figure 2.4: Matrix impulse responses for eight SIMO systems; four without ITD (top row) and four with ITD (bottom row). This is equivalent to the first 19 rows of the Hankel matrix of each system. The leftmost column shows the ideal matrices for this system (constructed from 19 contralateral HRIRs). The remaining three columns show response matrices for $N = (1, 6, 30)$ approximants.

than those of the system without ITD. Hence, if the ITD is included in the HRTFs, any low-order approximant with $N < 100$ will have larger Hankel error, and likely larger $\mathcal{L}^\infty$ error, than it would if the ITD is neglected. This performance degradation does not affect all 19 transfer functions equally, however.

The block impulse responses of the two high-order systems are shown at the left of Figure 2.4. Time is indicated by the horizontal axis, and azimuth is indicated by the vertical axis[5]. The right three columns of Figure 2.4 show impulse responses for low-order systems designed using HOA. From left to right the reduced system orders are $N = (1, 6, 30)$. For the system without ITD, the order $N = 1$ system retains most of the energy of all impulse responses. In contrast, for the system with ITD, most of the responses are nearly zero. For the order $N = 6$ approximants, the system

[5] Due to the SIMO architecture, the block impulse responses are equivalent to the top 19 rows of the Hankel matrices of the systems.

without ITD is visually well approximated for all azimuths. For the system with ITD however, the impulse responses far from the median plane remain nearly zero, and even the responses close to the median exhibit obvious distortion. As N increases, both systems are visually well approximated, as shown in the $N=30$ case (127).

The distortion seen in Figure 2.4 is problematic for spatial hearing: there is no longer a clear ITD. A typical example for a practical display is shown in Fig. 2.5. Measured left and right HRIRs for the direction $(az., el.) = (-120°, 0°)$[6] are shown by thin black lines. An order $N = 40$ MIMO system is designed using HOA from measured HRIRs for $D = 68$ directions, including the direction shown in Fig. 2.5. The state-space impulse responses for the same direction are shown by thick gray lines. The left ear response is accurately approximated, but the time delay of the right ear response is 'smeared.' The extent of this phase distortion varies with N, but is generally negligible for $N > 70$, even for large D. We seek systems with lower order, however. In this case the phase distortion is problematic for some directions: impulse responses with small time delay, less than 400μs, exhibit little smearing, whereas responses with large time delay exhibit substantial smearing. We have found that, perceptually, the smearing results in a sound object with increased *diffuseness*, displacement towards the median plane, and possibly even a split into two separate sound objects if the smearing is severe.

The observation above helps explain results presented in (9), where MIMO state-space systems for HRTFs in the horizontal plane were evaluated by human listeners. Localization errors were found to be small for systems with order $N > 80$. However, for systems with lower order, sources far from the median plane were heard to be

[6] Negative azimuth corresponds to the left. Vertical-polar coordinates are used throughout the present work. Azimuth angles vary from $-180°$ to $+180°$, and elevation angles vary from $-90°$ (below) to $+90°$ (above). The direction $(az., el.) = (0°, 0°)$ is straight in front of the listener.

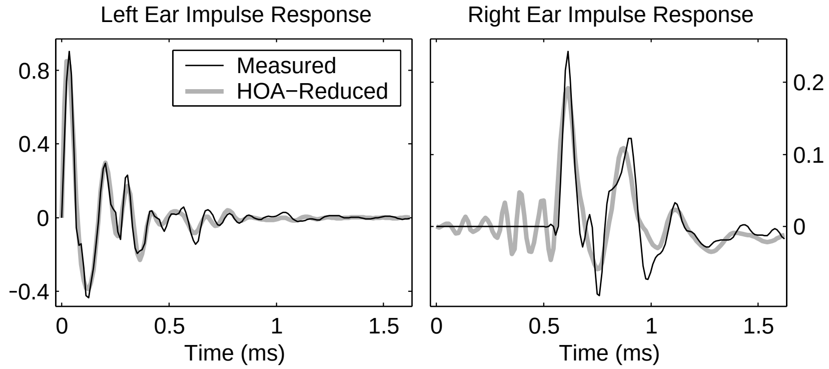


Figure 2.5: Measured HRIRs and HOA-reduced impulse responses for azimuth $\theta = -120°$ and elevation $\phi = 0°$. Note that the left (ipsilateral) ear response is about four time larger than than the right (contralateral) ear response.

located consistently closer towards the median. The poor localization performance far from the median may have been primarily due to smearing of the contralateral time-delay. Were it not for the time-delay distortion, the state-space order may have been reduced farther without incurring substantial error.

2.3.1 Modeling Time-Delay with State-Space Systems

Classical state-space systems theory provides few tools for maintaining multiple time-delays during order reduction. One trivial solution is to design a low-order state-space system that models the HRTFs *without* ITD , using either BMT or HOA. The system is then converted to a canonical form, and augmented with either pure delay terms or Padé approximations (128). However, this method increases the system order dramatically (multiple orders of magnitude), and is untenable in practice.

Time-delay state-space systems have been the subject of several recent studies (129). Unfortunately, attempts to model the HRTF filter array as a single time-delay system have been unsuccessful, as the HRTF filter array is a multi-time-delay system. An *ad hoc* alternative is to employ a state-space system with a single state-update, but let the output be a function of both the present state and T previous state vectors,

$$\begin{aligned}\mathbf{x}[n{+}1] &= \widehat{\mathbf{A}}\mathbf{x}[n] + \widehat{\mathbf{B}}\mathbf{u}[n] \\ \mathbf{y}[n] &= \widehat{\mathbf{C}}_0\mathbf{x}[n] + \widehat{\mathbf{C}}_1\mathbf{x}[n-\Delta_1] + \cdots + \widehat{\mathbf{C}}_T\mathbf{x}[n-\Delta_T]\end{aligned}$$

The system is designed by first constructing a low-order system $\widehat{\mathbf{\Sigma}'}$ that models the HRTFs *without* ITD. $(\widehat{\mathbf{C}}_0, \widehat{\mathbf{C}}_1, \cdots \widehat{\mathbf{C}}_T)$ are then computed by solving, in a least squares sense, an over-conditioned linear system using the measured HRTFs *with* ITD. This procedure was found to yield satisfying performance if $T \approx D$, but rapidly deteriorated for smaller T. Hence this method only yields satisfactory performance for systems with prohibitive computational cost.

A more successful alternative is a hybrid state-space/FIR system. In this case, a state-space system is augmented by FIR filters to reduce the time-delay distortion. The low-order state-space system is designed from a collection of HRTFs, including the ITD, and individual FIR filters connected to the input-output pairs that correspond to HRTFs with large time-delay ($> 0.4ms$). The FIR filter between the m^{th} input and p^{th} output is given by $\mathbf{h}_{pm}[n] - \widehat{\mathbf{h}}_{pm}[n]$, where $\widehat{\mathbf{h}}_{pm}[n]$ is the response achieved by the low-order state-space system alone. A heuristic rule is used to control the computational cost of the FIR filters; a delay operator is incorporated in each filter, and the filter length is restricted to the duration over which $\widehat{\mathbf{h}}_{pm}[n]$ poorly approximates $\mathbf{h}_{pm}[n]$. For example, for the system described in Fig. 2.5, an FIR filter would be added to the right-ear response, and the FIR response would be nonzero from about 0.3ms through 1.2ms. The hybrid system is designed so that the state-space system accounts for two-thirds of the net cost, and the array of FIR 'compensation filters' account for the remaining third.

The methods outlined above are all adequate for modeling ITD using a state-space system. However, inevitably each of these methods incurs a large increase in the computational cost. If the total cost must be minimized, then each of these methods yields mediocre time-domain performance. Of these methods, the hybrid state-space/FIR system yields the most promising performance, and this system is included in the comparisons in the next chapter. Ultimately, any state-space time-delay necessarily increases the underlying system order substantially. Hence, it is preferable to model the ITD outside the state-space system. Of course, this approach can only be applied to state-space architectures that allow the time-delay terms to be factored out of each transfer function. Alternative state-space architectures are described next.

2.3.2 State-Space Architectures

The HRTF filter array shown by the gray box in Figure 2.1 implements a total of $2D$ transfer functions. A MIMO state-space system that implements this filter array as shown has D inputs and 2 outputs. The matrix impulse response for this system is given by (2.3). Clearly, the $2D$ transfer functions must include the appropriate ITD between the left and right ears for each direction. The ITD smearing described above limits the potential for reducing the order the MIMO system, however.

Fortunately, modeling ITD in the state-space can be avoided with alternative architectures. If the arrangement of inputs and outputs is changed such that the state-space system has only a single input or a single output, then the time-delay terms in the contralateral HRTFs can be factored out and implemented externally. This approach was taken in the design of MISO systems in (8). In the present work we consider both SIMO and MISO architectures. Block diagrams of binaural displays with the three architectures are shown in Figure 2.6.

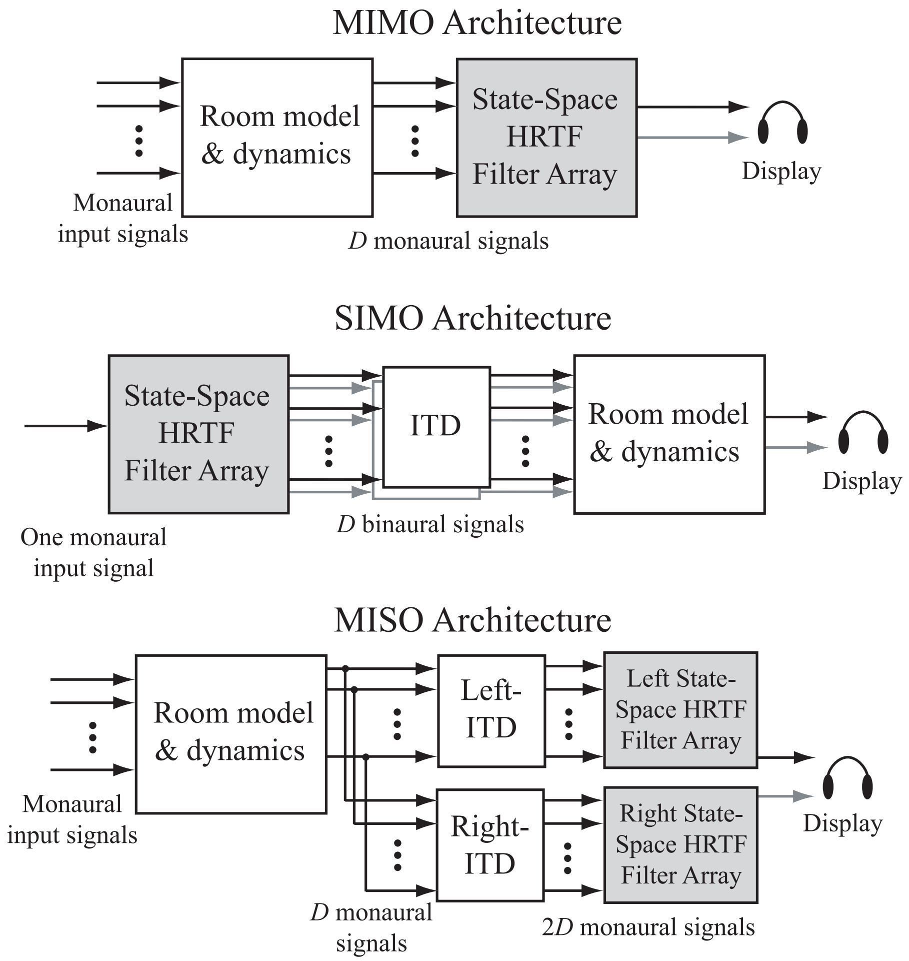


Figure 2.6: Three state-space architectures for binaural display.

If we simplify the binaural display, as shown in Figure 1.9, such that the display only displays a single virtual source, and swap the HRTF filter and the room modeling and dynamics[7], then a SIMO state-space architecture may be used. In this case only a single monaural signal can be input to the system, although spatial-extent, reflections and motion can still be rendered. Hence the SIMO architecture is not a complete solution to the general binaural display problem. Nonetheless, it is a reasonable alternative in the case that only one source signal needs to be rendered. For the SIMO architecture $M = 1$ and $P = 2D$, and matrix impulse response is

$$\mathbf{h}[n] = \begin{bmatrix} h_1^L[n] & h_1^R[n] & h_2^L[n] & h_2^R[n] & \ldots & h_D^R[n] \end{bmatrix}^T \tag{2.17}$$

Alternatively, rather than restrict the binaural display to one input and swap the internal components, we can simply break the MIMO HRTF system into two MISO systems. In this case there are two state-space systems with $M = D$ inputs, and $P = 1$ output, and two impulse response matrices

$$\begin{aligned} \mathbf{h}_L[n] &= \begin{bmatrix} h_1^L[n] & h_2^L[n] & \ldots & h_D^L[n] \end{bmatrix} \\ \mathbf{h}_R[n] &= \begin{bmatrix} h_1^R[n] & h_2^R[n] & \ldots & h_D^R[n] \end{bmatrix} \end{aligned} \tag{2.18}$$

In total, we consider three state-space architectures for the HRTF filter portion of the binaural display: MIMO, SIMO and MISO. Each architecture has one relative disadvantage that may favor the other architectures. The MIMO architecture requires that the ITD be modeled in the state-space, hence time-delay smearing will be a problem for low-order MIMO approximants. The SIMO and MISO architectures allow the ITD to be implemented outside the state-space. However, the SIMO architecture only solves a subset of the general binaural display problem: the single

[7] The order of LTI filters does not affect the overall transfer function of the system.

source case. And the MISO architecture has the disadvantage of requiring two separate systems that model similar transfer functions, due to symmetry of the head. This redundancy would seem to limit the computational efficiency of the MISO architecture. It is unclear a priori which architecture yields the best balance of flexibility and approximation quality.

For the MIMO architecture, every feature of the HRTFs must be implemented by the state-space system, including both ITD and ILD. In contrast, for the SIMO and MISO architectures, it is simple to implement both ITD and ILD externally. We have chosen to only factor the ITD out of the state-space systems, the frequency-independent ILD is left in the state-space. The natural ILD of the HRTFs is mathematically equivalent to weighting the ipsilateral HRTFs more strongly than the contralateral HRTFs during order reduction. Alternatively, the ILD can be implemented externally, in which case the ipsilateral and contralateral HRTFs are weighted equally. However, there is evidence that the spectral detail of the contralateral HRTFs is less important than that of the ipsilateral HRTFs (130; 131; 132; 133), hence the natural weighting due to the ILD may be desirable during order reduction. Weighting is discussed in more detail in Section 3.4.

2.4 Summary

This chapter formulated the HRTF filter array in the state-space and described two methods of order reduction based on the Hankel operator. The Hankel operator and Hankel error were developed in Section 2.2.1, and related to the more common convolution operator and $\mathcal{L}^\infty$ error. The ITD was found to be problematic for state-space order reduction, as any time-delays in the transfer functions require additional states with poles at the origin. Alternative state-space architectures were successfully applied to address this issue by allowing the ITD to be modeled externally. How the alternative architectures might affect the performance is unclear, however.

The performance of the proposed state-space methods is characterized in chapters III and IV. The next chapter describes a large numerical experiment in which state-space approximants are compared to truncated FIR arrays of equal net cost. Chapter IV then reports on a binaural listening experiment.

CHAPTER III

Approximation Performance

To characterize the performance of the state-space systems described in the previous chapter, an empirical experiment is conducted in which low-order state-space systems are constructed and their response is compared to measured HRTFs. The state-space methods are capable of modeling the measured HRTFs exactly, but we seek efficient approximants. In order to show that the state-space approximants are both low-cost and accurate, we include a simple and practical baseline approximation for comparison. The baseline approximation is an array of truncated HRIRs, and is described in Sections 3.1.1 and 3.4.5.

In order to compare the state-space systems with the baseline FIR array, we need a measure of computational cost C that is consistent for both tapped delay-line systems (a.k.a. FIR filters) and state-space systems. We use the arithmetic complexity as a measure of cost. The cost C for both systems is discussed in Section 3.1.2. For all configurations considered below, the state-space and FIR systems are constructed so as to have approximately the same cost.

Section 1.3 reviewed several methods of modeling measured HRTFs. These studies only considered modeling individual HRTFs, however. In the present work, we are interested in approximating collections of HRTFs. The number of directions D that are required for a binaural display depends upon the application. For some applications, such as displays that allow limited motion, only a small number of directions need to be included (5; 7). However, modeling acoustic reflections often requires a large number of directions surrounding the listener (2; 3). Hence, we view D as an independent variable. In the experiment below, we construct HRTF systems, both state-space and FIR, that model a varying number of directions D, but with fixed cost C.

The focus of the numerical evaluation is aggregate performance, although several examples are given as well. The main experiment is described in Section 3.2, and results are reported in Sections 3.2.1 and 3.2.2. The performance of the state-space/FIR hybrid system is reported separately in Section 3.2.3. A sample response is given in Section 3.3. The results are discussed in Section 3.4. The performance of the two order reduction techniques, BMT and HOA, is compared in detail, and approximation quality for ipsilateral versus contralateral HRTFs is considered. A headphone listening experiment is then reported on in the next chapter.

3.1 Basis for Comparison

Before describing the experimental procedures, we define the baseline filter array and the computational cost C for both state-space approximants and the baseline.

3.1.1 Baseline Filter Array

In the present work we follow the example of (60), and compare the proposed state-space approximants to truncated minimum-phase FIR filters. An array of order N FIR filters is constructed by truncating all but the first $N+1$ samples of each min.-phase HRIR. Hence the FIR filters are optimal FIR approximants in terms of $\mathcal{L}^2$ error (55).

For any given cost bound, two FIR arrays are constructed, one with cost bound $C_{\max}$, and the other with cost bound $2C_{\max}$. The 'double-cost' FIR array is included to gauge the significance of the improvement in the approximation quality of the state-space systems. In so doing, we will demonstrate that for some configurations, a state-space system not only outperforms an FIR array of equal cost, but also outperforms an FIR array of twice the cost.

3.1.2 Computational Cost

In the experiment below, state-space systems are compared to FIR filter arrays of *equal computational cost*. We define the cost C as the number of multiplication operations required per sample period, or equivalently, the total number of non-zero coefficients in the system. This measure of computational cost is common in filter design applications when comparing FIR and IIR filters (7; 60; 55). An FIR filter array of order N with M inputs and P outputs requires $C = PM(N+1)$ multiplies per sample period. For example, consider a 10-input and 2-output FIR filter array that implements $D=10$ HRTF pairs with order $N=255$. The cost of this array is $C=5120$.

For state-space systems, the computational cost depends on the choice of system realization, as there are many state-space systems with the same input-output behavior. In general, a state-space system of order N with M inputs and P outputs, and no feed-through path, requires $C = N^2 + (P+M)N$ multiplies per time step. However, after a low-order state-space system has been designed, it is possible to apply a similarity transform to the system matrices $(\mathbf{A},\mathbf{B},\mathbf{C})$ to reduce the number of non-zero elements in $\mathbf{A}$. For example, a *modal decomposition* can be used to diagonalize the $\mathbf{A}$ matrix (134), reducing the system cost to $C = (P+M+1)N$.

However, in this case the system matrices are complex. Alternatively, the $\mathbf{A}$ can be transformed to *Jordan canonical form*, a real and nearly diagonal form. Unfortunately, the Jordan form is difficult to compute in general, and notoriously sensitive to quantization error.

A more practical alternative is to employ a *Schur decomposition* to triangularize the $\mathbf{A}$ matrix (134). Because we seek system matrices that are strictly real, the new $\mathbf{A}$ matrix is only quasitriangular, with 1×1 and 2×2 blocks down the main diagonal. An algorithm for computing the real Schur decomposition can be found in (135). In this case the cost of the final state-space system is not greater than $C = N^2/2 + (P+M+1)N$. For example, consider a 10-input and 2-output state-space systems with order $N=20$. The cost of this array is $C=460$.

To speed computation, FIR filters are sometimes implemented in the frequency domain using the overlap-and-add method (55). For FIR filters with high order, this technique yields a substantial computational savings. However, the overlap-and-add method presents several complications that prevent direct cost comparisons. The method increases system latency, as the signal is processed in frames, and also increases memory requirements. For binaural displays, if the overall system latency becomes too large, localization performance as well as the sense of presence degrades (136). The method also requires hardware that can perform FFTs, and accommodate complex numbers. Furthermore, in the present study we are primarily concerned with truncated FIR filters of order $N < 100$, in which case the savings of the overlap-and-add method is modest. For these reasons we only consider direct implementation of the FIR filters using convolution.

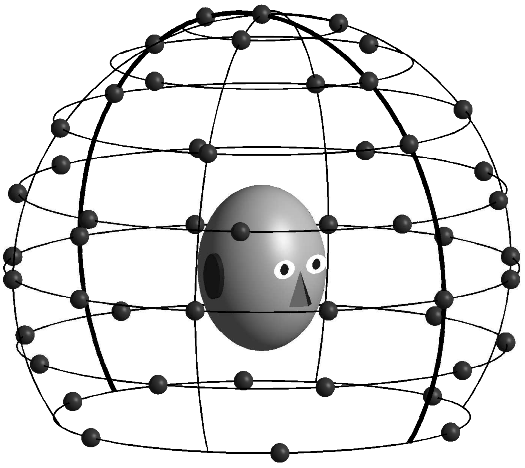


Figure 3.1: Sample collection of $D = 50$ directions.

3.2 Aggregate Performance

In the main experiment, approximants are constructed for a varying number of directions, and fixed cost. More precisely, for every D, the largest system order $N \in \mathbb{Z}_+$ is chosen such that the total computational cost of the system does not exceed $C_{\max}$. In the next chapter we consider the distribution of approximation error for different directions, and in the experiment below we focus on aggregate statistics. Separate systems are designed for each the eight individual HRTFs datasets, and averaged. Overall we observed little difference in performance for different individuals.

For a fixed number of directions, there are many possible arrangements of D directions around the listener. We use a simple rule for selecting D directions that is both practical for modeling common auditory scenes and limits the redundancy in the transfer function matrix. For every system constructed, D directions are chosen randomly subject to a constraint that they be approximately uniformly arranged around listener, but without perfect left-right or front-back symmetry. Figure 3.1 shows a sample collection of $D = 50$ directions.

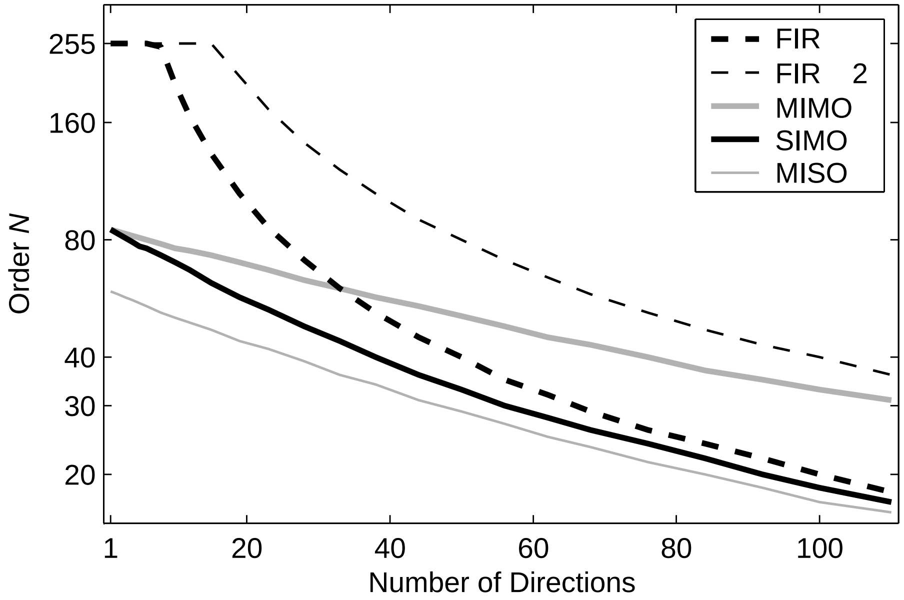


Figure 3.2: System order for five systems as a function of the number of directions D. The system labelled 'FIR ×2' has a cost bound of 8000, and the remaining four have a cost bound of 4000.

A cost bound of $C_{\max} = 4000$ is used below, which is approximately the cost of eight full-order HRIR pairs. State-space systems are designed that meet this bound for a varying number of directions $1 \leq D \leq 110$. Three architectures are considered: MIMO, SIMO and MISO, as described in Section 2.3.2. The ITD is included in the MIMO system, but not by the SIMO or MISO systems. Systems are designed using both BMT and HOA, as described in Section 2.2.2. Hence for every D, six state-space systems are designed. Two FIR arrays are designed for each D as well.

Figure 3.2 shows the order N for five systems: MIMO, SIMO and MISO state-space systems with cost $C \leq C_{\max}$, an FIR array with cost $C \leq C_{\max}$, and a second FIR array with cost $C \leq 2C_{\max}$. For the MISO architecture, N is the order of each state-space system. Note that for the two FIR filter arrays, it is not necessary to truncate the measured HRIRs if $D \leq (8, 16)$ in order to satisfy the cost constraint.

The hybrid state-space/FIR system for the MIMO architecture, as described in

Section 2.3.1, is also considered in the experiment. In the interest of simplifying the presentation, the performance of this system is given in Section 3.2.3.

3.2.1 Hankel and $\mathcal{L}^\infty$ Results

The Hankel error is defined as the largest Hankel singular value of the error system $\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}$ (117), and computation is straightforward. The $\mathcal{L}^\infty$ spectral error is estimated by finding the maximum largest singular value of the transfer function error matrix, $\mathbf{H}(\omega) - \widehat{\mathbf{H}}(\omega)$, over a finely sampled frequency grid. This method of estimating $\mathcal{L}^\infty$ is computationally expensive, but was not problematic for the experiment. Efficient methods of estimating $\mathcal{L}^\infty$ have been developed (137).

Figure 3.3 shows the Hankel and $\mathcal{L}^\infty$ errors of state-space and FIR systems with order given by Figure 3.2. Each panel corresponds to one architecture, from top to bottom: MIMO, SIMO and MISO. The $\mathcal{L}^\infty$ error is shown with black lines, and the Hankel error is shown with gray lines. Note that the truncated FIR filters are, individually, identical for all three architectures, but that the Hankel and $\mathcal{L}^\infty$ errors depend on the architecture. For the MISO architectures, the errors shown are the mean errors for the left-ear and right-ear systems.

Both errors increase monotonically as a function of D for all systems. The performance trends are very similar for all three architectures. As expected, the Hankel error lower bounds the $\mathcal{L}^\infty$ error in all cases. For the FIR systems, the Hankel error is a loose bound on the $\mathcal{L}^\infty$ error for $D > 40$. In contrast, for the state-space systems, the Hankel error is a relatively tight bound for all D.

For $D < 8$ the FIR array yields zero error, and for $D < 16$ the ‘double-cost’ FIR array yields zero error. For larger D however, the error of the FIR arrays increases significantly. Both state-space designs, BMT and HOA, outperform the FIR systems

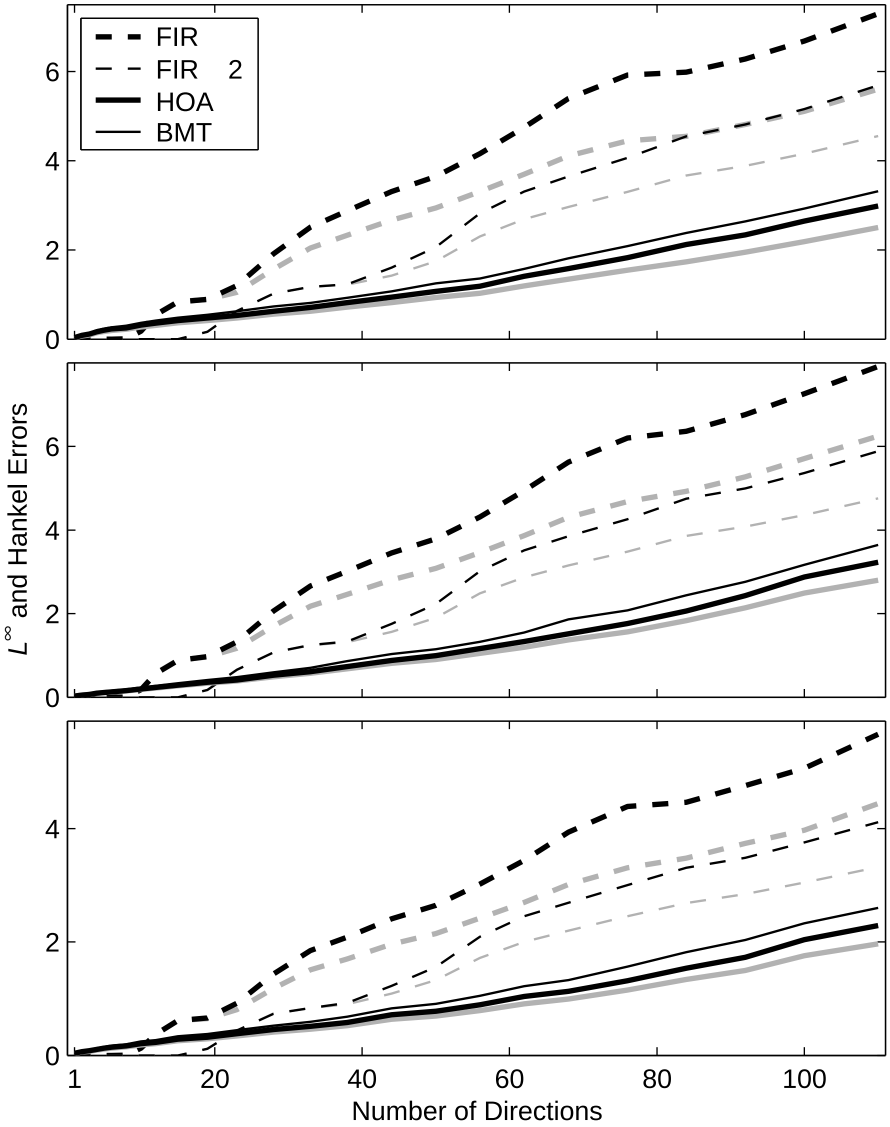


Figure 3.3: Hankel (gray) and $\mathcal{L}^\infty$ (black) error as a function of D. Errors are reported separately for each architecture: MIMO (top), SIMO (middle) and MISO (bottom). The Hankel error of the BMT systems is approximately equal to, and is hidden by, the $\mathcal{L}^\infty$ error of the HOA systems for all architectures.

for $D > 25$. The two order reduction methods show similar performance, although HOA yields lower Hankel and $\mathcal{L}^\infty$ errors than BMT. These results are promising, but system theoretic measures of performance are not necessarily suitable for auditory applications. Next we consider a modified $\mathcal{L}^2$ error as a simple model of auditory perception.

3.2.2 Auditory $\mathcal{L}^2$ Results

The 'auditory' $\mathcal{L}^2$ error that we report has been previously employed in HRTF approximation studies (60). This error measure is computed by warping the log-magnitude response of the ideal and low-order systems to a log-frequency scale. A fifth-octave smoothing filter is then applied to both responses to model the critical bands of the auditory system. The $\mathcal{L}^2$ error (RMSE) between the two modified responses is then computed over the range 300 Hz to 16 kHz. The auditory $\mathcal{L}^2$ error is computed for each input-output pair, and averaged.

The lower bound of the $\mathcal{L}^2$ integration is lower than in (60). In free-field conditions, ILD cues below 1 kHz do not dominate the localization of sound sources (1). However, it has been shown that low-pass sources can be localized in elevation (20). Furthermore, studies have shown that sensitivity to ILD cues extends below 1kHz in the presence of reflecting surfaces (114), and that the sense of 'externalization' is affected by ILD cues below 1 kHz (138). In fact, recently it has been shown for gerbils that the presence of a reflecting surface introduces perceptually salient magnitude features for localization as low as 500 Hz, whereas such features only appear above 10 kHz in the free-field case (89). Because we are interested in low-cost systems for binaural environment modeling, we include frequencies as low as 300 Hz in the auditory $\mathcal{L}^2$ error.

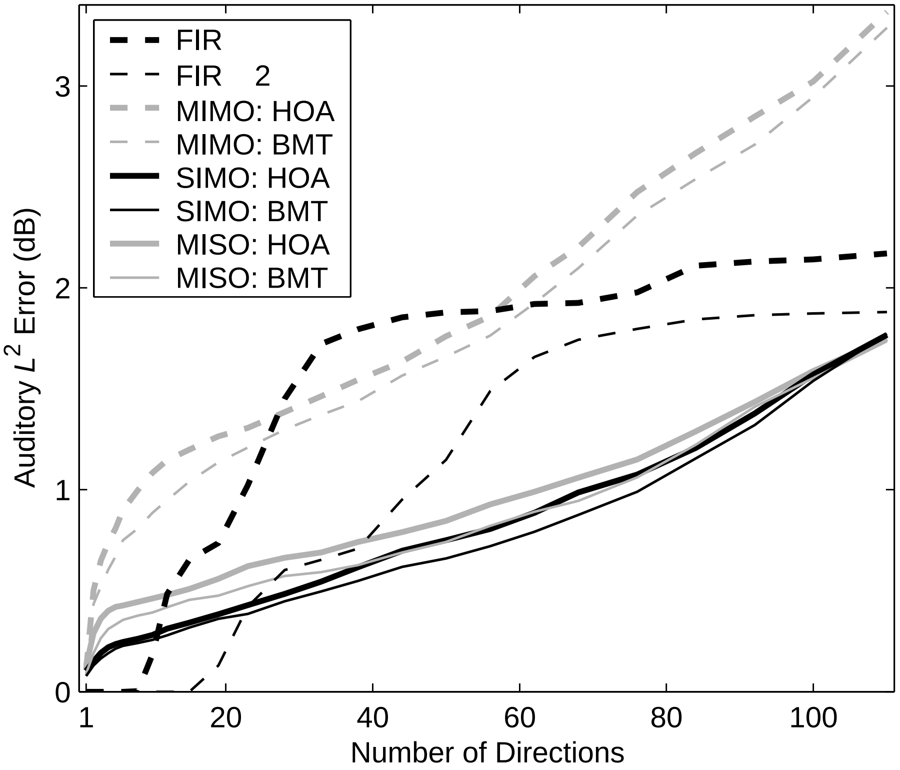


Figure 3.4: Auditory $\mathcal{L}^2$ error as a function of D. The auditory $\mathcal{L}^2$ error is computed for each input-output pair and averaged, and is independent of architecture for the FIR arrays.

Figure 3.4 shows the auditory $\mathcal{L}^2$ error for eight systems: six state-space and two FIR. Naturally, the auditory $\mathcal{L}^2$ error of the FIR systems does not change with architecture. For this error measure, the results of the different architectures can be directly compared, and are shown in a single panel. For $D>10$, the SIMO and MISO architectures yield lower approximation error than the FIR array, and for $D > 40$ the SIMO and MISO architectures outperform the 'double-cost' FIR array as well. The SIMO architecture yields slightly lower error than the MISO architecture, although this difference vanishes for $D > 100$. Furthermore, the BMT systems yield slightly lower error than the HOA systems for $D<100$.

In contrast to the SIMO and MISO architectures, the MIMO architecture does not perform well, for both BMT and HOA. For $30 < D < 60$, the state-space systems yield slightly lower error than the FIR array, but greater error for other D. As will be discussed in Section 3.4, the mediocre performance of the MIMO architecture is due to the ITD and contralateral HRTFs. We now report on the performance of the hybrid state-space/FIR method.

3.2.3 Hybrid System Performance

The SIMO and MISO architectures described above model the ITD externally. This approach ensures that the ITD is accurately modeled, and that all of the state-space poles are free to model the spectral detail in the HRTFs. However, the MIMO architecture requires that the ITD be modeled by the state-space system. Section 2.3 demonstrated that modeling the ITD with low-order state-space systems yields unacceptable performance, even if the $\mathcal{L}^\infty$ and auditory $\mathcal{L}^2$ errors are small. Accordingly, we consider augmenting the MIMO state-space systems so as to correct the distortion to the ITD, as described in Section 2.3.1. The performance of the SIMO and MISO state-space systems above surpasses that of the hybrid systems. Nonetheless, the performance of the hybrid is reported here for completeness.

The hybrid state-space/FIR system is designed such that the state-space portion consumes two thirds of the total computational cost, and the array of FIR 'correction' filters consumes the remaining third. The order of the state-space component as a function of D is shown in Figure 3.5, as well as the order of the MIMO state-space system and the two FIR systems.

Figure 3.6 shows the $\mathcal{L}^\infty$ error as a function of D for the hybrid approximants. Hybrid approximants were designed using both BMT and HOA for the state-space component. Figure 3.6 also shows the $\mathcal{L}^\infty$ error for the state-space and FIR systems

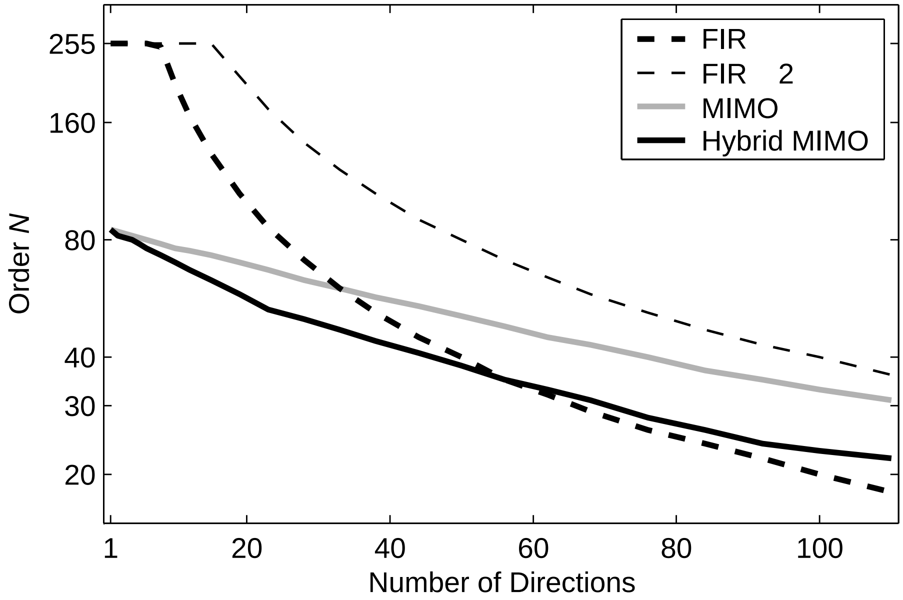


Figure 3.5: System order for four systems as a function of D. The order shown for the hybrid system is the order of the state-space component.

shown in the top panel of Figure 3.3. The $\mathcal{L}^\infty$ error of the hybrid systems is lower than that of the equal cost FIR array, but larger than that of the state-space systems.

Figure 3.7 shows the auditory $\mathcal{L}^2$ error for the same six approximants as Figure 3.6. In this case, the hybrid systems yield slightly lower error than the state-space systems, although the improvement is negligible in the region where the state-space systems outperform the equal cost FIR array. The advantage of the hybrid systems are not apparent in the $\mathcal{L}^\infty$ and auditory $\mathcal{L}^2$ errors, however. The motivation for the hybrid system is to reduce the time-delay smearing that results from state-space order reduction, and neither the $\mathcal{L}^\infty$ nor the auditory $\mathcal{L}^2$ errors strongly reflect the degree of time-delay smearing.

Time-domain distortion is often quantified from the phase spectra of the original and approximant responses. For binaural applications, such error measures are difficult to interpret. We are not concerned with the overall phase response, but only

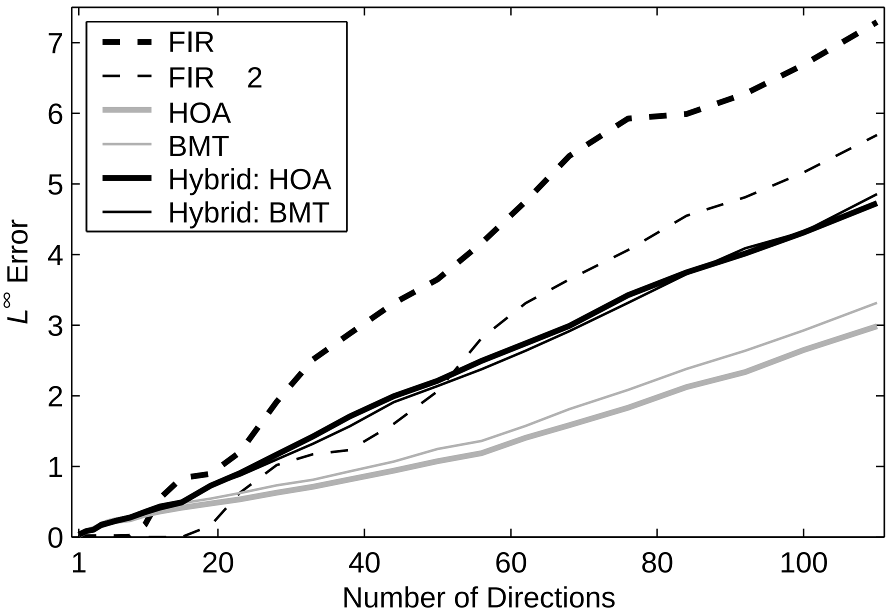


Figure 3.6: $\mathcal{L}^\infty$ error as a function of D.

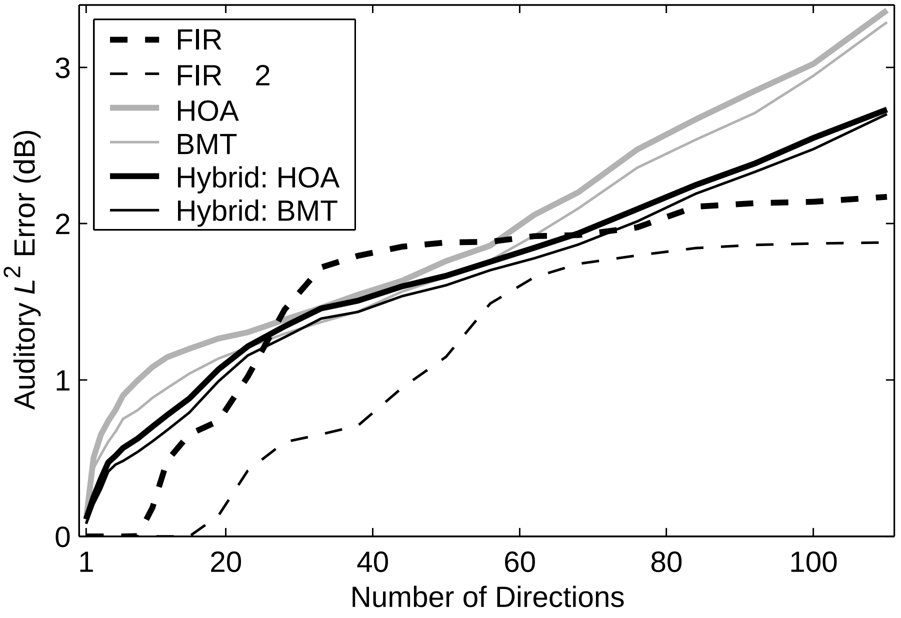


Figure 3.7: Auditory $\mathcal{L}^2$ error as a function of D.

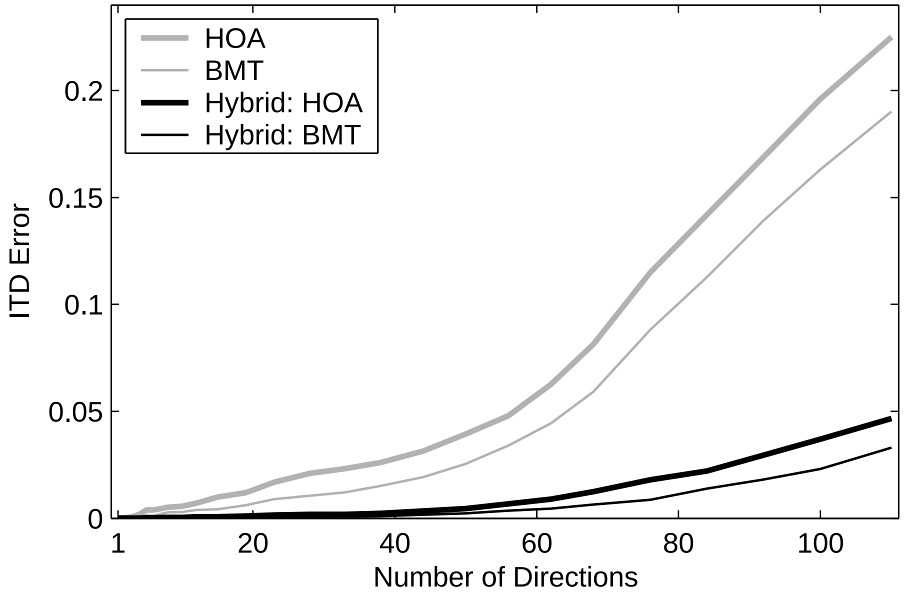


Figure 3.8: The fraction of approximant impulse response energy that appears prior to the time-delay given by ITD in the contralateral responses, as a function of the number of directions D.

the ITD. Accordingly, we quantify the ITD smearing using a simple and intuitive measure: the “ITD error” is defined as the fraction of the approximant impulse response energy that appears prior to the time-delay mandated by the ITD. The ITD error for each contralateral response is measured, and averaged.

Figure 3.8 shows the ITD error for the MIMO hybrid and state-space systems. Of course, the ITD error for the FIR arrays, as well as the SIMO and MISO state-space systems, is zero. From Figure 3.8, the hybrid systems have successfully reduced the ITD error by a factor of four. However, it is unclear whether this reduction is sufficient for binaural applications. Even small shifts in the ITD are problematic for binaural displays, but little is known about ITD *smearing*. For the hybrid systems, less than 5% of the impulse response energy appears prior to the time-delay, and the main ‘pulse’ of the response appears at exactly the right time-delay. This may be a sufficient condition for perceptually adequate HRTF models.

Finally, from Figure 3.8 it is apparent that BMT yields lower ITD error than HOA for both the hybrid and state-space systems. This is discussed in greater detail in Section 3.4.

3.3 Local Structure

Before discussing the aggregate performance below, we present an example. The local structure of the responses reveals additional differences between the FIR and state-space approximants.

Consider a SIMO system that models $D = 44$ HRTF pairs ($M = 1$, $P = 88$). One direction included in this system is the direction $(az., el.) = (+120°, 0°)$. The HRTF magnitude responses for this direction are shown in Figure 3.9. The left column of Figure 3.9 shows the response over a wide frequency range, from 100 Hz to 20 kHz. The right column gives the response over more narrow frequency range, as indicated by the dotted lines in the left column.

Two low-cost systems are constructed from the 88 transfer functions: a state-space system designed using the HOA method, and an array of FIR filters[1]. Both systems are designed not to exceed a cost bound of $C_{\text{max}} = 3000$. For the state-space system, order $N = 28$ is chosen, yielding a net system cost of $C = 2912$. For the FIR array, order $N = 33$ is chosen, yielding a net system cost of $C = 2992$.

The magnitude responses of the two approximants at $(+120°, 0°)$ are shown in Figure 3.9. At low-frequencies, the response of the FIR filters diverge from the desired response, especially below 300 Hz. Although, it is unclear if accurate low-frequency magnitude modeling is perceptually critical for spatial listening.

From the right column of Fig. 3.9, it is apparent that the spectral notches in

[1] A state-space system designed with BMT was also considered. The magnitude response of the BMT and HOA systems was visually similar, only the HOA response is shown here.

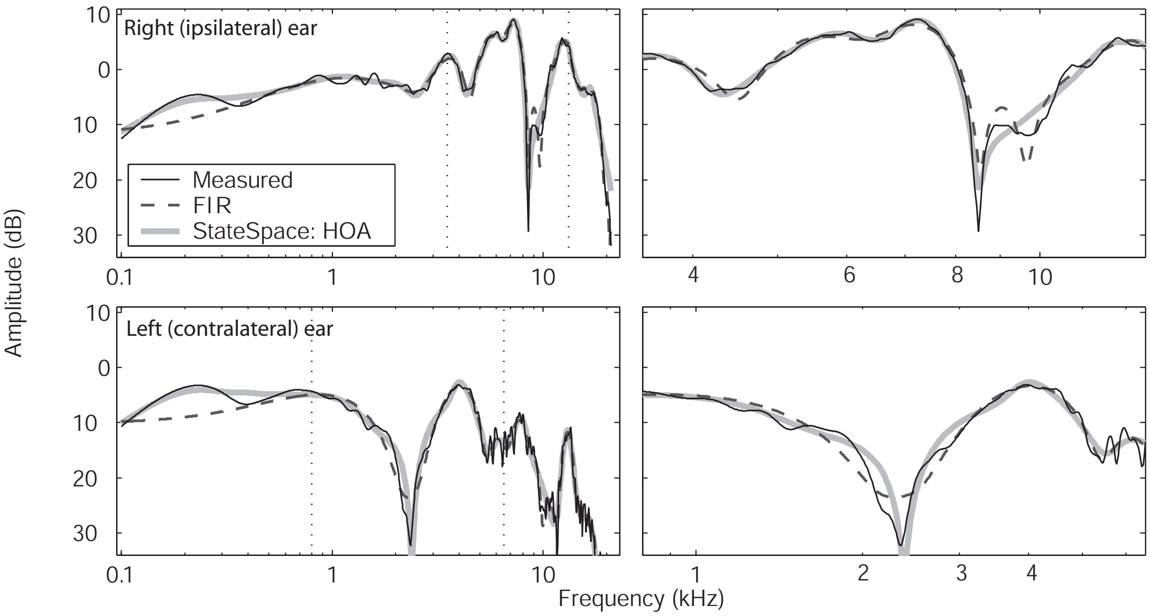


Figure 3.9: Magnitude responses for direction $\theta = 120°$ and $\phi = 0°$, for the right ear (top) and left ear (bottom). The vertical dotted lines in the left column indicate the frequency bounds of the right column.

the measured HRTF are more accurately modeled by the state-space system than the FIR array. For the right ear response, the shallow notch at 4.5 kHz is well approximated by the state-space system, but is shifted by the FIR system. A more significant difference is seen at 8.5 kHz, where the measured HRTF exhibits a sharp, lopsided notch. The state-space system also exhibits a lopsided notch at the same frequency, whereas the FIR system exhibits two notches with the same depth (-18 dB), one at 8.5 kHz and another at 9.7 kHz. The deep notch in the left ear response at 2.5 kHz is well approximated by the state-space system, whereas the FIR system exhibits only a shallow dip at this frequency. This trend is seen throughout the results: spectral notches are more accurately modeled by the state-space systems than the FIR arrays, particularly notches below 5 kHz.

3.4 Discussion

The main experiment above demonstrates that, when modeling collections of HRTFs, equal-cost state-space systems outperform truncated FIR arrays. The aggregate results also exhibit contradicting trends, however. For example, MIMO state-space system yield favorable performance in terms of the $\mathcal{L}^\infty$ error, but not in terms of the auditory $\mathcal{L}^2$ error. Other differences in the trends with these two error measures are apparent: HOA yields lower $\mathcal{L}^\infty$ error whereas BMT yields lower auditory $\mathcal{L}^2$ error. These disparate trends are explained below, along with several other issues that arise when comparing systems with different structures (an array of independent FIR filters versus a single state-space system).

The differences between BMT and HOA are most apparent in the design of 'simple' filter approximants, such as an ideal bandpass filter. Accordingly, we explore the approximation of simple ideal filters using both BMT and HOA in the next subsection. The following subsection explores BMT and HOA applied to HRTF approximation. The third subsection examines the distribution of approximation error between ipsilateral and contralateral HRTFs. The fourth subsection discusses weighting, and the final subsection describes the significance of the FIR baseline, and considers implementation issues when comparing FIR and state-space systems.

3.4.1 BMT versus HOA: Simple Filters

BMT and HOA have been employed in studies of simple SISO filters, although less is known about their relative strengths in the MIMO case (124; 125; 127). For simple filters that consist only of pass-bands and stop-bands, the differences between BMT and HOA are clear and significant. In this section we consider three example filter approximations, two SISO and one MIMO, and compare the error spectrum of

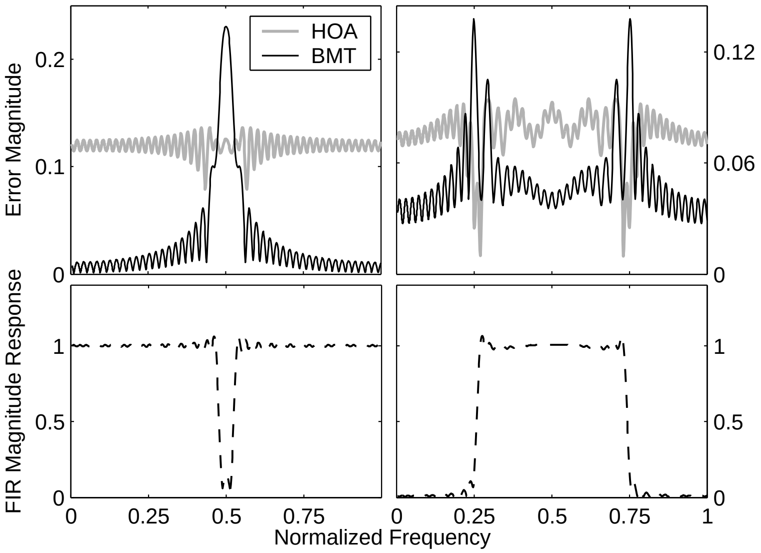


Figure 3.10: The bottom two panels show the magnitude response for two order 100 FIR filters; a notch filter (left) and a band-pass filter (right). The top two panels show the magnitude error of low-order state-space systems designed using either HOA or BMT. For the notch filter $N=6$ state-space systems are designed, and for the band-pass filter $N=10$ systems are designed.

BMT and HOA.

The two SISO examples below are similar to those presented in (125), which directly compares BMT and HOA for IIR filter approximation. The bottom row of Figure 3.10 shows the magnitude response of two $N=100$ FIR filters, a narrow notch filter and a wide band-pass filter. For both filters, two state-space approximants are constructed, one using BMT and one using HOA. The error magnitude is show in the top row of Fig. 3.10. For both filters, the HOA method exhibits relatively flat magnitude error. The BMT method yields lower magnitude error at most frequencies, but also exhibits substantial peaks above the error of the HOA method. For the notch filter example, the BMT method concentrates error in the narrow stop-band. For the band-pass filter, the BMT method concentrates error near the transition bands. These performance trends have been previously observed (124; 125): BMT often

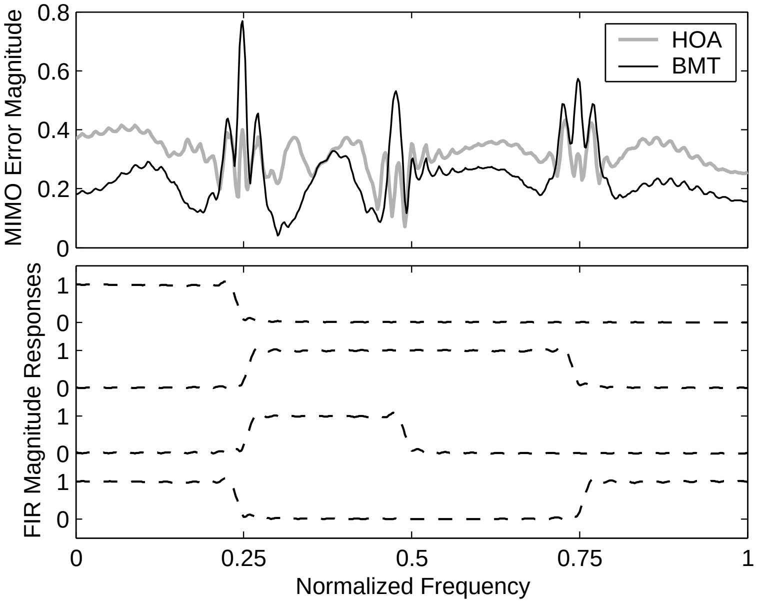


Figure 3.11: The bottom two panels show the magnitude responses for two-input, two-output FIR array with order $N = 100$. The top two panel shows $\sigma_1\big(\mathbf{H}(\omega) - \widehat{\mathbf{H}}(\omega)\big)$, for two $N = 20$ state-space systems designed using HOA and BMT.

concentrates error near stop-bands and transition-bands.

The trends outlined above for the SISO case can be observed in simple MIMO filters as well. Figure 3.11 shows the magnitude response of four $N = 100$ FIR filters. The filters are simple in shape: band-pass, band-stop and low-pass. The four filters are arranged as a 2-input, 2-output MIMO system, and two $N = 20$ state-space approximants are constructed using BMT and HOA. The resulting error magnitude responses are shown in the top panel of Figure 3.11. The error peaks of the BMT system again occur near the transition bands of the original filters. Indeed, the highest error peak exhibited by the BMT approximant occurs at the frequency where all four transfer functions exhibit transitions between stop bands and pass bands. However, in the MIMO case, this performance trend is less consistent than the SISO case. Other examples may be constructed in which the BMT method

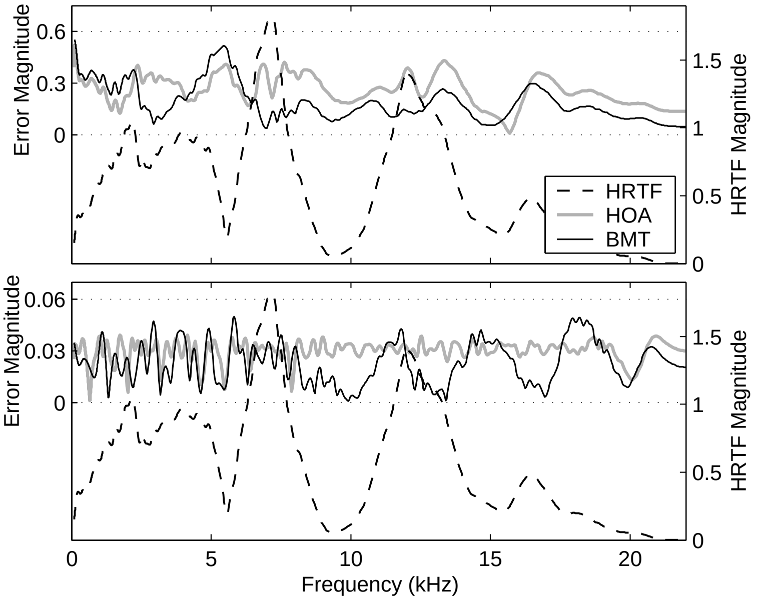


Figure 3.12: HRTF magnitude response for location $(\theta, \phi) = (30°, -36°)$, ipsilateral ear. Also shown are error magnitudes for four low-order approximations. The top panel shows the error responses for two $N = 6$ state-space systems designed using HOA and BMT. The bottom panel shows error responses for two $N = 30$ systems.

concentrates error at frequencies away from the transition-bands.

3.4.2 BMT versus HOA: HRTF Filters

For simple filters shapes, significant differences between BMT and HOA order-reduction methods are apparent. For more complicated filters, such as HRTFs, the two methods yield similar results. Nonetheless, the performance trends outlined above are still present, albeit more subtle.

Figure 3.12 shows the magnitude response of one HRTF[2]. The error magnitude for two $N = 6$ state-space systems is shown in the top panel, and for two $N = 30$ state-space systems is shown in the bottom panel. To facilitate comparison, the error magnitudes are shown in the same panel as the HRTF magnitude; the error

[2] Unlike Fig. 3.9, which shows HRTFs on a log-frequency, log-amplitude scale, Fig. 3.12 shows an HRTF on a linear-frequency, linear-amplitude scale.

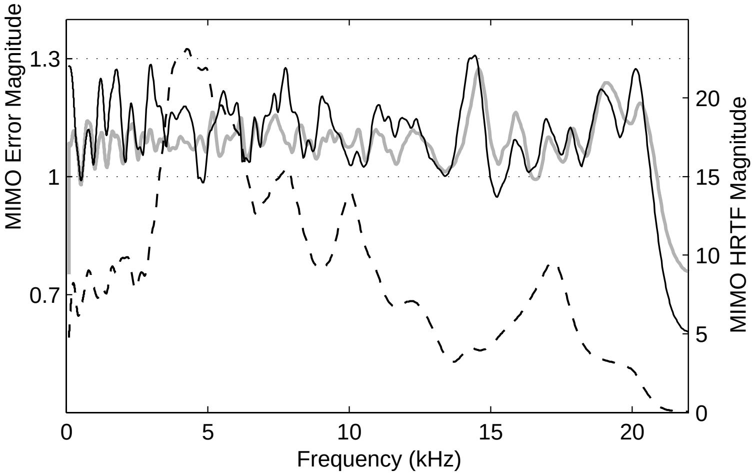


Figure 3.13: MIMO HRTF magnitude response for 44 directions surrounding the listener, $\sigma_1\mathbf{H}(\omega)$, where $H(\omega)$ is a 2×44 matrix function. Also shown are MIMO error responses, $\sigma_1\big(\mathbf{H}(\omega)-\widehat{\mathbf{H}}(\omega)\big)$, for two $N=40$ state-space systems designed using HOA and BMT.

magnitude values are given on the left of the figure, and the HRTF magnitude values are given on the right of the figure. The differences between the error magnitudes for the BMT and HOA systems are not as significant in this example as in the simple filter approximations above. The HOA approximants exhibit a flatter error magnitude than the BMT systems, although at most frequencies the BMT error magnitude is lower than the HOA error. The BMT error magnitude exhibits peaks in the error, some of which are at or near spectral notches in the original HRTF. For example, the errors of both BMT approximants exhibit a peak near the HRTF notch at 5.5 kHz.

HRTFs measured at different directions exhibit spectral notches at different frequencies. A collection of HRTFs is unlikely to exhibit many notches at the same frequency. But notches located at common frequencies across many directions may influence where the BMT method concentrates approximation error. Figure 3.13 shows an example for a MIMO system that models $D = 44$ HRTF pairs. In this case the magnitudes shown are the maximum singular value of the matrix transfer functions. State-space approximants with order $N = 40$ are constructed using both BMT and HOA. Both approximants yield error magnitudes between 1 and 1.3 for most frequencies. The BMT error magnitude again yields greater fluctuations than the HOA error magnitude. The BMT error exhibits several small peaks, some of which are located at or near notches in the global HRTF response. Nonetheless, the two methods yield highly similar results.

For collections of HRTFs, the difference in performance between approximants designed using BMT and HOA is small. Nonetheless, the trends described above explain the disparity between the aggregate $\mathcal{L}^\infty$ and auditory $\mathcal{L}^2$ results. The disparity is due to the difference between the ∞-norm and the 2-norm. The BMT method yields lower error at most frequencies, hence the auditory $\mathcal{L}^2$ is lower with BMT. However, the BMT approximants yields peaks in the error response that rise above the relatively flat error response of the HOA approximants. Hence HOA yields lower $\mathcal{L}^\infty$ error.

In practice, we have found no significant advantage to using either BMT and HOA for approximating collections of HRTFs. HOA yields slightly lower $\mathcal{L}^\infty$ error, whereas BMT yields slightly lower auditory $\mathcal{L}^2$ error. That BMT concentrates error near the spectral notches would be problematic for HRTF approximation if the trend were stronger. Spectral notches in the HRTF are known to provide important cues for spatial hearing (1; 21; 139; 140). In particular, accurate modeling of HRTF notches is thought to be critical for the perception of elevated sources. However, we found the error peaks resulting from BMT to be are small compared to the overall error. On the other hand, if we consider time-domain distortion, the HOA method suffers from a slight drawback. For the SIMO and MISO architectures, time-domain distortion is negligible. For the MIMO architecture however, the ITD smearing is problematic, and the HOA method appears worse in the regard. However, even the BMT method smeared the ITD sufficiently for the method to be unusable.

In summary, while there are relative pros and cons to using BMT versus HOA, the two methods yield very similar results for collections of HRTFs.

3.4.3 Ipsilateral versus Contralateral Approximation

The poor performance of the MIMO state-space systems, which include ITD, is reflected in the auditory $\mathcal{L}^2$ error, but not in the $\mathcal{L}^\infty$ error. This performance disparity is explained by considering the distribution of error between the ipsilateral and contralateral HRTFs. The MIMO approximants model the ipsilateral HRTFs relatively well, but not the contralateral HRTFs. This is shown in Figure 3.14, which gives the same results as Figure 3.4 except that the average ipsilateral and contralateral errors have been separated. The poor contralateral performance is due to the ITD, as discussed in chapter II. However, due to the ILD, the poor approximation quality of the contralateral HRTFs has little impact on the $\mathcal{L}^\infty$ error. In contrast, the auditory $\mathcal{L}^2$ error is computed from log magnitude spectra, hence the ipsilateral and contralateral approximations influence the average error equally.

Several psychoacoustic studies have shown that the magnitude response of the ipsilateral HRTF dominates the monaural cues during spatial hearing (130; 131; 132; 133). Hence it may not be problematic for the contralateral HRTFs to be poorly approximated. Nonetheless, with the MIMO state-space architecture, the time-delay of the contralateral response is also distorted, and this is an unacceptable distortion in most cases.

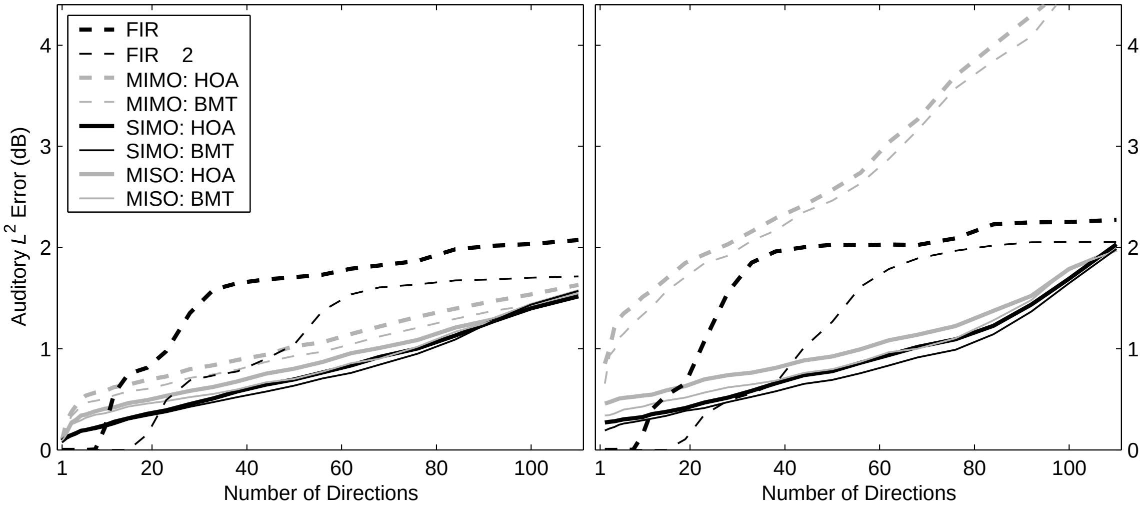


Figure 3.14: Auditory $\mathcal{L}^2$ error as a function of D, ipsilateral HRTFs only (left) and contralateral HRTFs only (right).

3.4.4 Weighting

The previous subsection raises the issue of weighting. For example, due to the perceptual weighting applied to the ipsilateral HRTFs, it may be desirable to weight ipsilateral HRTFs more heavily than contralateral HRTFs during order reduction. For the SIMO and MISO architectures, this is straightforward: simply increase the level of the ipsilateral HRTFs when constructing the original high-order system, perform order reduction, and then decrease the level of the appropriate rows of $\widehat{\mathbf{C}}$ for the SIMO architecture or the appropriate columns of $\widehat{\mathbf{B}}$ for the MISO architecture.

Other weights than ipsilateral versus contralateral may be applied as well. For example, weighting across direction has been proposed in modeling reflective environments (72). The auditory system gives *precedence* to the direct wave in perceiving the location of a sound source (16). Precedence arguments might suggest that some directions be weighted more than others, although the weighting must be time-varying if motion is included. Furthermore, localization performance alone may not be a suitable criteria for the design of a binaural display, as a listener may be able to correctly 'localize' a sound source without being perceptually immersed in the auditory scene[3]. In the interest of flexibility and generality, a uniform weight for both ears and all directions may be preferable.

Frequency weighting is common in the design of audio filters (60). Frequency weighting for audio applications usually heavily weight low frequencies, and weight less high frequencies. For example, Bark-scale frequency weighting is common. Several studies have proposed frequency-weighting extensions to state-space order reduction techniques (141). However, the BMT and HOA methods, as applied to collections of HRTFs, were found to naturally weight low frequencies, rendering frequency weighting unnecessary. Similar results have been reported in studies that model collections of HRTFs using PCA (48).

3.4.5 Baseline Interpretation

Of course, comparing FIR arrays to state-space systems is a bit like comparing apples to oranges. Nonetheless, an array of truncated HRIRs is a natural baseline for studies in HRTF approximation.

Previous studies in HRTF approximation make few comparisons to any other method of approximation. Rather, most studies define a criterion for an HRTF approximation to be adequate, and them demonstrate that the proposed approximation meets the criterion. Different studies have employed different criteria, and as such concluded that very disparate minimum system orders required for adequate HRTF approximation, from orders less than ten (64; 66) to orders greater than one hundred (9; 87).

[3] For example, studies have reported on instances in which a listener correctly localizes a virtual far-field sound source (i.e. its direction), but still perceives the virtual source as being inside the head (2).

One recent review article compares several HRTF approximation techniques (60). In particular, this study directly compares low-order IIR approximants with truncated min.-phase FIR approximants of equal net cost. In this case, IIR filters are shown to outperform FIR filters modestly. This comparison is revealing because any ‘adequacy criteria’ that an IIR approximant satisfied would likely also be satisfied by a truncated HRIR of slightly higher cost.

That truncated HRIRs are natural approximants is further evidenced by the variety of measured HRIR lengths in the published literature. In theory, the duration of an HRTF is infinite, hence the duration of a measured HRTF is restricted by measurement noise or perceptual considerations. Published lengths range from 200 samples at 44.1 kHz (142) to 1000 samples at 50 kHz (19). Furthermore, several studies have reported truncated HRIRs that are ‘perceptually adequate’ with filter orders between 30 and 70 (60; 61; 143). It would seem that for any HRTF approximant to be declared efficient and accurate, it must at least outperform a truncated FIR filter of equal cost.

The experiment above demonstrates that state-space approximants outperform arrays of truncated FIR filters by a significant margin. However, the structural differences between these two systems limits the impact of this result. For example, to reduce computational cost, the FIR array can be implemented in the frequency-domain using the overlap-and-add method (55). We do not consider this possibility here, as such an implementation requires more sophisticated hardware, additional memory, and introduces significant latency.

If the FIR filters are implemented directly with a tapped-delay system, an integer data type is sufficient. Fixed-point data types are not always sufficient for IIR filters or state-space systems, however. For these systems, floating-point data types are sometimes necessary. An informal study of the state-space approximants has revealed that the first- and second-order statistics of the state vector are similar to those of the input and output audio signals. Furthermore, the state-space system matrices have been found to be reasonably robust to small quantization error. These two observations support the use of an integer data type for the state-space approximants. Furthermore, a recent study of IIR filters revealed that fixed-point data types are sufficient for audio applications (144).

3.5 Summary

This chapter reported on a large numerical experiment that characterizes the performance of the state-space methods proposed in the previous chapter. State-space approximants of varying size D, but fixed cost C, were constructed. Three architectures and two order-reduction techniques were considered. A truncated FIR array was used as a practical baseline for comparison. The SIMO and MISO architectures were found to perform well. For $C = 4000$ and $D > 20$, the state-space approximants were found to outperform both the 'equal cost' and the 'double cost' FIR arrays. In particular, the state-space approximants more accurately modeled the spectral notches in the HRTFs, as well as the low-frequency envelope, than FIR arrays.

In contrast, the performance of the MIMO architecture suffered due to the necessity of modeling time-delay in the state-space. The performance with the MIMO architecture was improved somewhat by moving to a hybrid system in which FIR filters were used to mitigate the time-delay smearing that resulted from state-space order reduction. However, the performance of the hybrid system was still inferior to that of the SIMO and MISO state-space systems.

The two order reduction techniques, BMT and HOA, were found to yield similar performance for collections of HRTFs. For simple filter shapes, BMT is known to concentrate error near stop-bands and transition-bands. This is a potentially problematic property for modeling HRTFs, as spectral notches play an important role in spatial hearing. However, the shape of HRTFs is sufficiently complex that BMT does not concentrate error significantly. As such, both BMT and HOA are suitable for constructing low-order HRTF approximants.

CHAPTER IV

Psychophysical Validation

The previous chapter demonstrated that low-order state-space systems can efficiently model collections of HRTFs. For a fixed computational cost, state-space systems can be constructed that yield significantly lower approximation error than FIR filter arrays. While low approximation error is an encouraging property of the state-space systems, we ultimately seek the minimum approximant order for which the listener cannot tell the difference between the approximant and the measured HRTFs. This order threshold is estimated below with a series of three listening experiments.

We estimate order thresholds by using standard psychoacoustic discrimination tasks. In the experiments below, the order of the approximant will be adjusted systematically to determine the threshold at which listeners are able to discriminate a full-order rendering from an approximate one a certain percentage of the time. Stimulus conditions are included which estimate worst-case order thresholds. In addition, order thresholds are estimated for tpical conditions found in music recordings and complex auditory environments.

Observers are instructed to perform the discrimination based on any perceived difference, without paying specific attention to spatial attributes. This particular strategy potentially yields overly conservative order thresholds. On the other hand, this strategy is far less sensitive to interpretation by the listener, and provides a reliable upper-bound estimate of the system order. To estimate this upper-bound, one stimulus condition uses a single sample from a random noise process to generate all instances in each trial. In this case, small changes in the coloration of the sound introduced by the approximant's monaural characteristics can be used by the listener to perform the discrimination task, even if the spatial characteristics of the sound are not affected. Such coloration cues, though subtle, are reliable and we expect the listener to focus on such cues with sufficient training. We expect that when such coloration cues are not available to the listener, then the order thresholds will decrease. Accordingly, other stimulus conditions are included that draw independent samples of the noise process for each instance in a trial, include virtual acoustic reflections, or even employ complex moving sound sources as stimuli. These stimulus conditions allow us to estimate thresholds which are more typical of spatial audio applications.

The baseline array of FIR filters is included in the first two experiments. The baseline is included, in part, to show that the state-space approach is not only accurate, but also efficient. In so doing, we also find that the stimulus conditions affect the order thresholds of the FIR and state-space approximants differently. This allows us to further refine the relative advantages of the state-space approach.

Three experiments are reported in the following. The first experiment was conducted with a single participant, the author. The primary objective of this experiment was to determine if a more statistically efficient psychophysical procedure could be used in the subsequent experiments. The results of Experiment 1 were used to design an 'adaptive-level' procedure for Experiment 2, and a 'method of adjustment' procedure for Experiment 3. For Experiments 2 and 3, discrimination thresholds for five listeners, including the author, are estimated for a variety of stimulus conditions. The first two experiments use stationary wideband noise bursts as stimuli, while the third experiment uses complete virtual auditory scenes constructed from field recordings.

4.1 Experiment I: Psychometric Functions

Standard psychophysical discrimination tasks characterize the psychometric function of an observer, which describes the probability of correct discrimination as a function of some independent physical variable (145). In the present study, the independent physical variable is the approximant system order N. In general, for very small N, listeners easily discriminate between stimuli generated by the full-order system and those generated by the low-order approximant, and performance in the task is 100%. As N grows large, the differences between the stimuli grow small and listeners perform near 50% in the discrimination task.

Methods of estimating the psychometric function are distinguished by whether they estimate the function over the listener's operating range, or simply some reference point (e.g. the 50% correct point) on the function. Furthermore, methods are distinguished by whether they use a 'fixed-level' procedure or an 'adaptive-level' procedure[1]. Fixed-level approaches are generally used to estimate several points along the psychometric function and require that a few thousand discrimination trials be conducted to obtain statistically reliable results. In contrast, adaptive-level proce-

[1] The 'level' is the independent variable. The terms 'fixed-level' and 'adaptive-level' were first employed in signal plus noise experiments in which case the 'level' was the amplitude of the signal relative to the noise. In the present study, the 'level' is the order of the approximant, hence we could use the terms 'fixed-order' and 'adaptive-order.' However, to be consist with the psychophysical literature, we use the terms 'fixed-level' and 'adaptive-level.'

dures estimate one point in the psychometric function, that corresponds to a certain level of discrimination performance, and only require that a few hundred trials be conducted. However, in order to employ an adaptive-level procedure, the operating range of the psychometric function must be known, and the shape of the function must meet certain conditions (146; 147).

The first experiment estimates the psychometric functions of one listener, the author, for ten stimulus conditions. The stimulus conditions vary the approximant type, whether the HRTFs are individualized, and the consistency of the noise stimuli. By comparing the psychometric functions for different conditions, we assess the significance of these experimental variables and use these results in the design of Experiments 2 and 3.

4.1.1 Methods

Stimulus Conditions

Stimuli are constructed by filtering broadband noise bursts with either full-order HRTFs or low-order approximants (either truncated FIR or state-space). The sampling rate for all stimuli is 44.1 kHz. The monaural noise bursts are band-limited from 100 Hz to 16 kHz. The duration of the noise burst is 50 ms, with 5 ms raised cosine-squared ramps applied to the onset and offset of each burst[2].

Ten stimulus conditions are tested. The stimulus variables are:

1. Type of noise source (either ‘identical’ or ‘independent’).

2. Type of HRTF dataset (either individualized or non-individualized).

[2] At 44.1 kHz, the 50ms noise burst consists of 2205 samples. Prior to filtering with the HRTFs, an extra 1024 samples are appended to the end of the noise burst. Hence the total ‘duration’ of the filtered noise burst is 73.2ms. 1024 are more than enough appended samples for the full-order HRTF convolution. The extra samples are included only as a precaution: the state-space approximants do not, in general, yield finite impulse responses. However, we have found in practice that the state-space impulse responses reliably fall to zero quickly (< 0.00001 within 200 samples). In contrast, we have found that some IIR filter designs yield impulse responses that fall to zero slowly.

Table 4.1: System orders N for the three approximant types.

Approximant Type	Orders N						
FIR	6	12	22	38	62	94	140
SIMO State-Space	6	12	20	30	42	56	72
MISO State-Space	6	12	20	28	38	48	62

3. Type of approximant (either FIR, SIMO state-space or MISO state-space).

Of the $2 \times 2 \times 3 = 12$ possible variable combinations, two are eliminated; the MISO architecture is not tested with individualized HRTFs. For the noise sources, identical noises are generated for each trial by sampling a random noise generator once, and using that instance of the noise process for every interval in the trial. Statistically independent noises are generated by sampling a random noise process for each interval of the trial. Stimuli are generated using two distinct HRTF datasets, one measured from the participant in the experiment (the author), and the other measured from a different individual.

The full-order 'standard' stimuli are generated by convolving a noise burst with the $N = 255$ minimum-phase impulse responses of the measured HRTFs. For the FIR approximants, the 'comparison' stimuli are generated by convolving a noise burst with a truncated minimum phase impulse response. For the state-space approximants, the 'comparison' stimuli are generated by processing a noise burst with the low-order state-space system. State-space approximants are constructed using Hankel-norm optimal approximation (HOA). Two state-space architectures are included, SIMO and MISO. For all stimulus conditions, the psychometric functions are estimated at seven points. The seven values of N for the three approximant types are shown in table 4.1. These were chosen to span the range of the psychometric function in approximately uniform steps.

HRTF Collection

The approximants used to generate the comparison stimuli model $D = 50$ HRTF pairs. The collection is chosen randomly for each block such that the 50 directions are distributed approximately uniformly around the head, subject to a constraint that no left-right symmetry is allowed[3]. A sample collection of $D = 50$ directions is shown in Figure 3.1.

In the design of a practical binaural display, the choice of $D = 50$ is a reasonable compromise between flexibility and complexity. Interpolation and dynamic changes are more smooth if the distance between neighboring HRTFs is small. On average, with $D = 50$, each direction has neighbors at approximately 30° in all directions. In this case, simple weighted-average interpolation can be used to render a wide variety of acoustic phenomena. We have found that a more sparse collection of directions is more prone to interpolation and dynamic artifacts[4]. On the other hand, a more dense collection appears to offer little benefit and only adds to the complexity of the system. Furthermore, in Chapter III we found that state-space systems offered the greatest relative advantage for $D \approx 50$, as can be seen in Figure 3.4. Accordingly, we expect to see a clear advantage with the state-space approximants in discrimination performance.

[3] This constraint is included so as to force the collection of directions to be as ‘diverse’ as possible within the HRTF dataset. A decrease in the ‘diversity’ of the collection of HRTFs may improve the performance of an state-space approximant, but not an array of independent filters. Allowing symmetry in the collection may improve the relative performance of the state-space approximant.

[4] While the stimuli in the first two experiments are simple enough that interpolation and dynamics are not an issue, the third experiment considers complex moving sound sources. In the interest of estimating order thresholds that can be used in practice, we fix the number of directions at $D = 50$ for all three experiments.

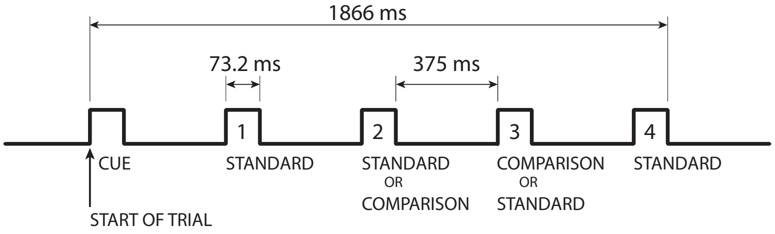


Figure 4.1: The timeline of each trial. The total duration of each observation interval is 73.2 ms. The stimulus consists of 50 ms noise bursts, with 23.2 ms of silence appended to the end of each burst prior to HRTF filtering.

Psychophysical Procedure

A four-interval, two-alternative forced-choice paradigm is used, in which the order N remains fixed during the block of trials, similar to (64). The timeline of a trial is shown in Figure 4.1. A trial consists of four intervals separated by 375 ms of silence. For each trial, either the second or third interval is the 'comparison' stimulus that is generated with the approximant, and the remaining three intervals are 'standard' stimuli that are generated with the full-order HRTFs. The interval for the comparison is randomized across trials. Following the conclusion of the fourth interval, the observer is asked to select which interval (either second or third) is most different from the other three. The user responds by selecting the appropriate button in the experiment's graphical user interface (GUI). A screen-shot of the GUI is shown in Figure 4.2. After the observer response, feedback is provided by lighting the GUI button that corresponds to the comparison stimulus for 100 ms. During the presentation of the trial, each observation interval is indicated by a light in the appropriate button in the GUI. In addition to the four observation intervals, a 'cue' light is provided to inform the listener that a new trial is about to begin. After each trial, there is a 2 s pause and the next trial begins automatically.

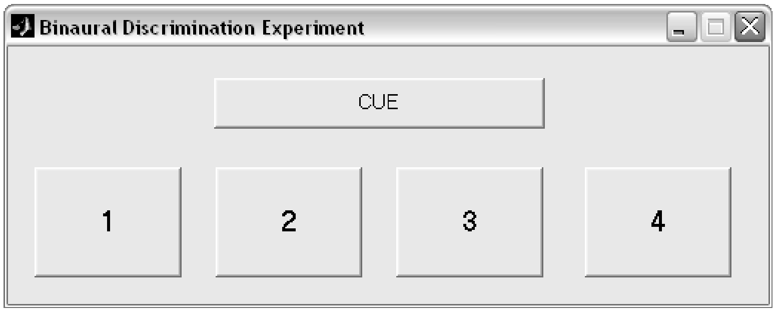


Figure 4.2: Screen shot of the GUI used for experiments 1 and 2.

Trials are presented in blocks of 54. The first four trials of each block are practice: the interval containing the approximant is indicated and the observer response is not recorded. Observer responses are recorded for the last 50 trials. Each block takes about 5 minutes to complete, and breaks are taken regularly after every 5-8 blocks.

The ten stimulus conditions are divided into four groups for testing based upon the noise stimuli variable and the HRTF dataset variable. The first and second groups use individualized HRTFs, and the first and third groups use identical noise instances. Within each group, the order in which blocks are presented, across both system type and system order, is randomized. For each stimulus condition and order, a total of five blocks are presented. Hence each point in each psychometric function is the result of 250 responses. The first two groups consist of 70 blocks each, and the last two groups consist of 105 blocks each. In total, 350 blocks are presented, and 17,500 responses are recorded.

Participants

One observer with normal hearing, the author, participated in Experiment 1. Individual HRTFs had been previously measured for the this observer. The entire experiment required about 28 hours of listening time.

Facility

Stimuli were presented over headphones with the participant seated in a double-walled sound-proof booth (Acoustic Systems, model 19460A) before an LCD monitor, keyboard and computer mouse. The booth is manufactured by Acoustic Systems, model 19460A. The experiment was run in MATLAB on an Apple G5 desktop. The observer indicated their responses by making button presses in the GUI using the computer mouse. The computer chasis was housed outside the booth.

Stimuli were presented at a 44.1kHz sampling rate using an M-Audio FireWire410 interface for D/A conversion and headphone amplification. Beyer-Dynamic DT931 headphones were used throughout the experiments. Stimuli were not equalized for the headphone transfer functions. All stimuli were presented at comfortable listening levels.

4.1.2 Results

Figures 4.3 through 4.6 show the psychometric functions for the ten conditions described above. In all cases, the psychometric functions that correspond to the 'identical' condition are shown with solid lines, and the functions that correspond to the 'independent' condition are shown with dashed lines. The FIR conditions are marked with circles (●), the SIMO state-space conditions are marked with left-pointing triangles (◀), and the MISO state-space conditions are marked right-pointing trianlges (▶).

Each point of each psychometric function shows the fraction of the 250 trials for which the observer made the correct decision. The error-bars for each point show the standard error, which is given by the standard deviation of the observer responses normalized by the square-root of the total number of trials. The standard error gives

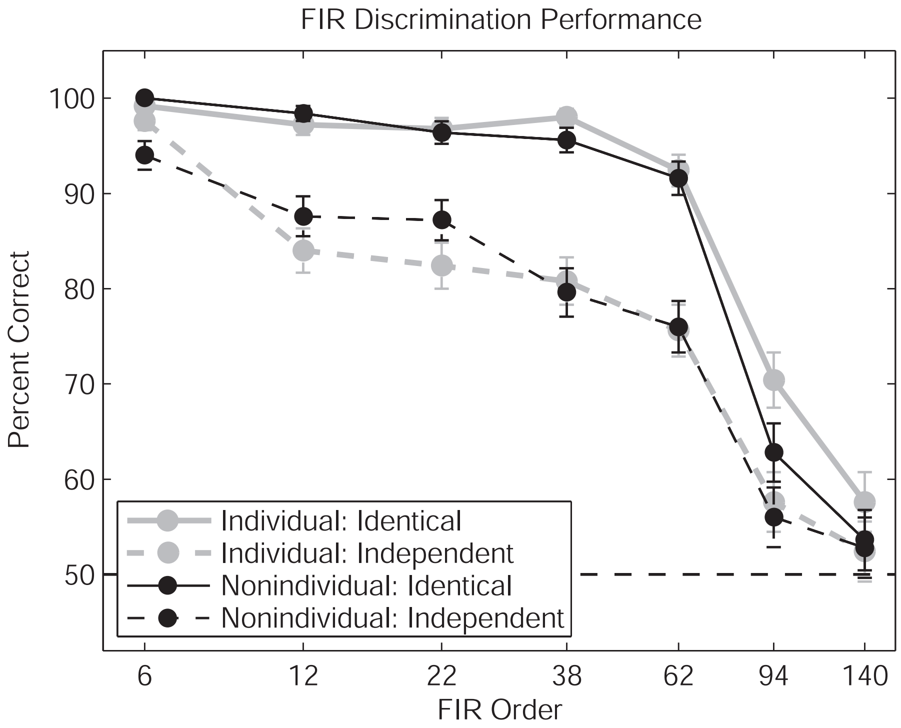


Figure 4.3: Psychometric functions for FIR approximants of both individualized and non-individualized HRTFs as a function of filter order for 'identical' and 'independent' noise instances.

the 68% confidence interval of the discrimination performance[5].

The psychometric functions for the truncated FIR approximants are shown in Figure 4.3. The abscissa plots the system order (N) on a logarithmic scale, and the ordinate plots the observer's percentage correct. The individualized HRTF conditions are shown with gray lines, and the non-individualized HRTF conditions are shown with black lines. For both identical and independent noise stimuli, there was no significant difference in discrimination performance for individualized versus non-individualized HRTFs. In contrast, there was a significant difference in discrimination performance for identical versus independent noise instances. For a wide

[5] If the experiment is repeated with the same number of trials, there is a 68% chance that the discrimination performance would be within one standard error of the estimate.

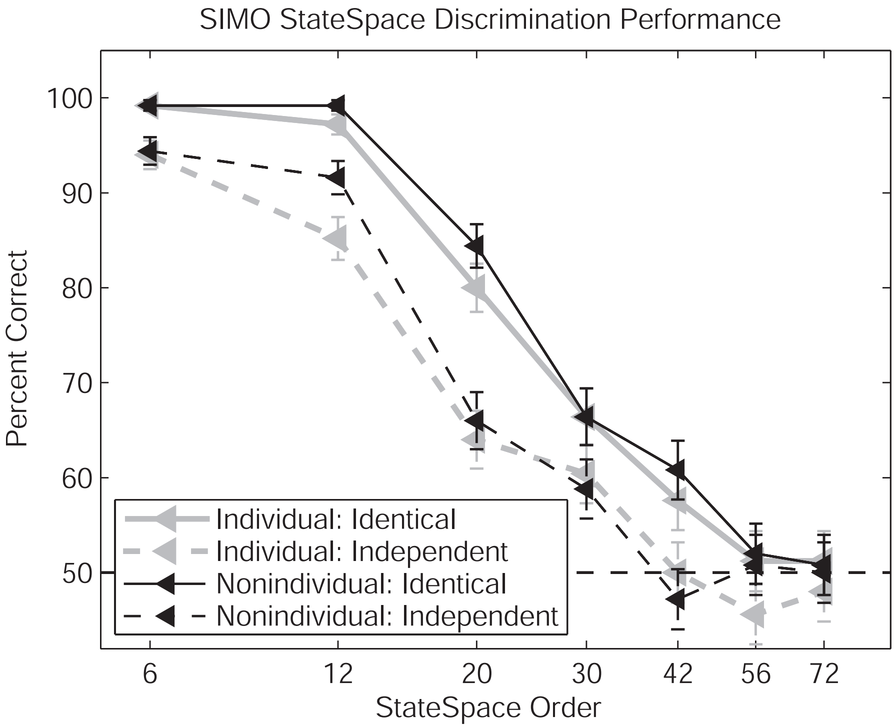


Figure 4.4: Psychometric functions for SIMO state-space approximants of both individualized and non-individualized HRTFs as a function of system order for identical and independent noise instances.

range of system orders, $12 \leq N \leq 62$, switching from identical to independent noise stimuli caused the discrimination performance to drop from over 95% to about 85%. However, for both noise stimuli, the psychometric functions did not fall to chance until the order was increased to about 150.

The psychometric functions for the SIMO state-space approximants are shown in Figure 4.4 using the same plotting convention as in Figure 4.3. As with the FIR approximants, there was little difference between the individualized and non-individualized HRTF conditions, although discrimination accuracy was slightly higher for the non-individualized condition than for the individualized condition. The noise stimuli affected the discrimination performance significantly. For both conditions, the shapes of the psychometric functions were the same, and only differed by a shift

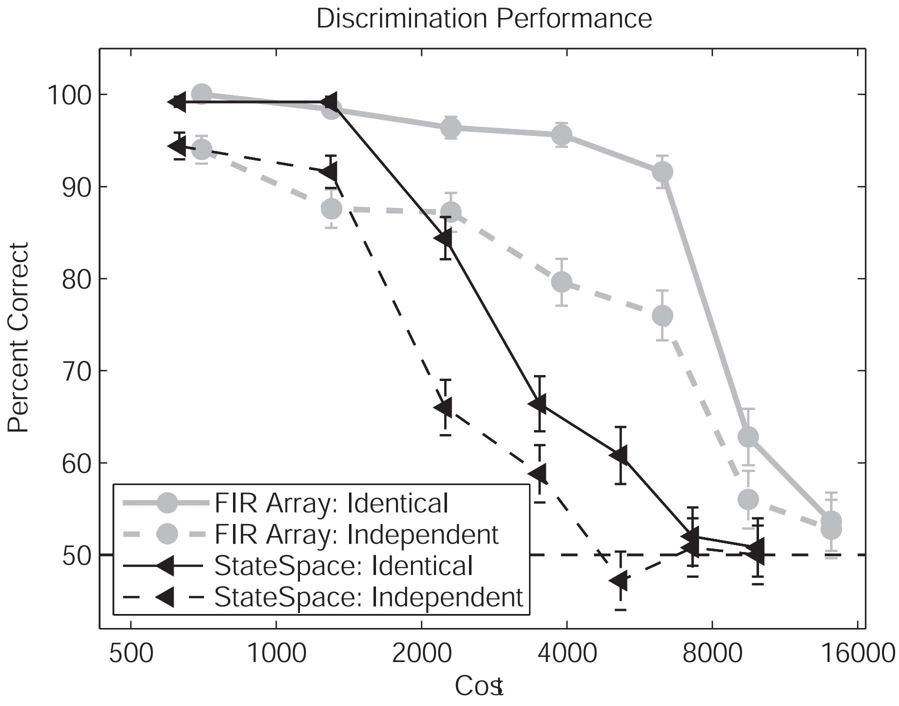


Figure 4.5: Psychometric functions for FIR and SIMO state-space approximants of non-individualized HRTFs as a function of net system cost.

in the independent variable, N. For the state-space approximants, the order at which the psychometric functions fall to chance depended on the noise stimuli: for identical stimuli, discrimination fell to chance at $N = 56$, whereas for independent stimuli, discrimination fell to chance at $N = 42$.

The discrimination performance of the FIR and state-space approximants can be directly compared if the psychometric functions are plotted as functions of net computational cost, rather than system order. The psychometric functions for the non-individualized conditions in Figures 4.3 and 4.4 are shown in Figure 4.5 as a function of computational cost. The cost for both systems is defined in Section 3.1.2. In Figure 4.5, the abscissa gives the net computational cost C on a logarithmic scale. It is clear that, for increasing computational cost, the discrimination performance fell to chance more rapidly with state-space approximants than FIR approximants.

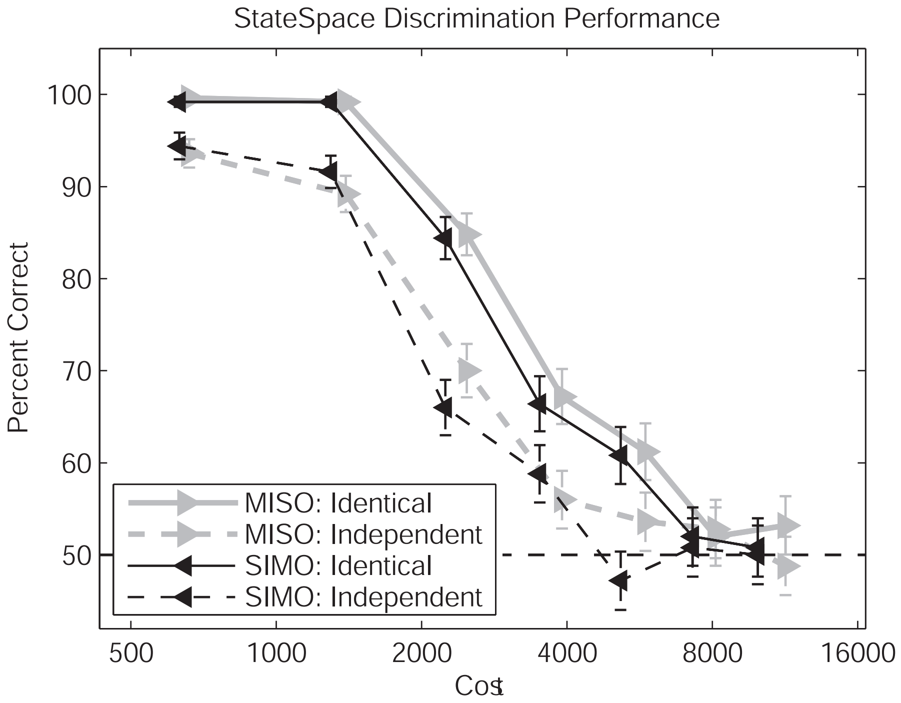


Figure 4.6: Psychometric functions for SIMO and MISO state-space approximants of non-individualized HRTFs as a function of net system cost for identical and independent noise instances.

For example, in the case of independent noise stimuli, discrimination performance fell below 60% for $C > 3500$ for state-space approximants, and for $C > 8000$ for FIR approximants.

The psychometric functions for the SIMO and MISO state-space systems, for the non-individualized condition only, are shown in Figure 4.6. The abscissa gives the net computational cost C on a logarithmic scale. Overall, discrimination performance was similar for both architectures for both noise stimuli. Discrimination accuracy was 2-4 percent higher for the MISO architecture than the SIMO architecture, which is in general agreement with the relative approximation errors shown in Figure 3.4.

4.1.3 Discussion

Experiment 1 used a fixed-level psychophysical procedure. For every point estimated along the psychometric function, a large number trials were required. Fortunately, Experiment 1 revealed several trends that enable the design of an adaptive-level experiment. All ten psychometric functions were reasonably smooth and monotonic in the independent variable. Furthermore, in the transition region from 90% correct to 60% correct, the psychometric functions were reasonably linear on a logarithmic scale. This is a convenient feature for an adaptive-level experiment, as the adaptive system orders may be spaced logarithmically, allowing a wide range of discrimination thresholds to be accurately and efficiently estimated.

Figures 4.3 and 4.4 demonstrate that discrimination performance was not degraded by the use of non-individualized HRTFs. There is no significant difference in discrimination performance between individualized and non-individualized conditions. Indeed, for the non-individualized HRTF condition, discrimination accuracy is slightly higher than for the individualized condition. Measuring complete HRTF datasets from multiple participants is costly and time-consuming. In lieu of individual HRTF measurements, we use the non-individualized HRTF dataset from Experiment 1 in Experiments 2 and 3.

From Figure 4.6, it is clear that the choice of SIMO versus MISO architecture had little impact on the discrimination performance, as a function of cost C. If the MIMO architecture had been included, it is likely that discrimination accuracy would have been significantly higher due to additional cues caused by the ITD smearing, as discussed in Section 2.3.1 (9). Comparing the SIMO and MISO discrimination performance, the MISO architecture yielded consistently higher discrimination accuracy, although the difference was small. This mirrors the approximation error results reported in Section 3.2.2, where the MISO architecture exhibited slightly higher auditory $\mathcal{L}^2$ error.

In Figure 4.5, contrasting trends can be observed in the change in psychometric function for identical and independent noise stimuli. In moving from identical noises to independent noises, additional stimulus *uncertainty* was introduced. This uncertainty reduced the reliability of timbral cues in performing discrimination. As expected, in the absence of such reliable cues, the discrimination performance degraded, and the psychometric functions shifted to lower orders. This shift was different for the state-space versus FIR approximants. For the state-space approximants, independent noise instances caused the psychometric function to shift to lower orders without changing the slope of the function. For the FIR approximants, however, the psychometric function changed slope. For the identical condition, discrimination accuracy was near 100% for $N \leq 62$, whereas for the independent condition, discrimination accuracy fell to the 80-90% range for $12 < N < 62$. However, for both conditions, the psychometric functions exhibited 'knees' near $N = 62$. This suggests that for a broad range of system orders, $12 < N < 62$, for 20-40% of trials, discrimination was performed primarily using subtle timbral cues that were unavailable when independent noise instances were used.

The observer noted subjectively that discrimination was performed using primarily timbral cues and not spatial cues for $N \geq 12$ for both state-space and FIR approximants[6]. Indeed, the observer noted that the task was perceptually demanding and required careful listening and concentration. For both FIR and state-space approximants, the $N = 6$ comparison stimuli often were 'collapsed inside the head'

[6] During the experiment itself, the observer did not know the order N for any block. However, as the author was developing the experiment, numerous informal tests were conducted in which the observer knew the system order.

relative to the full-order stimuli. For $N = 12$ both approximants yielded small spatial distortion that aided in discrimination. This distortion was larger for FIR approximants than for state-space approximants. For $N > 12$, both approximants yielded perceived sources at the correct location, and were equally well-externalized and well-focused.

In the next experiment, we utilize our results from Experiment 1 to design efficient experimental procedures that allow us to estimate perceptual thresholds for several conditions and listeners.

4.2 Experiment 2: Order Thresholds for Broadband Noise

Experiment 1 examined the relationship between perceptual fidelity and system order by directly sampling the psychometric function at several fixed orders. As can be seen from the experimental procedures, this required a very large number of trials. Having discovered that the psychometric functions are monotonic and linear allows us to use an adaptive-level stair-case procedure (146; 147) for Experiment 2. Furthermore, because Experiment 1 showed no difference in discrimination performance with individualized and non-individualized HRTFs, we employ only non-individualized HRTFs and recruit additional observers for Experiments 2 and 3. For Experiment 2, rather than estimate complete psychometric functions, we estimate one point on each psychometric function that corresponds to a fixed discrimination performance.

Some of the stimulus conditions from Experiment 1 are replicated in Experiment 2, and new stimulus conditions are introduced as well. The stimulus conditions vary the type of uncertainty of the stimulus. Two types of stimulus uncertainty are varied: a timbral uncertainty and a spatial uncertainty. The manipulation of timbral uncertainty is the same as found in the Experiment 1, through the use of statistically identical versus statistically independent noise bursts. The manipulation of spatial uncertainty in Experiment 2 is accomplished by the introduction of virtual reflections in the stimulus. By manipulating two types of stimulus uncertainty we can assess how robust the advantage of the state-space approach is, as well as compare the relative trends between the two types of uncertainty.

4.2.1 Methods

Stimulus Conditions

Stimuli are constructed in the same manner as Experiment 1, with the exception that some stimulus conditions include virtual acoustic reflections. Two types of stimulus uncertainty are varied, timbral and spatial. The timbral uncertainty is the same as in Experiment 1, 'independent' versus 'identical' noise bursts. The spatial uncertainty is introduced by using the first-order images sources of a medium-sized rectangular room (101) to model a simple 'reflective' environment, in contrast to the 'anechoic' environment that is effectively modeled by filtering the monaural noise burst with a single HRTF pair. For all four stimulus designs, both FIR and state-space approximants are tested. In total there are three experimental variables and eight stimulus conditions.

The 'anechoic' condition is the same as in Experiment 1: for each trial, one of the $D = 50$ directions is chosen, and the appropriate HRTF and approximant is used to compute the binaural stimulus. The 'reflective' condition includes five first-order acoustic reflections, along with the direct wave, which model a rigid-walled rectangular enclosure. The image source model was used to calculate the first-order reflections of a rectangular room of size 6m wide, 8m deep, and 4m tall. The reflection coefficients are $\beta = 0.7$ for the walls and ceiling and $\beta = 0$ for the floor (101). Note that the lowest elevation included in the HRTF dataset used in this study is $\phi = -36°$. As such, rendering reflections from the floor is not possible, hence the floor is modeled as entirely absorptive. Figure 1.8 shows a sample arrangement of image sources for a rectangular room.

The listener is positioned at the center of the room, and faces forward. For each trial, a source location is chosen from anywhere in the room with uniform probability subject to the constraints that the source not be within 1m of any boundary or the listener, and that the elevation of the source relative to the listener is at least $-36°$. For a given source location, five image sources (excluding the image below the floor) are computed, giving a total of six sources at various directions and distances[7]. Nearest-neighbor interpolation is used, which is mathematically equivalent for first-order image sources to tilting the walls of the room to be not perfectly rectangular. Hence the stimulus presented to the listener is the sum of six scaled-and-delayed binaural signals, where each binaural signal is the result of filtering the source noise with either an HRTF pair (for the standard stimulus) or an approximant (for the comparison stimulus). Note that only first-order reflections are included, and no diffuse reverberation is included.

The inclusion of the virtual reflections introduces a sort of spatial uncertainty into the stimulus. In most cases that included virtual reflections, the listener reported that they perceived the source direction with reasonable accuracy, but that the diffuseness of the sound was greater and the spatial cues are conflicted as compared to the ‘anechoic’ condition. Given that most commonplace auditory scenes include acoustic reflections, estimating discrimination thresholds for reflective environments is an appropriate step in setting the design specifications for immersive and flexible

[7] The distance between the listener and each source (either direct or image) is given by the distance between the source and the closer ear.

binaural displays. However, it is unclear what, if any, effect the inclusion of virtual reflections has on discrimination performance.

The SIMO architecture is used for all state-space approximants in Experiment 2. The stimuli in Experiment 2 consist of a single virtual sound source, and depending upon the stimulus condition, a collection of image sources. Recall that the drawback of the SIMO architecture is that only one source can be rendered, however for the stimuli in Experiment 2 this is not a problem. In both Experiment 1 and the empirical experiment in Chapter III, the SIMO architecture yield slightly better performance than the MISO architecture. In the interest of showing a clear difference in the threshold estimates for FIR and state-space approximants, we chose the SIMO architecture for the state-space approximants. Nonetheless, the difference in performance for the SIMO and MISO architectures was very small, hence we expect that the results would be little changed if we had used the MISO architecture instead. Furthermore, in Experiment 3 we consider stimuli that consist of multiple distinct sound sources, hence the MISO architecture is required for Experiment 3.

Psychophysical Procedure

The observer performed a four-interval, two-alternative forced-choice task in which the order of the approximant was adjusted according to their responses. A "2 down, 1 up" stepping rule (146) was used in which the order of the approximant increases whenever the observer was correct on the previous two trials, and decreases whenever the observer was incorrect on the previous trial[8]. The "2 down, 1 up" stepping rule has been shown to track the 70.7% correct point of the listener's

[8] The name of the stepping rule derives from classical 'signal + noise' discrimination experiments. In these experiments, 'down' refers to a decrease in the level of the 'signal' relative to the 'noise'. In the present experiment, the 'signal' is the distortion due to the filter approximation. Increasing the approximant order corresponds to a 'down' step and decreases in the 'signal'.

psychometric function.

Each block, or run, of the adaptive-level experiment begins with an $N = 6$ approximant, and ends after the observer has made 12 *reversals*. A reversal occurs on a given trial when the order update changes direction, either from increasing to decreasing or vice versa. The first four reversals are ignored, and the approximant order is estimated from the levels of the last eight reversals. The number of trials in each block varied, but is usually 50-60 trials, and the duration of each block is about 5 minutes. The observer takes a break after every four blocks.

From Figures 4.3 and 4.4 it can be seen that the psychometric functions are approximately linear functions of the logarithm of N in the transition region from 90% to 60% correct. Therefore a logarithmic spacing of approximant orders is appropriate. Logarithmic spacing also allows for a wide range of thresholds to be accurately measured. For the FIR approximants, the order increases in increments of 19%, and for the state-space approximants, the order increases in increments of 16% (rounded to the nearest integer). The increment sizes were chosen so that both approximants included 5-7 steps in the transition region, and yield blocks with approximately the same number of trials for all stimulus conditions. The specific system orders are

$$\begin{aligned} N_{FIR} &= [6, 7, 8, 9, 11, 13, 16, 19, 23, 27, 32, 39, \\ &\quad 46, 55, 66, 79, 94, 112, 134, 159, 190, 226] \\ N_{SS} &= [6, 7, 8, 9, 10, 12, 14, 16, 18, 21, \\ &\quad 24, 28, 32, 37, 42, 49, 56, 65, 74, 85, 98] \end{aligned}$$

The standard “2 down, 1 up” stepping rule was modified to accommodate specific features of the psychometric functions. We want the adaptive blocks to start near

100% discrimination accuracy, so that the observer almost always makes correct responses for the first few trials. As such, we choose to begin each adaptive run with order $N = 6$ for both FIR and state-space approximants. However, this slows the convergence to the 70.7% discrimination point, as the psychometric functions decrease slowly from 100% to 80%, but then rapidly from 80% to 60%. This trend is strong for the FIR approximants in particular. Preliminary tests showed that observers begin to make occasional errors for $N > 6$, hence reversals begin while the order is still far from the true 70.7% point. The order eventually converges to 70.7%, but requires up to 50 trials to step into the 70.7% neighborhood using the "2 down, 1 up" rule as described above.

To speed the convergence, two modifications were made to the stepping rule prior to the first two reversals. First, the step size is doubled. Second, a "2 down, 2 up" rule is employed in which a reversal from 'down' to 'up' occurs only after the observer has made two incorrect responses. This stepping rule converges at 50% discrimination, hence the adaptive track would tend to overshoot the 70.7% threshold if this stepping rule were used throughout the block. However, this rule is only applied to the first pair of reversals. The standard "2 down, 1 up" rule is used for the last ten reversals, and only the last eight reversal orders are recorded. We found that the adaptive track occasionally overshot the 70.7% threshold prior to the first reversal, but that this did not appear to bias the final eight reversals when comparing the author's results in Experiment 1 with preliminary results for Experiment 2.

A typical adaptive track is shown in Figure 4.7. Note that in trials 3 and 7 the observer made an incorrect response, even though the order of the approximant is much smaller than the 70.7% threshold. If the "2 down, 2 up" stepping rule had not been used for the first pair of reversals, reversals would be recorded before the track

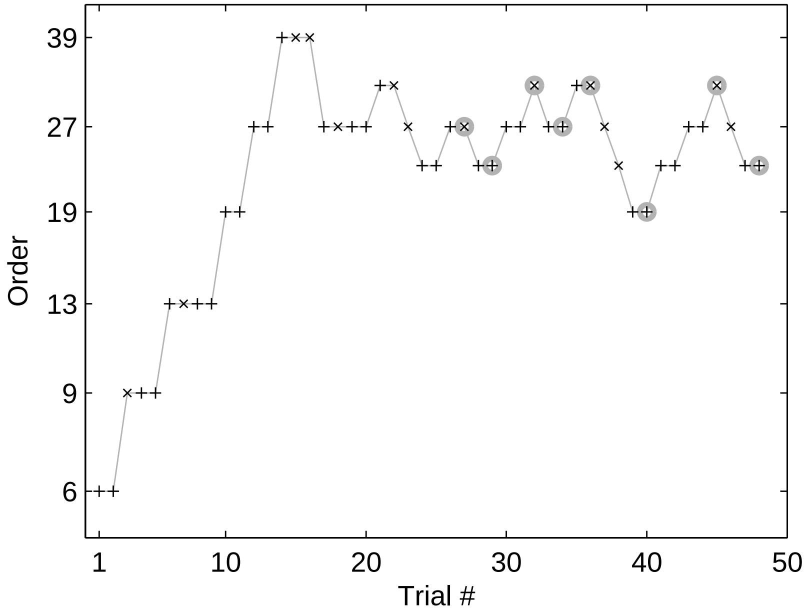


Figure 4.7: A sample adaptive track with an FIR approximant. Correct responses are marked with "+" signs, incorrect responses are marked by "×" signs, and the recorded reversals are marked with gray circles.

converged on the 70.7% threshold. Of course, the "2 down, 2 up" stepping rule allows the possibility of the observer stagnating, or plateauing, if the observer responds "one correct, one incorrect, one correct..." However, in practice such instances were rare during this experiment.

The eight stimulus conditions are divided into four groups of increasing net uncertainty. The first two groups test identical noise instances (no timbral uncertainty), and the first and third groups test anechoic environments (no spatial uncertainty). For each group, blocks were chosen at random with either FIR or state-space approximants. Each participant completed enough blocks to have a total of five 'stable' blocks for each condition, as defined below. For each group, participants performed fixed-level training prior to the adaptive tests. For the first group, participants performed about 3 hours of fixed-level training, and for the remaining groups participants performed about 1 hour of fixed-level training. The procedures for the fixed-level training follow those of Experiment 1.

Statistical Treatment

The last eight reversals of each block are recorded to estimate the order that corresponds to 70.7% correct in the discrimination task. Because the approximant orders are logarithmically spaced, the unbiased threshold estimate for each block is computed by taking the geometric mean of the eight reversals. Each observer performed enough adaptive blocks so as to have five 'stable' blocks for every stimulus condition, where a 'stable' block is defined as a block for which the standard deviation of the eight recorded reversals is less that 2 steps of the independent variable[9]. The geometric mean of the five 70.7% order threshold estimate are reported below, along with the standard deviation of the five estimates.

Participants

Five observers participated in the adaptive-level experiment, including the author. Four participants were recruited by email from the undergraduate program in the the Department of Performing Arts Technology in the School of Music at the University of Michigan, Ann Arbor. These four participants had experience with sound engineering and headphone listening, although none had prior experience with binaural listening or had participated in psychoacoustic studies. Among the participants, four were male, and one was female. All participants were between the ages of 19 and 30, and have 'normal' hearing as assessed by an audiogram measurement[10]. The entire experiment required about 14 hours of listening time from each partici-

[9] Unstable blocks were usually blocks in which the observer did not step up to the 70.7% neighborhood before reversals were recorded, in spite of the initial "2 down, 2 up" stepping rule. In such cases the eight recorded orders drifted up significantly from start to finish. Overall, approximately 80% of the blocks were acceptable.

[10] All five participants had audiograms that were within 4 dB of the lab average between 500 Hz and 4 kHz, and were within 10 dB of the lab average between 125 Hz and 16 kHz.

pant, divided into 1.5-2 hour sessions. Observers performed one session per day, and most sessions were performed on consecutive days.

4.2.2 Results

Figures 4.8 and 4.9 show the average $N_{70.7}$ threshold estimates for the eight conditions for each observer individually, and averaged across observers. The observers are labeled “S1” through “S5” along the abscissa. Observer S1 is the author. FIR thresholds are shown with light gray bars and state-space thresholds are shown with dark gray bars. For the individual observer thresholds, the error bars indicate the standard deviation of the five blocks. For the average thresholds, the error bars indicate the standard deviation of the five observer averages.

Threshold estimates for identical noise instances are shown in Figure 4.8. The top panel shows thresholds for the anechoic condition, and the bottom panel shows thresholds for the reflective condition. Thresholds for the FIR approximants were in the range $48 \leq N_{70.7} \leq 92$, and thresholds for the state-space approximants were in the range $16 \leq N_{70.7} \leq 32$. Both approximant types yielded thresholds that vary by a factor of two across observers. Nonetheless, it is clear that the state-space approximants yielded significantly lower thresholds for each observer.

Threshold estimates for independent noise instances are shown in Figure 4.9, for both the anechoic condition (top panel) and the reflective condition (bottom panel). Thresholds for the FIR approximants were in the range $13 \leq N_{70.7} \leq 76$, and thresholds for the state-space approximants were in the range $9 \leq N_{70.7} \leq 23$. For each observer, the thresholds with state-space approximants were consistently lower than those with FIR approximants.

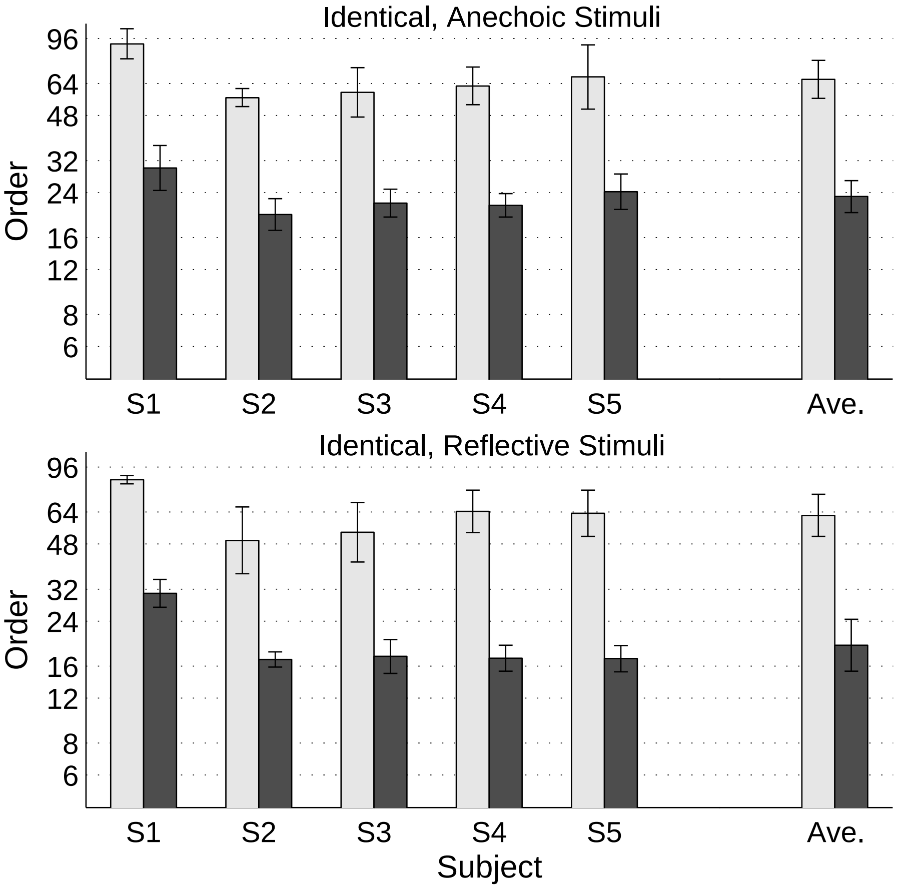


Figure 4.8: 70.7% thresholds for identical noise instances. Thresholds for all 5 participants are shown, along with average thresholds to the far right. Thresholds for the anechoic condition are given in the top panel, and for the reflective condition in the bottom panel. Thresholds for the FIR approximants are shown with light gray bars, and thresholds for the state-space approximants are shown with dark gray bars. The standard deviation of the five threshold estimates for each condition is indicated with errorbars.

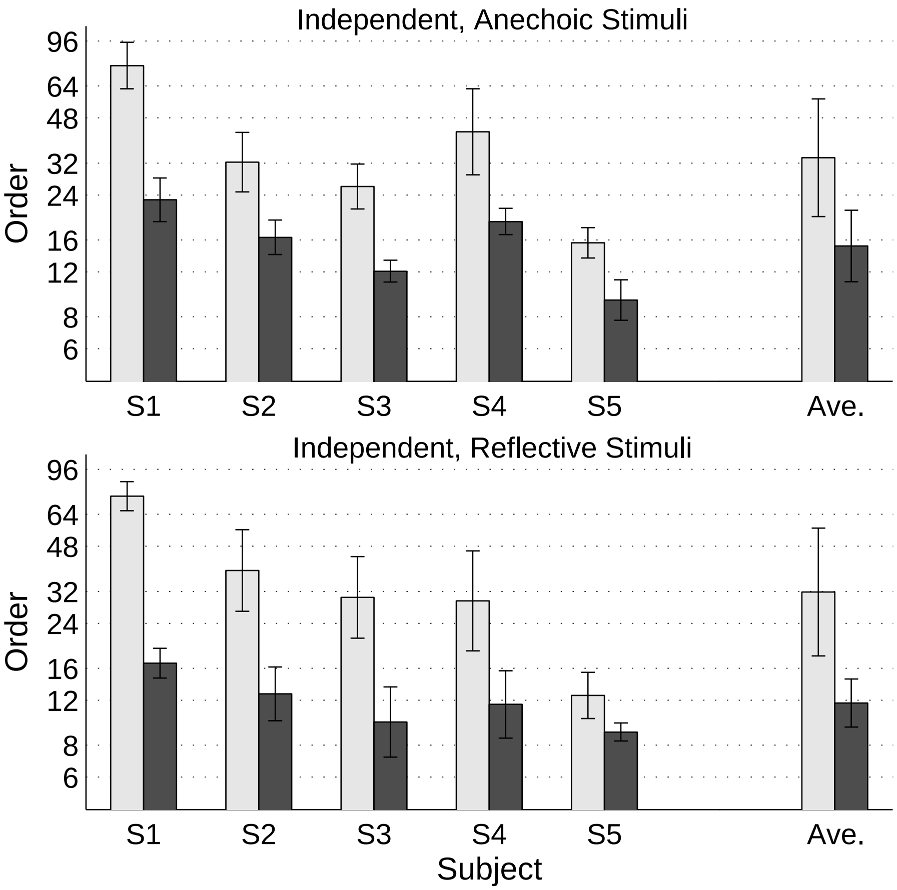


Figure 4.9: 70.7% thresholds for independent noise instances.

4.2.3 Discussion

The results above exhibit several interesting trends. The most obvious is that the $N_{70.7}$ thresholds for the state-space approximants are significantly lower than those for the FIR approximants. This is clear for all observers and for all stimulus conditions. Other significant trends are evident as well, but require additional interpretation. A preliminary discussion of several important points is given below, while a detailed analysis of the average $N_{70.7}$ trends with respect to stimulus uncertainty is postponed until section 4.4.

Analysis of Variance

The statistical significance of the $N_{70.7}$ threshold estimates was assessed with a four-factor, repeated measures ANOVA. The four factors were: observer (S1-S5), approximant type (FIR and state-space), noise stimulus (identical and independent), and stimulus environment (anechoic and reflective). The first three factors were found to be clearly significant ($p < 0.0001$), but the environment factor was not ($p = 0.106$). ANOVA was also performed separately for the FIR and state-space conditions, in which case the environment factor was significant with state-space approximants ($p < 0.0001$), but not with FIR approximants ($p = 0.433$). This disparity is discussed in greater detail in section 4.4. The p values change little if the observer factor is removed from the ANOVA.

Individual Differences

Observer S1 was the author. Observer S1 yielded significantly higher thresholds than the other observers for all stimulus conditions. This was likely due to the extended training that this observer received while designing and performing Experiment 1. Nonetheless, the relative trends for different stimulus conditions were similar for S1 and the other observers.

The $N_{70.7}$ estimates for observer S1 were in agreement with the psychometric functions estimated in Experiment 1, indicating that the adaptive procedure used in Experiment 2 was unbiased. For FIR approximants, observer S1 yielded $N_{70.7} = 92$ and $N_{70.7} = 76$ for identical and independent noise instances, respectively. Referring to Figure 4.3, the psychometric functions cross the 70.7% point at approximately $N = 94$ and $N = 70$ for the corresponding conditions. For the state-space approximants, observer S1 yielded $N_{70.7} = 30$ and $N_{70.7} = 23$ for identical and independent noise instances, respectively. Referring to Figure 4.4, the psychometric functions cross the 70.7% point at approximately $N = 28$ and $N = 19$ for the corresponding conditions.

Considering all five observers, the threshold estimates showed more observer variation with independent noise instances than with identical noise instances. This suggest that the timbral uncertainty introduced in the independent noise condition degraded the discrimination performance of some observers more than others. The drop was greatest with observer S5, who exhibited higher thresholds than most other observers in the identical condition, but exhibited the lowest thresholds among all participants in the independent condition. Indeed, observer S5 reported subjectively that the discrimination task with independent noise instances was difficult even with very low-order approximants. Overall, the individual differences seen here are consistent with studies in spectral-shape discrimination (148).

Subjective Observations

All five observers reported subjectively that the discrimination task was more 'difficult' with independent noise instances than identical noise instances. The observers were instructed to select the interval that was *most* different from the other three. For the identical condition, this selection was essentially equivalent to selecting the interval that was simply different, as the other three intervals were physically identical. In contrast, for the independent case, all four intervals were physically different and the observer had to weigh the differences so as to select the interval that was most different.

Furthermore, three observers reported subjectively that the discrimination task was more 'difficult' with reflective stimuli than anechoic stimuli. No observers reported that the discrimination task was easier with reflective stimuli. In the case of reflective stimuli, observers generally reported that the sound objects were better externalized, but also more diffuse and 'enveloping' and that directional cues were more ambiguous.

Order Versus Cost

Recall that, in general, there is a difference between the relative system orders and the relative computational cost of the two approximant types. As defined in Section 3.1.2, the total cost C of the FIR approximant is proportional to the order N, whereas for the state-space approximants the relationship between C and N is quadratic. Nonetheless, for the order thresholds $N_{70.7}$ that were estimated in this experiment, the relative advantage of the state-space approximants would change little if expressed in terms of cost $C_{70.7}$ instead.

Consider two systems that model $D = 50$ HRTF pairs, and FIR system and a SIMO state-space system. Figure 4.10 shows the ratio of the computational cost of the state-space system to that of the FIR array of *equal order* for $1 \leq N \leq 50$. For $N > 50$, a state-space system yields significantly higher cost than an equal-order FIR array. However, the highest order threshold estimated for state-space approximants in Experiment 2 was $N_{70.7} = 32 < D$, hence the $\mathbf{C}$ matrix dominates the computational cost of the state-space system and the $\mathbf{A}$ and $\mathbf{B}$ matrices contribute little to the net cost. At $N = 32$, a SIMO state-space system yields a computational cost that is about 15% larger than the cost of an FIR array of equal order. At $N = 12$, a SIMO state-space system yields the same computational cost as an FIR array of equal order.

To relate this back to the results presented in Figures 4.8 and 4.9, consider how the relative differences between the FIR and state-space approximants would change

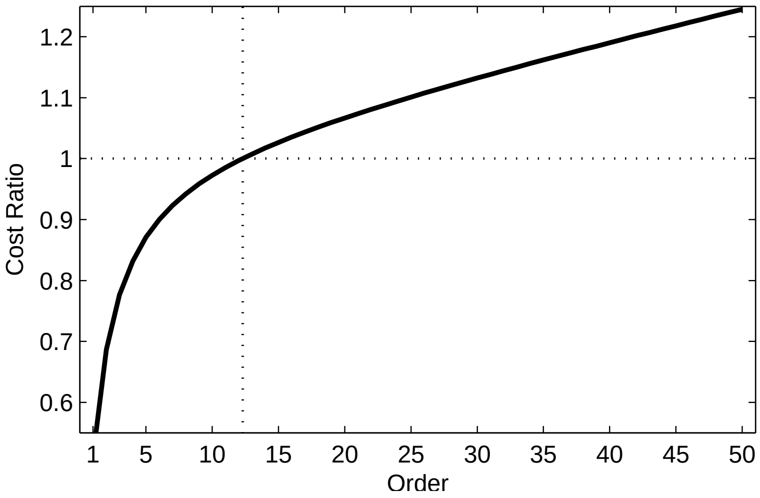


Figure 4.10: Ratio of the computational cost of a SIMO state-space system to an FIR array of equal order N. Both system model $D = 50$ HRTF pairs.

if the Figures showed $C_{70.7}$ instead of $N_{70.7}$. For example, the ratio of $N_{70.7,\mathrm{FIR}}$ to $N_{70.7,\mathrm{SS}}$ for the "identical, anechoic" condition is 2.91, whereas the ratio of $C_{70.7,\mathrm{FIR}}$ to $C_{70.7,\mathrm{SS}}$ is 2.60. This is the largest change of any of the conditions considered in Experiment 2. For the "independent, reflective" condition, the ratio of orders is 2.67 whereas the ratio of costs is 2.55. Indeed, all of the relative trends evident in the order thresholds can also be observed in the cost thresholds.

4.3 Experiment 3: Order Thresholds for Auditory Scenes

In Experiment 2, a variety of stimulus conditions were explored and a robust procedure was used to estimate order thresholds $N_{70.7}$. However, all conditions, even those with increased stimulus uncertainty, consisted of individual wideband noise bursts. Discrimination is relatively easy with such stimuli, hence the order thresholds may be unduly pessimistic. In everyday listening, sound sources are spectrally complex and temporally nonstationary, may be narrowband. Furthermore, typical auditory scenes consist of many such sound sources acting simultaneously. In the final experiment we present listeners with auditory ‘scenes’ and determine the lowest approximant order without causing a discriminable loss in perceptual fidelity.

Experiment 3 considers stimuli that are complex virtual auditory scenes (VAS) consisting of commonplace sound sources. The objective adaptive-level procedure employed in the previous experiment is replaced with a ’method-of-adjustment’ search for the threshold of subjective equality. The observer is given control of the “Approximation Quality,” which is simply the order N of the approximant, and instructed to find the poorest “Approximation Quality” that yields an “Approximate VAS” that is identical to the “Ideal VAS.” The FIR reference approximant that was used in the previous experiments is abandoned in Experiment 3 and only state-space approximants of differing orders are considered below.

4.3.1 Methods

Stimulus Conditions

Three virtual auditory scenes were constructed by the author. Each VAS is 20-25s in duration and consists of several sound sources, some of which occur simultaneously. The sound sources were culled from a Bang and Olufsen collection of anechoic recordings (149), a CBS collection of field recordings (150), and field recordings made by the author. Note that only the sounds from the Bang and Olufsen collection were recorded in anechoic conditions. The other recordings include natural reverberation and modest background noise.

The first VAS consisted of a short trumpet solo in front of the listener followed by several independent applause sources surrounding the listener followed by three firework whistles above the listener. The second VAS consists a speech source and a percussion source on either side of the listener at 0° elevation, and a diffuse recording of tropical birds above the listener. The third VAS consists of several animal sounds, from the bleat of a sheep to the buzz of a flying insect, presented in an overlapping sequence. For the first and third scenes, the virtual listener is stationary within the scene, and the sources move slowly with straight trajectories and constant speed throughout the duration of the scene. For the second scene, the listener is rotating slowly (with a 5s period) and the sources are stationary. Soundfiles of the three virtual auditory scenes are available online[11].

For each VAS, the virtual listener is positioned inside a 6m wide by 8m deep rectangular enclosure. The virtual enclosure has no floor or ceiling, and the vertical walls extend to infinity. For the first two auditory scenes, the reflection coefficient of all four walls is $\beta = 0.7$, and for the third scene the walls to the front and right have a reflection coefficient of $\beta = 0.7$, and for the remaining two walls $\beta = 0.1$. Sound sources are positioned inside the room, and 1st- and 2nd-order reflections are modeled with image sources. However, for sources located far above the listener, such as passing birds, no reflections are included. Finally, no diffuse reverberation is modeled, as reverberation is largely independent of the listener position and is best rendered using a separate subsystem.

The auditory scene is constructed in overlapping 50 ms frames, with 5 ms raised-cosine ramps applied to the beginning and end of each frame[12]. For each frame, the location of the source relative to the listener is updated and the magnitude and delay adjusted according the the distance to the nearer ear[13]. VBAP interpolation (86) is used to map each source direction to the nearest three HRTF directions included in the state-space model that surround the true source direction. In this way, $D = 50$ monaural signals are generated for both the left and right ears, including constant

[11] http://www-personal.umich.edu/ nhadams/auditoryScenes/index.html

[12] Note that the 50 ms frame period is small compared to the speed with which the sources move relative to the listener, such that no source moves more than a few degrees within each frame.

[13] The radius of the virtual listener's head is 8.75 cm.

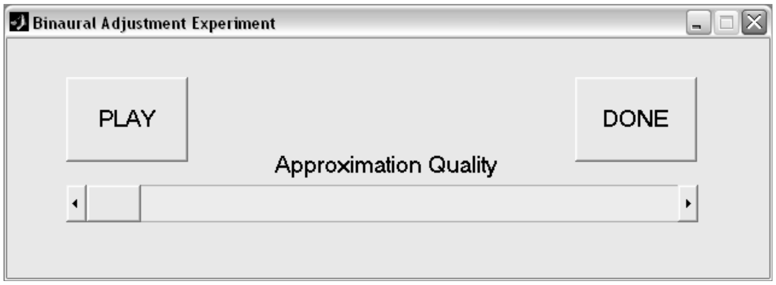


Figure 4.11: Screen shot of the GUI used for Experiment 3.

sample delays corresponding to the ITD. The 2×50 monaural signals are then filtered with the appropriate full-order HRTFs, yielding the “Ideal VAS,” and are also filtered with two order-N MISO state-space approximants, yielding the “Approximate VAS.”

The test stimulus that is presented to the listener is constructed by switching between the “Ideal VAS” and the “Approximate VAS” every 2.4 s. For each switch, non-overlapping 5 ms raised-cosine squared ramps are used. The brief gaps created in the test stimulus due to the switches are obvious, but not distracting.

Psychophysical Procedure

The observer is instructed to use a variable-mapping slider that controls “Approximation Quality” to find the poorest quality setting at which he or she perceives no difference between the “Ideal VAS” and the “Approximate VAS.” The observer need not recognize the “Ideal VAS” segments from the “Approximate VAS” segments in deciding whether to increase the “Approximation Quality”, the observer need only perceive a difference before and after any switch in the test stimulus. A screen-shot of the GUI is shown in Figure 4.11.

Each block of the experiment begins with the observer listening to the full “Ideal VAS” once. The observer then listens to and adjusts the test stimulus as many times as he or she prefers until the poorest approximation is found that yields subjective equality. The observer is allowed to stop or restart the playback of the test stimulus at anytime.

During each presentation of the test stimulus, the approximation quality of the "Approximate VAS" is fixed. Between each playback of the test signal, the observer may adjust the "Quality" of the "Approximate VAS," which is simply the order N of the two MISO systems. The order N is controlled by a slider that has a random exponential warp[14] applied to the mapping between slider position and N for each block of the experiment. The collection of orders N is spaced logarithmically with a step size of 20%. The full set of approximant orders N is

$$\begin{aligned} N &= [1,2,3,4,5,6,7,8,10,12,14,17, \\ &\quad 21,25,30,36,43,51,62,74,89] \end{aligned}$$

The space of possible mappings between slider position and order N is shown in Figure 4.12.

Each block begins with $N = 1$. The observer may take as much time with each block as he or she prefers. When the observer decides that the "Approximate VAS" is identical to the "Ideal VAS", and the "Approximation Quality" slider is not higher than necessary, the observer selects "Done," the order N_{se} corresponding to this "Approximation Quality" is recorded, and a new block is initiated. The three auditory scenes are presented once each in sequence, and the sequence is repeated five times. For each observer a total of fifteen blocks are performed, with breaks taken after every 4-6 blocks.

[14] The random warp is applied to discourage the listener from using a visual reference point to aid the task.

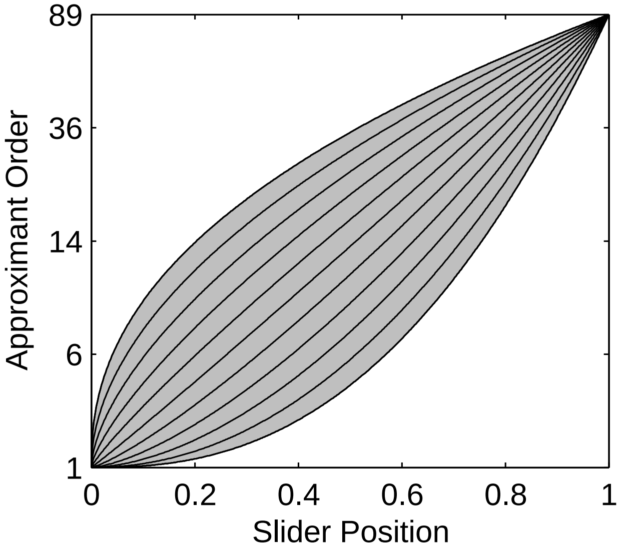


Figure 4.12: The range of possible mappings between slider position and order N. The black lines show ten example mappings, all with equal probability.

Statistical Treatment

For each scene, and each observer, the geometric mean and standard deviation of the five threshold estimates, N_{se}, are computed. The average threshold for each scene is given by the geometric mean of the averages for each observer.

Participants

The five observers from the previous experiment participated in Experiment 3. No training was performed prior to starting Experiment 3. The entire experiment required about 1.5 hours of listening, and was completed in a single session.

4.3.2 Results

Figure 4.13 shows the average thresholds of subjective equality for the five observers and three auditory scenes. The top panels, and bottom left panel, show the average thresholds for each observer. Error-bars indicate the standard deviation. The bottom right panel shows the average thresholds for each VAS, and the error

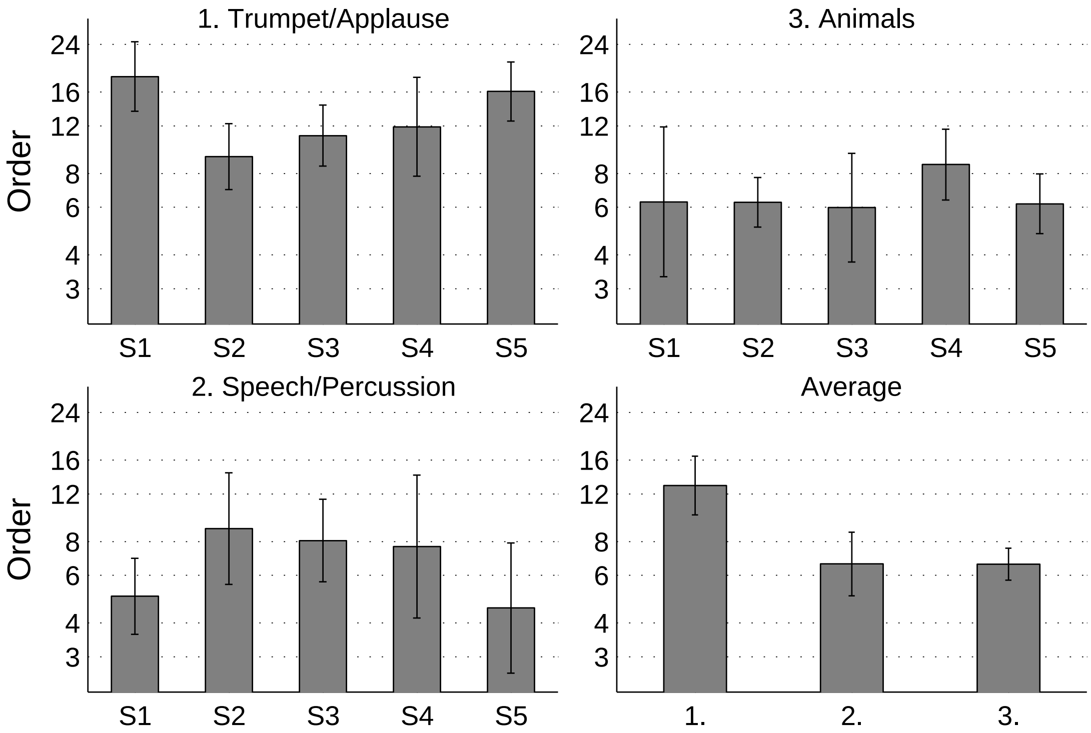


Figure 4.13: Discrimination thresholds for three virtual auditory scenes: 1. Trumpet followed by applause and fireworks, 2. Revolving speech and percussion with birds overhead, 3. Animal sounds. The bottom right panel shows the average thresholds across the five observers.

bars show the standard deviation of the observer averages.

4.3.3 Discussion

For the first VAS the average threshold is $N_{se} = 13$, and for the other two scenes the average threshold is $N_{se} = 7$. That the thresholds are significantly lower for the second two scenes is likely due to the monaural source signals used to create the scenes. The first VAS contains several recordings of diffuse applause in reverberant environments. The applause signals are the most stationary and wideband of the source signals included in the three scenes. As such, the first VAS provides more consistent spectral cues that aid in discrimination and is most similar to the noise bursts used in Experiment 2. Indeed, all five participants reported that they focused their attention upon the applause portion of the first VAS in performing the task. Furthermore, the participants generally reported subjectively that the discrimination of applause was primarily timbral rather than spatial.

The second two scenes yield lower thresholds. The standard deviation is larger for the second VAS than the third VAS, however. The percussion source that is included in the second VAS may have provided cues that aided in discrimination for some blocks, but not consistently. The percussion signal provides sharp impulses that are relatively wideband compared to the other source signals. However, the percussion articulates a clear rhythmic pattern that is invariant to the approximation quality. That is, the specific timbre of individual percussive hits may reveal subtle timbral cues that improve discrimination, but only if the observer chooses to the concentrate on such individual minutiae rather than the rhythmic performance. Thus the listening strategy of the observer may have affected thresholds more for the second VAS than the third VAS.

4.4 Stimulus Uncertainty

We now return to Experiment 2 and explore the dependence of $N_{70.7}$ on stimulus uncertainty in greater detail. Figure 4.14 shows the average threshold estimates for all eight conditions. For each condition, the average threshold is given by the geometric mean of the thresholds for each observer. State-space approximants yielded lower $N_{70.7}$ estimates than FIR approximants for all stimulus conditions. However, the type of stimulus uncertainty affected thresholds for the two types of approximants differently. The timbral uncertainty variable had a clear influence on the $N_{70.7}$ estimates for both approximants. For state-space approximants, thresholds fell about 35% from identical to independent conditions, whereas thresholds fell about

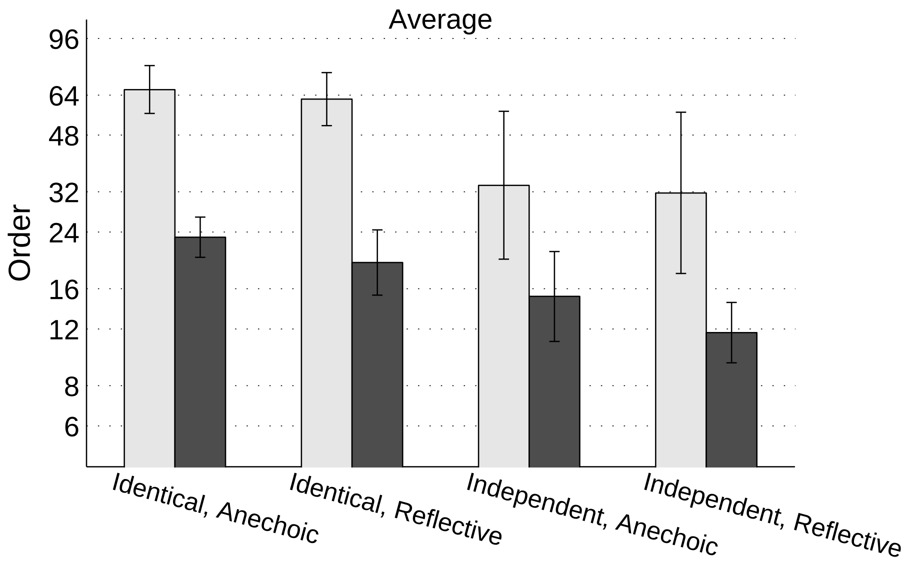


Figure 4.14: Average 70.7% thresholds. Error bars show the standard deviation of the threshold estimates across observer.

50% for FIR approximants. In contrast, spatial uncertainty only affected the $N_{70.7}$ estimates for state-space approximants. Where thresholds fell about 18% from anechoic to reflective conditions for the state-space approximants, thresholds did not change significantly for the FIR approximants.

4.4.1 Auditory $\mathcal{L}^2$ Error Distribution

That the threshold estimates fall for the reflective condition with state-space approximants, but not with FIR approximants was an unexpected result. This result may be explained, in part, by considering the distribution of approximation error across the collection of $D = 50$ directions modeled by each approximant.

Recall the auditory $\mathcal{L}^2$ error reported in Section 3.2.2. The distribution of this error across the $D = 50$ directions is different for FIR and state-space approximants. Figure 4.15 shows average empirical distributions for two FIR approximants and two state-space approximants. The orders of the approximants are given by the average $N_{70.7}$ estimates in the anechoic condition, as shown in Figure 4.14. The left panels of Figure 4.15 show the distributions for the identical noise condition, and the right panels show the distributions for the independent noise condition. Specifically, the orders of the FIR approximants are $N = 67$ and $N = 34$, and of the state-space approximants are $N=23$ and $N=15$.

Consider a simple binary model for a listener's ability to discriminate an HRTF approximant from a full-order HRTF, in which for a given HRTF approximant at a fixed direction, a listener is either always able to perform the discrimination task, or never able to perform the task. Suppose the auditory $\mathcal{L}^2$ error of the approximant is a sufficient statistic for whether or not a listener can perform the discrimination task. That is, there exists a threshold, $\epsilon_{\mathcal{L}^2}$, such that a listener can discriminate the approximant from the full-order HRTF if and only if the auditory $\mathcal{L}^2$ error of the approximant is greater than $\epsilon_{\mathcal{L}^2}$. The 70.7% discrimination threshold is the point at which the listener guesses for $2 \times (1 - 0.707) = 58.6\%$ of the trials, and is able to perform the task for the remaining 41.4% of the trials. Therefore, $N_{70.7}$ is the approximant order for which 58.6% of the auditory $\mathcal{L}^2$ error distribution is less than $\epsilon_{\mathcal{L}^2}$, and 41.4% of the auditory $\mathcal{L}^2$ error distribution is greater than this threshold. $\epsilon_{\mathcal{L}^2}$ is shown by the vertical line in each panel of Figure 4.15. For both identical and independent noise instances, $\epsilon_{\mathcal{L}^2}$ is in general agreement for FIR and state-space approximants. For the identical noise condition, $\epsilon_{\mathcal{L}^2} = 2.91$dB for FIR approximants and $\epsilon_{\mathcal{L}^2} = 2.35$dB for state-space approximants. For the independent noise condition, $\epsilon_{\mathcal{L}^2} = 3.90$dB for FIR approximants and $\epsilon_{\mathcal{L}^2} = 4.00$dB for state-space approximants.

For the reflective condition, the stimulus consists of six binaural signals summed together: the direct wave plus five virtual reflections, where the reflections are identical to the direct wave except for having been attenuated, delayed, and processed with a different HRTF pair. Suppose the listener aggregates the six distinct auditory

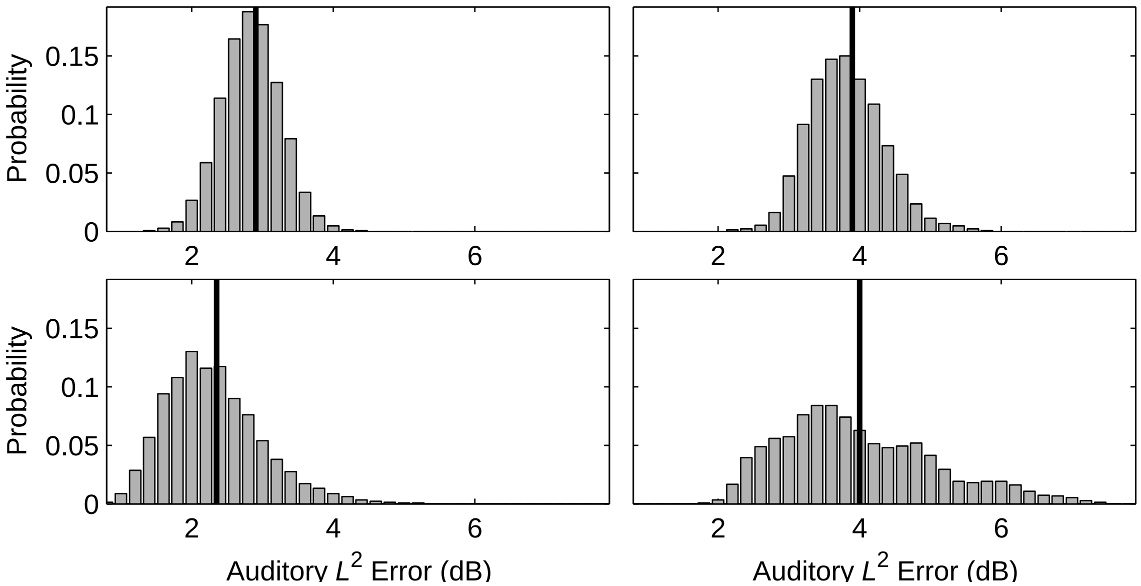


Figure 4.15: Average empirical auditory $\mathcal{L}^2$ error distributions for two FIR approximants (top panels) and state-space approximants (bottom panels). The orders of the FIR approximants are $N = 67$ (left) and $N = 34$ (right), and of the state-space approximants are $N = 23$ and $N = 15$. The vertical line in each panel indicates the error for which the cumulative probability is 58.6%

$\mathcal{L}^2$ errors into a new sufficient statistic, and that if and only if the aggregate statistic falls above $\epsilon_{\mathcal{L}^2}$, as shown in Figure 4.15, can the listener perform the discrimination task.

The aggregation rule used by the listener is unknown, but there are several possibilities. For example, if we assume due to precedence that the listener ignores the five reflected waves entirely, then 41.4% of the trials would yield an aggregate statistic above the error threshold and discrimination performance would be unchanged for both approximants. On the other hand, if we assume that the aggregate statistic is given by the maximum of the six errors, then $1 - 0.586^6 = 96\%$ of the trials would yield a statistic above the threshold[15]. Indeed we can easily verify numerically that,

[15] Assuming that the six errors are independent random variables.

for any of the empirical distributions shown in Figure 4.15, sampling six instances from the distribution yields a maximum auditory $\mathcal{L}^2$ error that is greater than $\epsilon_{\mathcal{L}^2}$ with probability 96%.

Suppose the aggregate statistic is given by a weighted combination of the six individual auditory $\mathcal{L}^2$ errors, where the weights are given by the inverse of the relative distance from each image source to the listener. If the aggregate statistic is the weighted linear mean of the six errors, then the fraction of trials in which aggregate statistic falls above $\epsilon_{\mathcal{L}^2}$ increases for all four approximants in Figure 4.15 to $45-49\%$. However, if the aggregate statistic is the weighted geometric mean of the six errors, then the fraction of trials for which the aggregate statistic exceeds $\epsilon_{\mathcal{L}^2}$ differs for the FIR and state-space approximants. For the FIR approximants, 44% of the trials exceed $\epsilon_{\mathcal{L}^2}$, whereas as only 34% of the trials exceed $\epsilon_{\mathcal{L}^2}$ for state-space approximants. Hence this model of discriminability predicts that, when virtual acoustic reflections are added to the stimulus, $N_{70.7}$ drops with state-space approximants, but not with FIR approximants.

Unfortunately, this model of reflective stimulus discriminability does not accurately predict the size of the drop in $N_{70.7}$ with state-space approximants. That is, if we consider the error distribution with approximant order given by $N_{70.7}$ for the reflective conditions ($N = 19$ and $N = 12$ for identical and independent conditions, respectively), the aggregate statistics exceed $\epsilon_{\mathcal{L}^2}$, as computed from the distributions in the anechoic condition, for $80 - 90\%$ of the trials, rather than the 41.4% that we might expect. Furthermore, the author is unaware of any psychophysical rational for this model. Nonetheless, this model predicts the direction of change in $N_{70.7}$ even if it does not predict the size of the change.

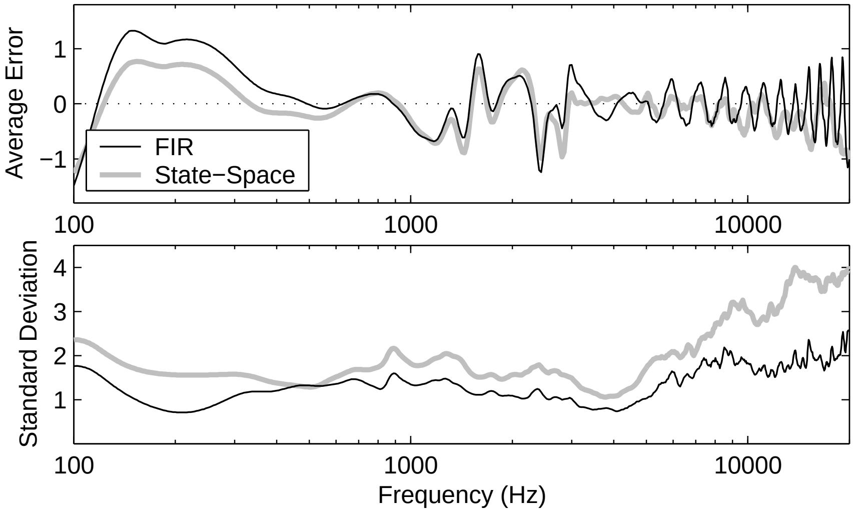


Figure 4.16: Average decibel error as a function of frequency for FIR and state-space approximants with order $N = 34$ and $N = 15$, respectively. The top panel gives the average error, and the bottom panel gives the average standard deviation of the error across the $D = 50$ directions.

4.4.2 Spectral Error Distribution

The different trends of $N_{70.7}$ for FIR and state-space approximants can be further explained by considering the distribution of error across the $D = 50$ directions as a function of frequency. Figure 4.16 shows the average error in decibels (top panel) and the average standard deviation of the error (bottom panel) for FIR approximants with order $N = 34$ and state-space approximants with order $N = 15$. The orders are given by the $N_{70.7}$ estimates for the independent/anechoic condition. The error is computed from the critical-band smoothed spectra of the measured HRTFs and approximant responses.

The state-space approximants yield lower absolute error for 150-400 Hz and 3-9 kHz, and yield similar error to the FIR approximants at all other frequencies. However, the average standard deviation of the error is greater with the state-space approximants at all frequencies. That is, the error of the FIR approximants is more consistent across the $D = 50$ directions than it is for the state-space approximants. Therefore, if the stimulus consists of several directions added together, then the error at each frequency with the FIR approximant is more likely to be reinforced than with the state-space approximant. This would seem to be the case for the 150-400 Hz frequency range in particular, where the FIR approximant yields an average error about 0.5 dB greater than the state-space approximant, and a standard deviation about 1 dB lower than the state-space approximant.

Overall, the result above would seem to favor state-space approximants for the display of reflective stimuli. However, the observation above also indicates that it may be easier to apply a global corrective filter to the FIR approximants than to the state-space approximants.

4.4.3 ‘Controlled’ versus ‘Everyday’ Uncertainty

Experiments 2 and 3 estimated different perceptual thresholds. Experiment 2 yielded thresholds in the range $12 \leq N_{70.7} \leq 24$, whereas Experiment 3 yielded thresholds in the range $7 \leq N_{se} \leq 12$. This is due, in part, to the psychophysical procedures. The second experiment used an adaptive procedure to estimate the $N_{70.7}$ threshold, whereas the third experiment used a method-of-adjustment to estimate the threshold of subjective equality, N_{se}. The latter is known to be subject to observer bias, since it asks the observer to judge when “good” is good enough. Nonetheless, the differences between the $N_{70.7}$ and N_{se} estimates are likely dominated by the differences in the stimulus between Experiments 2 and 3.

Experiment 3 considered everyday stimuli consisting of multiple complex sound sources that were ‘nonstationary,’ both in the sense that the sources evolved over time, and in the sense that the sources moved relative to the listener. With complex time-varying sound sources, listeners may tend to concentrate on 'higher-level' attributes of the sound objects, such as how the sound objects evolve over time, rather than fine timbral detail. For example, when listening to a speech signal, the listener may be distracted by what is said. This tendency is further evidenced by the threshold estimates for individual observers.

In Experiment 2, the first observer, S1, yielded consistently higher thresholds than the other observers. This disparity is not evident in Experiment 3. For Experiment 2, S1 performed significantly more fixed-level training than the other observers. For Experiment 3, none of the observers had any training. S1 may have received a modest amount of implicit training as he was developing Experiment 3, but he conducted no training or testing prior to performing the experiment. From Figure 4.13 it can be seen that S1 yields higher thresholds for the first VAS only. Due to the applause, this VAS is perhaps most similar to the noise stimuli used in Experiment 2. This implies that the higher discrimination thresholds evidenced by S1 in Experiment 2 are a result of the additional training that S1 performed with wideband noise stimuli. That is, observer S1 may be less prone to focusing on the higher-level attributes of the stimuli, and is relatively more focused on timbral detail.

4.5 Measures of Perceptual Fidelity

Different experimental methodologies have been employed to characterize the perceptual fidelity of HRTF approximations. The methodology employed in the present work was chosen because of the modest physical resources required, and the guarantee that threshold estimates are perceptually sufficient for the binaural display of a variety of virtual auditory scenes. For the worst-case scenario, in which the timbral minutiae of broadband noise bursts must be accurately rendered, we found that the minimum order is approximately $N = 24$. For more pragmatic auditory scenes with 'real-world' sources, in which a small amount of timbral distortion is allowed, we found that the minimum order is approximately $N=7$.

Other studies of HRTF approximations employ different methodologies and conclude that vastly different minimum orders are required for adequate fidelity, ranging from $N=6$ to $N=100$ (9; 60; 64). To a large extent, this broad range of minimum orders reflects the specific psychophysical criteria used for assessing perceptual fidelity.

Many studies of HRTF approximations consider localization performance, rather than discrimination performance. A common paradigm is to compare the directional errors when HRTF approximations are used with the errors when full-order HRTFs are used (9; 48). Timbral differences between the approximate and full-order HRTFs are ignored so long as they do not affect the location estimates. This paradigm is useful, for example, when searching for directional hearing cues. However, this approach does not yield thresholds that guarantee no loss in fidelity. It is unclear to what extent timbre influences the 'immersiveness' of the display. For example, a virtual source may be heard at the correct direction, but may not be well externalized. Furthermore, for some binaural display applications, such as concert hall modeling and prediction, timbral accuracy is critical.

In contrast to localization performance, discrimination performance reveals whether or not the listener hears any difference between the approximate and full-order HRTFs. Such experiments may be divided into two broad groups: those with a virtual reference condition (headphone display using full-order HRTFs) (64) versus those with a physical reference condition (an appropriately positioned loudspeaker in physical space) (69; 138; 151). The latter group requires a physical space, an array of loudspeakers, and sensitive measurement and equalization equipment. Furthermore, a ‘physical space’ reference necessarily limits the variety of auditory scenes that may be used as stimuli. In the interest of convenience and flexibility, we use a virtual reference condition in the experiments below.

Some HRTF discrimination experiments attempt to force listeners to use only spatial cues, and not timbral cues, by adopting a technique from auditory profile analysis (152), in which a random amplitude rove is applied to each instance of the stimulus[16]. Sometimes the rove is applied to the overall signal level (64), and sometimes it is applied independently to each critical frequency band (48; 151). Furthermore, some studies instruct the listener to only use spatial cues in performing the discrimination task, and to ignore timbral cues (64). However, given that there is no clear distinction between spatial cues and timbral cues[17], any attempt to force listeners to use only spatial cues may allow a significant loss in overall fidelity. In particular, this may be case when non-individualized HRTFs are used, as a spatial distortion with a listener’s own HRTFs may be perceived as a timbral distortion when presented to a different listener. In the present study, no random roves are employed and listeners are instructed to use any cue that aids in discrimination.

Another methodology that has been employed to evaluate HRTF approximations is subjective similarity (10; 60). In this case, listeners are presented with a sequence of binaural stimuli rendered using both the full-order HRTFs and approximate HRTFs, and asked to quantify the similarity between the two renderings. Usually a scale from

[16] In studies of auditory profile analysis, a random rove is applied so as to force the listener to integrate spectral shape across frequency channels, rather than simply detecting the level change in any one channel. Of course, a change in level at any frequency is tantamount to a change in spectral shape, but the listener can detect this change by either the absolute level at that frequency or the change in spectral shape. In auditory spectral analysis it is important that the procedure only measure detection based on the later cue

[17] For example, consider the median plane, where a change in source elevation is tantamount to a change in timbre.

one to five is used, with a five meaning ‘identical’, four meaning ‘very similar’ and so forth. Clearly, this method is less formal, and there is no guarantee that any threshold estimate is the true perceptual threshold and not the threshold where the listener simply decides “close enough.” Furthermore, this method often yields results that are noisey and difficult to interpret[18]. In Experiment 3 above, the listener was asked to search for the poorest “Approximation Quality” that yields identical approximate renderings. While this method only estimates subjective equality thresholds, the listener does not quantify the subjective similarity of any rendering that he or she feels is less than ‘identical’ to the “Ideal” rendering.

4.6 Summary

Three psychophysical experiments have been conducted to validate the state-space methods proposed in previous chapters. Experiment 1 estimated psychometric functions for several stimulus conditions for one observer. This experiment demonstrated that an adaptive-level method is feasible with the discrimination task, and answered several preliminary questions necessary for the design of an adaptive-level experiment. For a wide range of system costs, C, and all stimulus conditions, the FIR approximants were found to be significantly more discriminable than equal cost state-space approximants. Further, it was found that the use of individualized versus non-individualized HRTFs has little influence on discrimination performance. SIMO state-space approximants were found to outperform equal-cost MISO state-space approximants by a small margin. SIMO state-space approximants were considered in Experiment 2, and MISO state-space approximants were considered in Experiment 3.

[18] One such difficulty is ‘calibrating’ across listeners

Experiment 2 estimated $N_{70.7}$ order thresholds for several stimulus conditions and five observers. The state-space approximants were found to yield significantly lower thresholds that the FIR approximants for all stimulus conditions. The difference in thresholds was largest for the "identical, reflective" condition; that is, the condition with the greater timbral uncertainty and the lesser spatial uncertainty. Conversely, the difference in the thresholds is smallest for the "independent, anechoic" condition; that is, the condition with the the lesser timbral uncertainty and the greater spatial uncertainty. This trend is explained, in part, by recognizing that the spectral magnitude error is more consistent across direction with FIR approximants than with state-space approximants. Thus, subtle timbral cues are more helpful in the discrimination task for FIR approximants, where the "independent" condition reduced the reliability of these cues. Furthermore, as virtual reflections are added to the stimulus, the net spectral magnitude error is more likely to compound with the FIR approximants. Overall, Experiment 2 showed that $12 \leq N_{70.7} \leq 24$ for state-space approximants depending upon stimulus uncertainty.

Experiment 3 introduced even more uncertainty into the stimulus. For Experiment 3, the stimulus was a complex, non-stationary virtual auditory scene. The thresholds, in terms of system order N, drop further. For two of the three scenes used in Experiment 3, the average discrimination threshold was $N_{se} = 7$, and for the other scene the threshold was somewhat higher, likely because fine timbral cues of wideband applause were used to aid in discrimination.

We have shown that indiscriminable state-space approximants present a significant cost savings over conventional HRTF implementations. For example, the total cost of two MISO systems of order $N = 7$ that model $D = 50$ HRTF pairs is $C = 778$ (34.3 MIPS). This is about one and a half orders of magnitude lower than the cost of convolving all 100 signals with full-order HRTFs, $C = 25,600$ (1129 MIPS). Clearly, state-space approximants provide a low-cost and high-fidelity method of constructing virtual auditory scenes for binaural display.

CHAPTER V

Summary Remarks and the Framework in Practice

The main contribution of this dissertation was the design and evaluation of state-space approximants of collections of HRTFs, and how such approximants naturally fit into a powerful framework for flexible and immersive binaural auditory display. The next section summarizes the key results of this dissertation, as well as limitations and avenues for future research. Section 5.2 outlines several informal synthesis recommendations that the author developed while working with the proposed framework. The final section discusses general implications of this framework and concludes the dissertation.

5.1 Summary

Binaural auditory displays seek to immerse a listener in a virtual auditory scene. Great progress towards this goal has been made throughout the last two decades. Nonetheless, contemporary displays suffer from burdensome computational demands, and are insufficiently flexible for the display of immersive scenes. This dissertation addresses these shortcomings with a state-space approach that both reduces the computational demands of the display, and improves the flexibility of the display.

5.1.1 Binaural Display Framework

Chapter I introduced spatial hearing and the head-related transfer function (HRTF), and described prior research on HRTF approximantion. HRTFs are the cornerstone of binaural auditory displays. The HRTF for a particular direction models the acoustic interactions of a plane wave enroute to a listener's two ears. However, individual plane waves occur only in primitive auditory scenes. Such scenes are rarely experienced in everyday listening. As such, binaural displays that render sound sources with a single HRTF pair often yield auditory scenes that lack *presence*, in which the sources are not well-externalized and do not move appropriately relative to the listener.

Typical everyday auditory scenes are complex. There are often multiple sound sources, sources that are moving relative to the listener, as well as spatially-extended sources. Commonplace environment cause acoustic reflections and diffuse reverberation, which further complicates the auditory scene. Many of this phenomena may be easily modeled with a multi-HRTF framework. By choosing the HRTFs to surround the listener with sufficient density then any plane-wave direction can reasonably be presented by a linear combination of nearby directions included in the HRTF collection. In this case a source signal is rendered by filtering the source signal with the full collection of D HRTF pairs, and the D binaural signals are combined so as to display the auditory scene to the listener. However, this framework compounds an already formidable computational load.

5.1.2 State-Space Formulation

Fortunately, collections of HRTFs exhibit much redundancy. Therefore a system that models a collection of HRTFs may be able to accurately implement the entire collection with lower cost than a system that implements each HRTF independently. One such system that can model multiple transfer functions simultaneously is a state-space system. State-space systems are attractive for several reasons. Being a pole-zero structure, state-space systems can efficiently model sharp spectral peaks and notches. State-space systems are relatively simple to implement in hardware and do not introduce latency. And finally, robust state-space order-reduction methods exist that guarantee stability and causality and minimize spectral error in some sense. Accordingly, we formulate the HRTF system in the state-space and describe the design of efficient and accurate approximants in Chapter II.

Hankel Order-Reduction

It is straightforward to design a state-space system that models a collection of HRTFs exactly. Such a system is high order and very costly, higher cost even than implementing each HRTF individually with convolution. However, the state-space form admits a convenient method for dramatically reducing the order of the system.

Two order-reduction techniques were explored, both of which are based on the Hankel operator of the state-space system. The Hankel operator is similar to the convolution operator except that it only considers the relationship between past inputs to future outputs. The Hankel error lower bounds the maximum convolutional error (the $\mathcal{L}^\infty$ spectral error), and with the HRTF data under study the Hankel error is found to be a fortuitously tight bound on the $\mathcal{L}^\infty$ error.

The two order-reduction methods considered in this dissertation are Hankel-norm optimal approximation (HOA) and balanced model truncation (BMT). On one hand, the former method is optimal in the Hankel-error sense, whereas the later is not optimal in any sense. On the other hand, the HOA approximants are relatively burdensome to compute, whereas BMT approximants are simple to compute.

State-Space Architectures

Three state-space architectures are considered. A MIMO architecture with D inputs, one for each direction, and 2 outputs, one for each ear, is perhaps most obvious. However, the performance of this architecture is found to suffer due to interaural time delay. Modeling time delay in state-space is inefficient, we found that attempts to reduce the order of such systems smeared the time-domain response of the approximants, even for modest reductions in order. Due to the perceptual sensitivity to ITD, the MIMO architecture was judged untenable.

If the arrangement of inputs and outputs is changed such that the system has either one input or one output, then the ITD can be factored out of the state-space system. In the case of the SIMO architecture, only one distinct sound source can be rendered, and the room/motion model is applied after the state-space filtering. Hence the SIMO architecture only solves a subset of the general binaural display problem. In contrast, the MISO architecture allows the display of simultaneous distinct sources. However, the MISO architecture requires two separate state-space systems, one for each ear.

5.1.3 Empirical Evaluation

Chapter III characterized the performance of the two order-reduction methods and three architectures, as well as an array of truncated FIR filters of equal net cost. In order demonstrate that the state-space approximants are *efficient* it was necessary to show that the approximants are simultaneously accurate and low cost. Accordingly, the state-space systems were shown to be both more accurate and lower cost than a reference approximant. We chose an array of truncated FIR filters as the reference system because HRTFs are nearly minimum-phase, and hence admit an obvious and reasonably accurate FIR approximant for any order N. In order to compare state-space and FIR approximants of equal net cost, the cost of each system was defined to be the total number of filter coefficients in the system.

Aggregate Performance

The average approximation error was computed for both state-space and FIR approximants with a fixed computational cost. The cost bound was set to the total cost of implementing eight HRTFs with convolution. The number of directions modeled by the approximants was treated as an independent variable in the range $1 \leq D \leq 110$. Numerous error metrics were considered, including the Hankel error, $\mathcal{L}^{\infty}$ error, and auditory $\mathcal{L}^2$ error. For all metrics, the state-space approximants outperformed the equal-cost FIR reference for $D > 10$, and the double-cost FIR reference for $D > 20$.

The MIMO architecture yielded favorable performance in terms of $\mathcal{L}^{\infty}$ error, but poor performance in terms of auditory $\mathcal{L}^2$ error. The poor auditory $\mathcal{L}^2$ performance was found to be due to the ITD, with error concentrated in the contralateral HRTFs. The SIMO and MISO architectures yielded similar performance overall, although the SIMO architecture yielded slightly lower error.

Spectral Detail

Upon closer inspection, the state-space approximants were found to more accurately model perceptually important features of the HRTFs, such as spectral notches. The state-space approximants were also found to yield more accurate approximants at low frequencies. Low frequency accuracy is often described as unimportant for spatial hearing, although listeners are sensitive to timbrel distortions at low frequencies.

The two order-reduction methods, HOA and BMT, were found to yield similar approximation quality overall. HOA was found to distribute error more uniformly across frequency, whereas BMT was found to concentrate error near spectral notches. However, this trend was slight and not likely to be perceptually significant in practice.

5.1.4 Psychophysical Evaluation

Having empirically characterized the efficacy of the state-space approach, Chapter IV reported on a headphone listening experiment that estimated the minimum order N required such that the listener could not tell the difference between the approximant and the full-order HRTF. The threshold was estimated with a series of discrimination experiments that considered varying types of stimulus uncertainty. A preliminary experiment was conducted with one observer, the author. This was a fixed-order experiment that directly estimated psychometric functions for several conditions. The results of this experiment were used to design an adaptive-order experiment in which five observers participated.

Both state-space and FIR approximants were included. Minimum order thresholds were found to be consistently lower with state-space approximants than FIR approximants, although the cost-savings with state-space approximants was found to depend on stimulus uncertainty. All approximants considered in these experiments modeled $D = 50$ HRTF pairs that surrounded the listener approximately uniformly.

Threshold Estimates

The second experiment estimated objective 70.7% discrimination thresholds. For FIR approximants the average thresholds were within $32 \leq N_{70.7} \leq 67$, and for state-space approximants the thresholds were within $12 \leq N_{70.7} \leq 23$. The highest thresholds were estimated for the condition of least timbrel and spatial uncertainty. Timbrel uncertainty caused the FIR thresholds to drop by 50%, whereas the state-space thresholds dropped by 35%. The larger drop with FIR approximants implies that subtle timbrel minutia were more apparent than they were for state-space approximants. In contrast, spatial uncertainty had no affect on the FIR thresholds, whereas the state-space thresholds dropped by 18%. FIR approximants were found to yield more consistent spectral error across the $D = 50$ directions. The reflective stimulus condition, which corresponds to increased spatial uncertainty, was modeled with the combination six binaural signals. The more diffuse stimulus impaired discrimination with state-space approximants, causing the threshold to drop. With the FIR approximants however, the spectral error across the six directions was more likely to be reinforced, and the thresholds were unchanged.

The final experiment estimated thresholds of subjective equality with complete virtual auditory scenes used as stimulus. Only state-space approximants were considered. The average thresholds were within $7 \leq N_{se} \leq 12$. The highest threshold, N_{se}, was estimated for an auditory scene that contained diffuse, reverberant applause. This stimulus provided reliable, wideband timbrel cues that aided the observer in discrimination. The remaining two auditory scenes yielded thresholds of $N_{se} = 7$.

5.1.5 Contributions

This dissertation made several contributions to the research community. The main contributions are listed below.

- A unified framework for the binaural display of a wide variety of virtual auditory scenes was described. This framework is based on a model of a large collection of HRTFs surrounding the listener. With this framework, a binaural display can render auditory scenes that include multiple simultaneous sources, spatially-extended sources, acoustic reflections, and source and listener motion.

- A state-space model of collections of HRTFs was designed in order the reduce the cost of the entire collection. Order-reduction techniques based on the Hankel-operator were applied to design low-order approximants.

- An empirical experiment was conducted in which low-order state-space systems are compared to truncated min.-phase FIR arrays of equal net cost. Two approximation errors were reported, the $\mathcal{L}^\infty$ error and the auditory $\mathcal{L}^2$ error. For a cost bound equal to implementing eight HRTFs directly, the state-space systems yield lower error than the FIR arrays for $D > 10$. Furthermore, the state-space systems yield lower error than double-cost FIR arrays for $D > 20$.

- The modeling of ITD in the state-space was explored, but no computationally efficient state-space solutions were found. This problem was solved by using state-space architectures, SIMO and MISO, that allowed the ITD to be implemented externally.

- Two state-space order reduction techniques, BMT and HOA, were compared for large-scale systems. These two methods have previously been compared for SISO filters, but few comparisons have been made for the MIMO case. The relative differences between the two methods that were reported for SISO filters were observed in MIMO approximants; BMT tends to concentrate error in the vicinity of spectral notches and transition regions, whereas HOA spreads the error more uniformly. However, in practice the two methods yield roughly equivalent performance in approximating collections of HRTFs.

- Three psychoacoustic experiments were conducted to estimate minimum order thresholds for indiscriminable state-space approximants. The first experiment was a fixed-level test that estimated the psychometric function for several conditions. The main experiment showed that order thresholds for state-space approximants are less than half those of FIR arrays for several stimulus conditions. With broadband noise stimuli, state-space thresholds were found to be $9 \leq N_{70.7} \leq 23$. The final experiment considered stimuli that consisted of complete dynamic auditory scenes with 'real-world' sound sources and acoustic reflections. In this case, state-space thresholds were found to be $7 \leq N_{se} \leq 12$.

5.1.6 Limitations and Future Work

The proposed multi-direction framework, couple with state-space approximants, enables the binaural display of a wide variety of auditory scenes. The framework has some disadvantages however. The framework is physically motivated rather than perceptually motivated. From a perceptual viewpoint, acoustic reflections may not need to be rendered as accurately as the direct wave. The computational cost of the display may be reduced if a more perceptually efficient model of acoustic reflections and reverberation is used. Nonetheless, the framework is convenient for scene design precisely because it is a simple physical model, and the use of state-space approximants keeps the computational cost of this framework within reason.

Diffuse reverberation was not considered in this dissertation. The acoustic response of an enclosure is usually divided into early reflections and diffuse reverberation. Modeling individual reflections for the diffuse portion of room response is both computationally untenable and prone to obvious distortions if modeled using standard linear acoustics. The diffuse portion of the room response should be implemented as a separate module, perhaps using the HRTF filterbank, but not necessarily.

The image-source model often yields unnatural sounding virtual enclosures, and is not easily generalized to enclosures with complex shape and diffusing surfaces. The image-source model can be replaced with a ray-tracing procedure without increasing the cost of the HRTF filterbank, although the ray-tracing procedure itself would be more expensive. The advantage of ray-tracing would be greater generality, as well as potentially more natural sounding virtual enclosures. In fact, any ray-based room auralization system can be coupled with the proposed binaural display framework. Not all methods of room auralization can be coupled with the proposed framework, however. BEM and FEM are popular and powerful alternatives for auralization and room prediction. These methods are based on a different structure, and the effect of the room and the effect of the listener cannot be separated as with ray-based methods. In this case the HRTF is incorporated into the model of the room and the BEM/FEM directly computes the final signal at the listener's ears.

From the perspective of hardware implementation, one obvious drawback to the proposed framework is the necessity of a large number of parallel audio channels for the HRTF filterbank. A large number of parallel channels inevitably increases the manufacturing cost. On the other hand, the state-spate approximants require no memory to work with all $2D$ signals, aside from a few memory cells to store the state vector. The overall system may need to generate a large number of signals for the HRTF filterbank, but these signals are immediately combined without requiring additional storage or processing. The hardware costs of the proposed framework and state-space implementation are by no means trivial, but they are within the reach of contemporary embedded DSP systems.

5.2 Immersive Auditory Scene Design

Before concluding the dissertation, we wish to summarize several subjective 'rules-of-thumb' that appear to improve the presence, or immersiveness, of virtual auditory scenes synthesized using the methods developed in the previous chapters. In practice, synthesizing an immersive virtual auditory scene requires more than accurate HRTF filters. As discussed in Chapter I, the HRTF is based on strict acoustical assumptions that are rarely experienced in everyday listening. The framework for binaural display that was considered in this dissertation models a dynamic complex auditory scene as a collection of static primitive point sources in free-space. Some scenes are easier to model and more immersive than others when using this framework. A binaural sound designer must be aware of what auditory scenes are simple to render and perceptually convincing.

The recommendations below are the result of the author's experience designing several complete binaural auditory scenes for the experiment reported on in the previous chapter. These auditory scenes included multiple simultaneous sound sources, source and listener motion, and acoustic reflections. Overall, the scenes are quite compelling, given the simple physical model that they are based upon.

The recommendations concern the multi-HRTF framework, and not the HRTF implementations. It is assumed that a collection of D HRTFs is available with sufficient fidelity. The recommendations are both restrictions on the auditory scene that generally yield the best sense of presence, as well as system parameters in the rendering of the scene. The recommendations are divided into two groups. The first group concerns static auditory scenes; the nature of the source signals, the number of directions D, and the image-source model for room reflections. The second group concerns the incorporation of source and listener motion, and frame segmentation issues that result.

5.2.1 Static Synthesis Recommendations

- The monaural source signals influence the presence of the scene, and the spatial cues of the source should compliment the content of the source signal. We found that everyday sounds, such as samples taken from field recordings, were externalized more convincingly than synthetic sounds, such as broadband noise bursts. The presence of the scene is further improved if the spatial cues of the source compliment the listener's expectation for where the source should be positioned and move. For example, a monaural signal of a fly can be quite effective if the virtual fly is located near the listener, but is unlikely to be heard far from the listener[1].

- The image-source model is only appropriate for certain environments. The original authors of the method recommend the method for 'small' enclosures (101). We found, in practice, that for large enclosures the early reflections predicted by this model are perceived as shrill echoes that break down the percept of a coherent acoustic environment. Furthermore, if higher-order image sources are included the sound object acquires an unnatural ringing timbre[2]. Finally, large reflection coefficients tend to compound the ringing/echoey quality of the model. Thus we found that: reflection coefficients should be small, $\beta < 0.5$, including reflections greater than second-order may introduce timbrel distortion,

[1] The percept of a fly buzzing nearby a listener can be further improved if the sound is brief and fades in and out. A fly is not an omni-directional sources, its radiation pattern consists of a 'flower' of lobes that the listener would pass into and out of as the fly passes by.

[2] Ignoring the distinct HRTFs filters for each image-source, including all image-sources is tantamount to an array of comb-filters, hence the ringing. Filtering each reflection with a distinct HRTF tends to reduce the ringing, but this was still an obvious distortion.

and no room dimension should be greater than about 10m.

- Low-frequency components ($\lesssim$ 200 Hz) are common in everyday sounds, and are important to the sense of presence, but can be problematic to include in the binaural auditory scene. Low-frequency acoustic energy tends to be physically diffuse and contributes little to the percept of source direction. Measured HRTFs are often attenuated at low frequencies due to limitations in the measurement process. Nonetheless, removing low-frequency components from a source signal yields a sound object that appears unnaturally thin and brittle. Furthermore, low-frequency room modes are perceptually obvious. But modeling room modes or rendering sources with much low-frequency energy sometimes yields virtual scenes that appear unnaturally bloated. Overall, it may be preferable to render low-frequency content with a separate diffuse model, analogous to the N.1 surround-sound formats.

- Including a large number of directions D in the display is a design convenience, but not a perceptual necessity. In the experiments reported on the previous chapter, all displays modeled $D = 50$ directions. For a fixed auditory scene, we informally found that often $20 < D < 30$ HRTFs spread uniformly around the listener was adequate. However, larger D tended to reduce the frequency and severity of the frame and motion artifacts described below.

5.2.2 Motion and Frame Segmentation

- Motion improves the externalization of a sound object. Virtual sources that are stationary relative to the listener are usually perceived as being either on or near the head. The sound object is focused and localized at the appropriate direction, but is not localized in the far-field. The percept of an externalized source becomes more convincing if the virtual source moves relative to the listener. This is true of both source motion and listener motion. Pure source motion improves the externalization somewhat, although the improvement is more dramatic with listener-controlled motion. For example, the moving source demo described in Appendix A allows the user to move a virtual source within a restricted area in real-time. Furthermore, a head-tracking device improves the presence of the display significantly.

- Complementing motion in the binaural dimension improves the percept of source motion in the other dimensions. Previous authors have observed that purely front-back or up-down motion is often perceived as a change in timbre rather than source motion (17). We have found that including acoustic reflections helps the listener resolve the "motion versus timbre change" ambiguity. Nonetheless, including a binaural (left-right) component to the source motion improves the percept of source motion significantly, even if this component is small compared to the other components of the source motion.

- Motion must be slow relative to the frame duration. The rendering engine used to generate the dynamic virtual auditory scenes in the listening experiments was frame-based. Every 50ms the engine updated the spatial positions of the sources and listener. As such, the motion of the sources had to be slow[3] relative to the 50ms interval. Including fast source or listener movements requires a short frame duration. In the limiting case of a frame that is a single sample period in duration, the only limitation on source motion is that a source move slowly relative to the sample period and speed of sound. In this case even the

[3] The angular velocity relative to the listener must be slow. For example, a source that is moving slowly along a straight-line trajectory may briefly have a large angular velocity it the trajectory passes near to the listener.

well-known doppler effect is accurately rendered.

- Frame boundaries can introduce glitches and other artifacts into the signal. This is a common problem in any audio compression problem, and great care is taken in the design of window shapes that minimize this distortion. In the present application, we found that the coupling of source motion and acoustic reflections will sometimes exacerbate frame boundary distortions. The changes in the relative delays of each reflection at frame boundaries (due to source motion) can cause distracting glitches to become audible. This artifact often appears if a sharp transient occurs during a frame boundary and the source is moving rapidly. In some instances this distortion can even cause the sound object to momentarily jump between distinct spatial locations. We found the slight adjustments to the source's position and timing were usually sufficient to remove such artifacts. Nonetheless, the potential for artifacts at frame boundaries would seem to imply that the frame duration should not be shorter than necessary.

Note that some of the recommendations above are essentially restrictions on the design of the auditory scene itself. That is, some scenes can be more effectively rendered than others. For many applications, such as video games and musical composition, these restrictions and recommendations simply become part of the craft. For example, a sound engineer that is synthesizing the binaural audio track for a film would be primarily concerned with making the scene immersive and entertaining. Hence, the exact positions of sound sources and reflecting surfaces can be shifted from where these objects appear in the video if it yields a more compelling audio track.

For other applications however, these restrictions are untenable. In architectural acoustics and concert hall prediction, the ability to accurately synthesize any arbitrary acoustic scene is paramount. Modifying any aspect of the scene would destroy the legitimacy of the prediction. Nonetheless, the binaural display framework proposed in this dissertation may still be suitable for room prediction if a sufficiently accurate ray tracing procedure is used. The image-source model has well-known limitations and was included in the present work primarily because of its simplicity.

5.3 Implications

Binaural auditory displays have many potential applications, from virtual-reality entertainment to information display, from electroacoustic composition to sonar interfaces. Binaural displays also have the potential to be a convenient tool for analyzing the human auditory system. The methods developed in this dissertation can be applied to any of these applications.

The multi-direction framework is powerful, flexible and convenient. For example, we can imagine a futuristic iPod that can not only play audio files, but can also binaurally auralize a physical model of an acoustic scene. That is, the final mixdown to a binaural signal occurs in the iPod itself depending upon the computational resources of the device, and the audio file itself need only consist of the source signals and physical model. Such an iPod could be customized to the individual listener, and even incorporate a head-tracking device to improve realism and presence.

Surround sound techniques are becoming common in sound recording and reproduction, and are now employed by composers as well. Music produced in this manner requires a multi-speaker reproduction system. Such systems are expensive, difficult to setup, and still relatively uncommon compared with two-channel home stereos.

Furthermore, such systems require a precise arrangement of the speakers within the room in order for spatial cues to be properly reproduced. A flexible headphone-based system could display the same scene without such elaborate hardware and installation. Indeed, Beyer-Dynamic has recently released a commercial headphone system that binaurally simulates a 5.1 home theater, including a general room model and head-tracking system (153).

Broadly speaking, headphones may become the dominant method of listening to music in the future, either with iPod's or personal computers. There is widespread speculation about this on internet discussion boards and blogs, from casual listeners to audiophiles. In an extreme example of this, Headphone Disco has recently been introduced in Europe, where wireless headphones are used in dance clubs instead of loudspeakers[4]. As binaural display technology improves in quality and drops in price, it may fuel the dominance of headphone listening. It appears that binaural technology may ultimately be a double-edged sword that, similar to other entertainment and virtual-reality technologies, increases the enjoyment of media and the accessibility of information, but also further isolates individuals from each other.

5.4 Conclusion

Headphone listening has become a prevalent part of everyday life, for both work and play. Yet the sound stage perceived with conventional headphones is still contained within the listener's head. Binaural display technology promises to expand this constrained sound stage to an unconstrained virtual stage. Many of the necessary tools for such a headphone system are in place, but have they not been collected into a single system and are too computationally burdensome to be widely distributed.

[4] http://www.headphonedisco.com/

This dissertation considered a powerful and flexible, but potentially expensive, framework for binaural display. We found that with state-space approximants the cost of this framework could be reduced to approximately the same cost as direct convolution with a single full-order HRTF pair. We validated the state-space approximants both empirically and psychoacoustically. For a wide variety of virtual auditory scenes, only a small number of states are required to render the scenes indiscriminably from full-order renderings. In conclusion, low-order state-space systems may be the missing link that enables binaural displays to become flexible, perceptually convincing, and affordable.

APPENDICES

APPENDIX A

The Binaural Display of Clouds of Point Sources

The previous chapters of this dissertation considered a general framework that renders a monaural signal at many locations surrounding the listener. This appendix reports on a preliminary study in which special case of the multi-direction framework: a system that filters a monaural signal with a collection of D *nearby* HRTFs, yielding a 'cloud' of point sources (7).

There is interest within the binaural community in the design of binaural displays that can render real acoustic environments, where sources may have spatial extent, be in the near-field, be present in highly reflective environments, or be moving (1; 88; 3). Using a single pair of head-related transfer functions (HRTF) is insufficient to render a convincing and flexible sound object in this context. The computational burden, along with filter interpolation and dynamic updating, presents a serious obstacle. To mitigate these problems, we propose a binaural display in which a monaural source is auralized at many locations simultaneously. Moving sources could then be rendered through amplitude panning (5; 86), circumventing filter interpolation and updating altogether. Acoustic reflections could easily be incorporated as well. Furthermore, rendering clouds of point sources could ease problems associated with individuation by providing relative cues and reducing the importance of fine spectral detail (139; 32; 28; 2).

Binaural displays typically render a monaural source, $S(\omega)$, by[1]

$$B_{L,R}(\omega) \;=\; H_{L,R}(\omega \,|\, \theta, \phi) \,\cdot\, S(\omega) \tag{A.1}$$

where $H_{L,R}(\omega \,|\, \theta, \phi)$ is the HRTF for left and right ears for azimuth θ and elevation ϕ. This chapter explores the most primitive of spatially-extended sources, a cloud of identical uniformly-spaced point sources that subtend some angle in space. A monaural signal is rendered at D nearby points in space by

$$\begin{aligned} B_{1\;L,R}(\omega) &= H_{1\;L,R}(\omega) \,\cdot\, S(\omega) \\ B_{2\;L,R}(\omega) &= H_{2\;L,R}(\omega) \,\cdot\, S(\omega) \\ &\;\vdots \\ B_{D\;L,R}(\omega) &= H_{D\;L,R}(\omega) \,\cdot\, S(\omega) \end{aligned} \tag{A.2}$$

where $H_{k\;L,R}(\omega) = H_{L,R}(\omega \,|\, \theta_k, \phi_k)$. The D binaural signals $\{B_{1\;L,R}(\omega), \cdots B_{D\;L,R}(\omega)\}$ can then be combined or further filtered[2] to achieve a wide range of spatially-extended sound objects.

Clearly, direct implementation of such a binaural display yields a computational cost that scales linearly with cloud size. Given the high-order of measured head-related impulse responses (HRIRs), such a binaural display exacerbates an already formidable computational burden. Accordingly, we explore perceptual coding and filter design techniques so as to reduce the computational cost to that of a binaural display for a single point-source. We propose a cascade of two filters for each point

[1] For simplicity, we neglect the influence of the headphone-to-ear-canal transfer function (HpTF).

[2] A fundamental result of linear systems theory gives us that the order of linear filters is irrelevant. Hence, not all points in a cloud need auralize the same signal. For example, a monaural signal could be rendered for a cloud of nearby points, and then each cloud signal could be filtered to yield a composite sound object in which the low-frequency components come from the bottom of the cloud and the high-frequency components come from the top of the cloud, simulating the sound object perceived from many home stereo speakers.

in the cloud, in which the first filter is common to all spatial locations in the cloud and the second filter provides the spectral cues for each point within the cloud. The goal is to ensure that the net filter response for each point in the cloud is similar to the measured HRTF, while shifting as much of the computational burden to the first filter. We implement two filter-order reduction techniques, truncation of the measured HRIRs, and balanced model truncation (BMT), for the design of low-order IIR filters from the measured HRIRs (65; 66; 67). We verify the binaural display using an objective perceptual metric (60), as well as two informal listening tests.

The next section introduces the two-stage filter implementation for the auralization of binaural clouds, as well as describes the filter designs in greater detail. We evaluate the method numerically using several sets of measured HRTFs in § A.2 and discuss observations from informal listening tests in § A.3.

A.1 Methods

We employ two approximation techniques to reduce the computational burden of direct implementation of (A.2), a perceptual approximation for sound spatialization coupled with filter order reduction. For simplicity, we focus on modeling the magnitude response of the HRTFs, and implement our binaural display using a minimum-phase reconstruction of the phase response plus a linear phase term for the interaural time difference (ITD).

A.1.1 Perceptual Approximation

Rather than filter the monaural source with the measured HRIRs for every point in space, we propose a two-stage filter bank. The two-stage model is motivated by the observation that small perturbations in source location can be affected by relatively

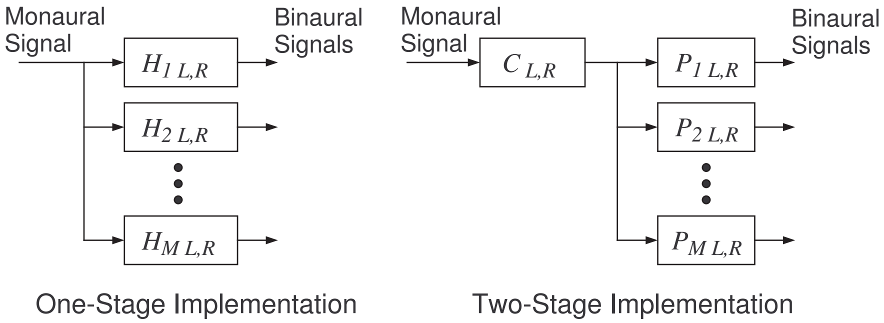


Figure A.1: The direct model for cloud rendering via HRTF filtering (left). The proposed two-stage model (right).

low-order changes in the spectrum of the signal (139). For example, a spectral notch might move a few hundred Hz, but otherwise the spectrum is unchanged. Furthermore, from a perceptual standpoint, it not necessary to render each spatial location in full detail, rather the cloud must be centered at the correct location, and then given a degree of spatial extent. We found during informal listening tests, that the precise spatial extent of a source is rarely perceived. The center, size, and oblongness are perceived, but not the precise shape (for example, a diamond versus a square).

Fig. A.1 shows two possible binaural display architectures. The one-stage architecture uses the measured HRTFs, $H_{1\cdots D\ L,R}(\omega)$, or some approximation thereof, to render the monaural source for D nearby spatial locations. The two-stage architecture first filters the monaural source with $C_{L,R}(\omega)$, providing cues that localize the center of the sound object, and then filters the output with $P_{1\cdots D\ L,R}(\omega)$, providing cues that define the extent of the sound object.

We propose two implementations of $C_{L,R}(\omega)$. The first is simply $C_{L,R}(\omega) = H_{k\ L,R}(\omega)$ where location k is the center of the source. The second is the average of the log magnitude spectra,

$$\log C_{L,R}(\omega) = \frac{1}{D} \sum_{k=1}^{D} \log H_{k\ L,R}(\omega). \tag{A.3}$$

For either case, the second filter is then given by

$$P_{k\ L,R}(\omega) = \frac{H_{k\ L,R}(\omega)}{C_{L,R}(\omega)}. \tag{A.4}$$

The hope is that $C_{L,R}(\omega)$ is sufficiently similar to $H_{k\ L,R}(\omega)$ so that the amount of spectral detail that $P_{k\ L,R}(\omega)$ must model is reduced(67). This architecture provides a framework for offloading the computational burden of rendering each cloud point individually to the common filter $C_{L,R}(\omega)$.

A.1.2 Filter Order Reduction

Measured HRTFs typically yield FIR filters on the order of hundreds of samples. Many filter design techniques have been explored in the context of the binaural displays (60). In the present chapter we employ one FIR and one IIR filter design. We construct an N^{th}-order FIR filter by simply truncating the minimum-phase impulse-response derived from the measured HRTF, yielding the optimal L_2 N^{th}-order FIR filter, in terms of spectral mismatch.

We construct an N^{th}-order IIR filter via balanced model truncation (BMT) (65), a convenient IIR filter design technique that has yielded promising results in the context of binaural displays (66; 67). BMT converts the ideal filter into state space form, where a transform is applied yielding an input-output equivalent filter with balanced controllability and observability grammians. The state space representation is then truncated to the N states with the largest Hankel singular values, yielding a solution that is *close* to the Hankel-norm optimum, which lies between the L_2 and L_∞ norms (125; 60). However, the transform matrix is often ill-conditioned, yielding

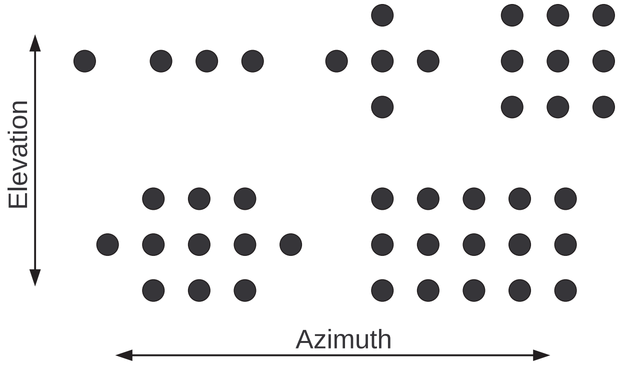


Figure A.2: Spatial cloud distributions for $D = (1, 3, 5, 9, 11, 15)$.

unstable filters. A simple algorithm for circumventing this problem in the case of approximating FIR filters is given in (65).

A.2 Characterization

A.2.1 Filter Implementations

A total of nine binaural implementations are considered below. A direct system, using the measured HRIRs to render each cloud point, provides an 'ideal' baseline for comparison. The remaining eight systems either use low-order filters ($N = 12$ in the case of FIR filters, and $N = 6$ in the case of IIR filters) exclusively, or a combination of a high-order $N = 256$ FIR filters for $C_{L,R}(\omega)$ and low-order ($N \in \{12, 6\}$) filters for $P_{1 \cdots D\ L,R}(\omega)$. Four architectures are implemented; a conventional one-stage system, a two-stage system in which the first filter is given by a low-order approximation to (A.3), a two-stage system in which the first filter is given by a high-order FIR implementation of (A.3), and a two-stage system in which the first filter is given by the HRIR of the center of the cloud. For each of the four architectures, the low-order filters are implemented either by FIR or IIR filters, as described in § A.1.2, yielding a total of eight different systems to compare to the baseline.

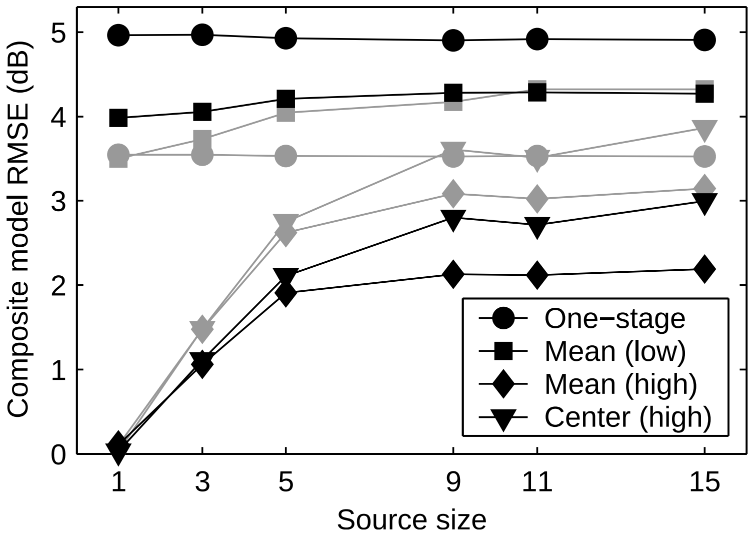


Figure A.3: Composite RMSE for eight binaural implementations. Black curves give the RMSE of 12th order FIR filters and grey curves give the RMSE of 6th order IIR filters designed via BMT.

A.2.2 Stimulus Conditions

We compare the eight systems described above to the baseline system for six cloud sizes, $D = (1, 3, 5, 9, 11, 15)$. The point-source locations are shown in Fig. A.2, where the azimuthal separation between adjacent cloud points is 10°, and the elevation separation is 18°. Clouds are rendered with azimuths from 0° to 160° in 20° increments, and elevations from $-18°$ to 54° in 18° increments, for a total of 45 cloud center locations.

A.2.3 Objective Evaluation

We use a perceptual RMS error metric based on the spectral mismatch between the measured HRTF and the composite response of the one- or two-stage filters (60). The metric is computed by warping the two spectra to a logarithmic frequency scale, smoothing with a 0.2 octave filter, converting to a decibel scale, and computing the RMS difference. Fig. A.3 gives the composite RMSE averaged across both ears,

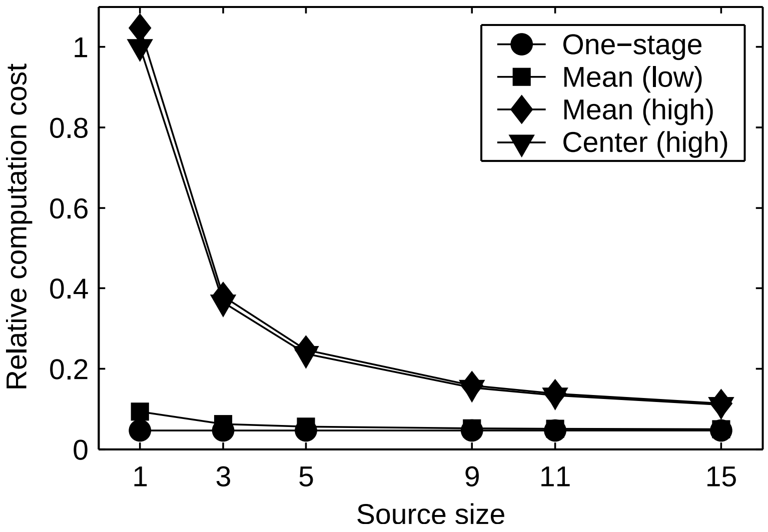


Figure A.4: Ratio of the computational cost of four binaural implementations to the cost of a full-length one-stage implementation.

45 center locations, and eight HRTF sets. The abscissa represents the cloud size, D, and the ordinate represents the composite RMSE. The parameter is the filter architecture and type. FIR filter implementations are shown in black and BMT IIR filter implementations in grey. The four architectures are indicated by marker shape.

The composite RMSE of the six two-stage implementations included in Fig. A.3 increase with cloud size, and the composite RMSE of the two one-stage implementations are constant with cloud size, as expected. The one-stage FIR implementation yields the poorest performance, with a constant RMSE of about 5 dB. For the one-stage implementation, the BMT IIR filter implementation yields a consistent 1.5 dB performance improvement. For the two-stage implementations however, there is either little change in performance (for the low-order case), or substantially worse performance (for the high-order case) when using BMT. This result is somewhat unexpected and is discussed further below. The four implementations that use a high-order implementation of $C_{L,R}(\omega)$ yield the best performance. As the cloud

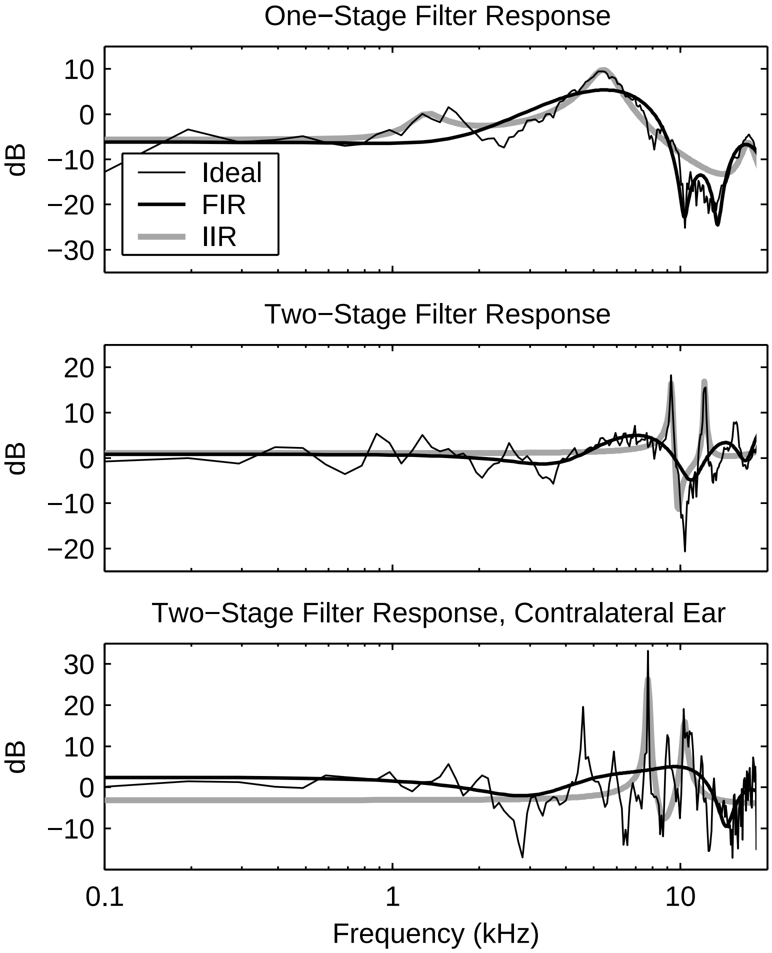


Figure A.5: The ideal magnitude response (thin line), the low-order FIR response (thick black line), and the low-order IIR response (grey line) for three filter implementations.

size shrinks to one source, all four implementations reduce to realizing the single, measured HRTF.

Fig. A.4 gives the computational savings of the eight filter implementations relative to a direct $N = 256$ implementation. Note that because $N = 12$ for the FIR filters and $N = 6$ for the IIR filters, the net computational cost for the two is equivalent. The one-stage and low-order two-stage implementations yield a computational cost that is less than one tenth that of the direct implementation for all cloud sizes.

The high-order two-stage implementations offer no computational savings for $D = 1$, but as D increases, the computational savings improve, such that for $D \geq 9$ the computational cost is less than one fifth that of the direct implementation.

Let us return to the comparison between FIR and IIR filter designs. Fig. A.5 gives the log magnitude response of both low-order FIR and BMT IIR filter approximations for three 'ideal' filter responses. For each panel, magnitude responses for three filters are shown: ideal (thin black), FIR approximation (thick black) and IIR approximation (thick grey). The top panel gives the log magnitude HRTF for $(\theta, \phi) = (0°, 36°)$. In this case, the IIR filter clearly approximates the 'ideal' response more accurately, particularly below 10 kHz. The middle panel shows the response of the second stage of a two-stage implementation in which $C_L(\omega) = H_L(\omega \,|\, 0°, 18°)$ for cloud point $(0°, 36°)$. The IIR filter again approximates the ideal response more accurately, particularly the narrow spectral peaks and notches at 9, 10 and 12 kHz. Indeed, BMT is well-suited to modeling HRTFs due to its ability to realize narrow spectral peaks and notches even at low-order.

As the ideal response becomes more complicated, particularly in the contralateral ear where measurement SNR is problematic, the BMT method begins to break down. The bottom panel of Fig. A.5 gives one such case, showing the response of the second stage of a two-stage implementation for the contralateral ear with $C_L(\omega) = H_L(\omega \,|\, 80°, -18°)$ for cloud point $(80°, 0°)$. The ideal response, $H_L(\omega \,|\, 80°, 0°)/ H_L(\omega \,|\, 80°, -18°)$ is quite noisy in this case. The BMT IIR accurately approximates the narrow spectral peak at 8 kHz, but at the expense of the rest of the spectrum. In this case, the FIR filter more accurately approximates the general shape of the ideal response, especially at low frequency.

A.3 Perceptual Observations

We have conducted two informal listening tests that verify the objective results presented above. For both tests the stimulus was a 20ms burst of pink noise windowed with the falling half of a Hanning window. Binaural clouds were created by filtering the noise burst for each spatial location within the cloud, and then using a Poisson process to determine each event's occurrence time within a five-second observation interval. The resulting source is that of a cloud of "popcorn pops" with a variable number of events over the five-second period.

We first conducted a similarity test, in which a cloud rendered using a minimum-phase reconstruction of the measured HRTFs was compared with one rendered using one of the eight implementations described above. Similarity was measured on an absolute scale from one to ten, with ten being the most similar. The four high-order two-stage implementations yielded the best results for all cloud sizes and locations, with similarity consistently between eight and ten. The FIR and IIR filter implementations generally yielded equivalent performance, although for some spatial locations the IIR implementation created a quiet, yet noticeable, high-frequency ping. Of the low-order implementations, both one- and two-stage, the worst performer was the one-stage FIR design (similarity ~ 4) and the best performer was the one-stage IIR design (similarity ~ 8).

We also conducted an informal stimulus sample discrimination task (SSD) and found that the high-order two-stage designs were difficult to distinguishable from the direct $N = 256$ implementation, whereas the other designs were easily discriminated. These results were observed when using identical noise bursts for rendering each cloud. We also conducted a listening test in which a new instantiation of the noise burst was used for each cloud. In this case, detecting the use of low-order approximations essentially fell to chance. This suggests that the spectral coloration obtained as a result of errors in the low-order filter approximation is on the same order as the spectral variation across different instantiations of the noise process.

A.4 Sample Application: Moving Sound Source

Auralizing clouds of 'popcorn pops' is not the only application of this binaural display. The system is also well suited to the auralization of moving sound sources. By auralizing a monaural signal for several nearby locations in space, we can display a moving sound source by taking a time-varying linear combination of the binaural signals output by the system. The weights used in the linear combinations can be computed using VBAP (86; 67). Let the desired source location in rectangular coordinates be given by $\mathbf{d} = [d_x, d_y, d_z]'$. Let the three closest HRTF sample locations that form a triangle around $\mathbf{d}$ be given by $\mathbf{p_1} = [p_{1,x}, p_{1,y}, p_{1,z}]'$. The weights for the three binaural signals to combine are then given by

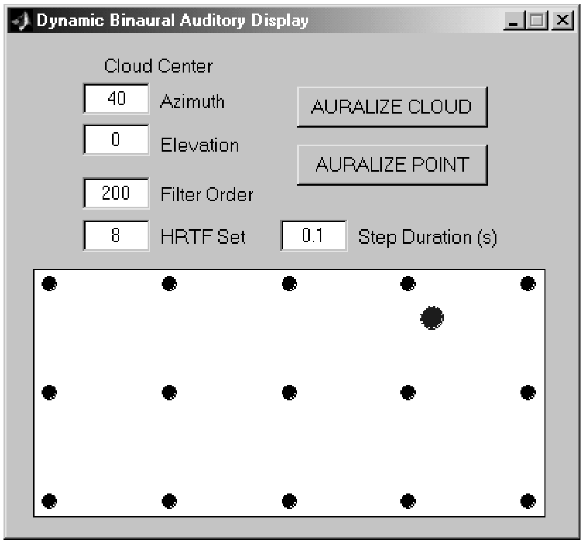
$$\mathbf{w} = \mathbf{d'} \cdot [\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}]^{-1} \tag{A.5}$$

and normalized by,

$$\mathbf{w'} = \frac{\mathbf{w}}{\|\,\mathbf{w}\,\|_p} \tag{A.6}$$

where $1 \leq p \leq 2$ is an adjustable parameter dependent upon the degree of coherence between the combined binaural signals.

Using this method, a realtime moving-source binaural display was implemented in MATLAB. A screen shot of the GUI is shown in figure A.6. The display filters a monaural audio file for 15 nearby spatial locations off-line. The user can drag the blue circle around the region of space supported by the 15 sampled locations. The



Dynamic Binaural Auditory Display

Cloud Center

40 Azimuth

0 Elevation

200 Filter Order

8 HRTF Set

0.1 Step Duration (s)

Figure A.6: Screen shot of a MATLAB GUI that auralizes a monaural audio file with user-controlled position in realtime.

perceived sound object moves smoothly through space with the blue circle. Informal listening tests verify that the sound object moves smoothly through virtual space with modest system latency[3].

The HRTF filtering is performed off-line in the current implementation. The filtering could be incorporated into the realtime engine, but the MATLAB implementation places tight computational limits on the number of operations per audio sample. Nonetheless, a preliminary investigation indicates that a low-order filtering can be performed in realtime. Another limitation of the current implementation is the incorporation of ITD. The ITD is added independently to all 15 binaural signals. This gives the final interpolated signal a more diffuse quality than it would otherwise have if the ITD was interpolated separately (67). Nonetheless, this application demonstrates the utility of filtering a monaural signal for multiple points in space in

[3] The latency for this MIATLAB GUI is about 100 ms, and is due to the requirement that the audio buffer never empty. The maximum rate at which MATLAB (running on an IBM ThinkPad T30) can sample the GUI state and output a frame of audio data appears to be about ten times per second. Wenzel has found that latencies up to 100 ms are relatively benign for dynamic binaural auditory displays (136).

order to facilitate realtime moving sound source auralization.

A.5 Conclusion

We have described a binaural display that renders clouds of nearby point sources by filtering a monaural source with multiple HRTF pairs. We argued that such a display can create more convincing and flexible virtual sound objects, by alleviating HRTF individuation and facilitating moving sources via the phantom source technique. We proposed a two-stage filter implementation in which the first filter is common to all spatial locations in the cloud, and the second filter provides the spectral cues unique to each location within the cloud. We found that much of the computational burden can be shifted to the first filter, in which case the net computational complexity of the display is reduces to that of a conventional binaural display.

APPENDIX B

Order Reduction Algorithms

Algorithms have been previously published for both Balanced Model Truncation (BMT) (116; 9) and Hankel-norm Optimal Approximation (HOA) (117; 154)[1]. In the interest of completeness and reproducibility the algorithms have been adapted to the binaural display application and are described below in detail. In particular, the HOA algorithm is somewhat involved. Published HOA algorithms typically only provide a broad overview that does not describe several difficult steps. Each step of the HOA method is described below, with additional references for individual sub-algorithms as necessary.

The principal step in BMT is the computation of the SVD of Hankel matrix $\mathcal{H}$. This computation is straightforward, but $\mathcal{H}$ is often a large matrix and memory consumption may be an issue. In contrast, HOA operates entirely on the system matrices, hence memory consumption is not an issue. However, HOA may be more time consuming to compute, as solving Lyapunov equations and stable projections are computationally intensive steps for large systems.

[1] Both algorithms are included in the Robust Controls Toolbox, an extension of the Controls Toolbox, which is itself an extension of the MATLAB software package.

B.1 Balanced Model Truncation

Balanced Model Truncation (BMT) is a state-space model reduction technique that operates by discarding all but the N largest singular values of a balanced system (116). If the original system is given in state-space form, then it is *balanced* before truncation. For HRTF modeling, this is not necessary as the original system is a collection of transfer functions. In this case the Hankel matrix $\mathcal{H}$ is given by the block impulse response of the HRTFs. A balanced system is then constructed from the SVD of $\mathcal{H}$ (116).

The BMT algorithm described below constructs an order N state-space system from a collection of measured HRIRs of order N_0. Both the originial HRIR filter array and the reduced-order state-space system have M inputs and P outputs.

1. Prepend a single zero to the beginning of each HRIR, as the state-space system in (2.1) has no feed-through path (116; 9). For the MIMO architecture, the time-delay of the contralateral HRIRs must be included, but is removed for the SIMO and MISO architectures. In all other respects, the BMT algorithm is independent of system architecture.

2. Arrange the HRIRs into the desired matrix impulse response, $\mathbf{h}[n{=}0, 1, \cdots N_0{+}1]$, where $\mathbf{h}[n]$ is $P \times M$, and construct the finite Hankel matrix $\mathcal{H}$

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}[1] & \mathbf{h}[2] & \mathbf{h}[3] & \dots & \mathbf{h}[N_0{+}1] \\ \mathbf{h}[2] & \mathbf{h}[3] & \mathbf{h}[4] & \dots & \mathbf{0} \\ \mathbf{h}[3] & \mathbf{h}[4] & \mathbf{h}[5] & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}[N_0{+}1] & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} \tag{B.1}$$

The size of $\mathcal{H}$ is $P(N_0{+}1) \times M(N_0{+}1)$[2].

3. Compute the singular value decomposition (SVD) of $\mathcal{H}$

$$\mathcal{H} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{B.2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices with dimensions $P(N_0{+}1) \times P(N_0{+}1)$ and $M(N_0{+}1) \times M(N_0{+}1)$, respectively, and $\mathbf{S}$ is a diagonal matrix of size $P(N_0{+}1) \times M(N_0{+}1)$ with the singular values, $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R)$, arranged along the main diagonal, where $R = \min\big(P(N_0{+}1), M(N_0{+}1)\big)$.

4. Weight the singular vectors, given by the rows of $\mathbf{U}$ and $\mathbf{V}$, with the square root of the singular values[3]: $\widetilde{\mathbf{U}} = \mathbf{U}\mathbf{S}^{1/2}$ and $\widetilde{\mathbf{V}}^T = \mathbf{S}^{1/2}\mathbf{V}^T$. $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ both have dimensions $P(N_0 + 1) \times M(N_0 + 1)$. Hence $\mathcal{H} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$, where $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are orthogonal but not orthnomornal.

5. Partition the SVD so as to retain only the N largest singular values[4]

$$\widetilde{\mathbf{U}} = \begin{bmatrix} \widetilde{\mathbf{U}}_1 & \ldots \\ \widetilde{\mathbf{U}}_2 & \ldots \\ \widetilde{\mathbf{U}}_3 & \ldots \end{bmatrix}, \quad \widetilde{\mathbf{V}} = \begin{bmatrix} \widetilde{\mathbf{V}}_1 & \ldots \\ \vdots & \ddots \end{bmatrix} \tag{B.3}$$

where $\widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{U}}_3$ have dimension $P \times N$, $\widetilde{\mathbf{U}}_2$ has dimension $P(N_0 - 1) \times N$, and $\widetilde{\mathbf{V}}_1$ has dimension $M \times N$.

[2] $\mathcal{H}$ may be very large depending upon (P, M, N_0), and computing its singular value decomposition may require a prohibitive amount of memory. However, almost half of $\mathcal{H}$ is zero, and given that the HRIRs are minimum-phase, most of the energy in $\mathcal{H}$ is contained in the upper-left corner. We found that the results are not affected substantially if only the upper-left quarter of $\mathcal{H}$ is used for BMT. All results given in the present study are computed using the full Hankel matrix.

[3] The square-root operation is 'element-wise,' as $\mathbf{S}$ is not necessarily square.

[4] The SVD is the most computation-intensive part of the BMT algorithm. Because only the N largest singular values are used to construct the final state-space system, there is no need to compute the entire SVD. Only the N largest singular values, along with their singular vectors, need to be computed.

6. Construct an order N state-space system from $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$. The system matrices $\big(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}\big)$ are given by

$$\begin{aligned} \widehat{\mathbf{A}} &= \left(\begin{bmatrix} \widetilde{\mathbf{U}}_1 \\ \widetilde{\mathbf{U}}_2 \end{bmatrix}^T \begin{bmatrix} \widetilde{\mathbf{U}}_1 \\ \widetilde{\mathbf{U}}_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} \widetilde{\mathbf{U}}_1 \\ \widetilde{\mathbf{U}}_2 \end{bmatrix}^T \begin{bmatrix} \widetilde{\mathbf{U}}_2 \\ \widetilde{\mathbf{U}}_3 \end{bmatrix} \\ \widehat{\mathbf{B}} &= \widetilde{\mathbf{V}}_1^T, \qquad \widehat{\mathbf{C}} = \widetilde{\mathbf{U}}_1 \end{aligned} \tag{B.4}$$

where $\widehat{\mathbf{A}}$ is $N \times N$, $\widehat{\mathbf{B}}$ is $N \times M$, and $\widehat{\mathbf{C}}$ is $P \times N$.

B.2 Hankel-Norm Optimal Approximation

The Hankel-norm Optimal Approximation (HOA) algorithm also designs a low-order system by discarding all but the N largest singular values of the original system, albeit in such a way as to minimize the Hankel error of the resulting system. For the SISO case, closed-form HOA algorithms have been published (123; 125). These studies use HOA to construct low-order IIR filters from FIR filters. For the MIMO case the algorithm is considerably more dense. The HOA algorithm is further complicated if the original system contains repeated singular values. This complication is neglected in the algorithm below, as measured HRIRs invariably contain sufficient white observation noise to preclude any repeated singular values (119). The observation noise also ensures that the system is minimal.

The HOA algorithm described below constructs an order N state-space system from a collection of measured HRIRs of order N_0. Both the original HRIR filter array and the reduced-order state-space system have M inputs and P outputs.

1. Prepend a single zero to the beginning of each HRIR, as in the BMT algorithm. In addition to removing the need for a feed-through path in the state-space

system, prepending a zero also transforms the transfer functions so as to be strictly proper, a requirement for HOA (117). For the MIMO architecture, the time-delay in the contralateral HRIRs must be included, but is removed for the SIMO and MISO architectures.

2. Construct a high-order state-space system that implements the measured HRIRs exactly. The HOA algorithm operates directly on the system matrices ($\mathbf{A}$,$\mathbf{B}$,$\mathbf{C}$), hence it is necessary to realize the HRIR filter array as a state-space system prior to performing order reduction. This is readily accomplished with the controller canonical form (155). Without loss of generality, consider a system with more outputs than inputs[5], $P \geq M$. The controller canonical realization of this filter array is an order $M(N_0+1)$ state-space system. The $\mathbf{A}_0$ matrix is $M(N_0+1) \times M(N_0+1)$, the $\mathbf{B}_0$ matrix is $M(N_0+1) \times M$, and both matrices are block diagonal

$$\mathbf{A}_0 = \begin{bmatrix} \mathbf{I}' & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}' & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}' \end{bmatrix} \qquad \mathbf{B}_0 = \begin{bmatrix} \underline{1} & \underline{0} & \dots & \underline{0} \\ \underline{0} & \underline{1} & \dots & \underline{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{0} & \underline{0} & \dots & \underline{1} \end{bmatrix} \tag{B.5}$$

where

$$\mathbf{I}' = \begin{bmatrix} 0 & 0 & 0 & 0 & & \cdot \\ 1 & 0 & 0 & & \cdot & \\ 0 & 1 & & \cdot & & 0 \\ 0 & & \cdot & & 0 & 0 \\ & \cdot & & 1 & 0 & 0 \\ \cdot & & 0 & 0 & 1 & 0 \end{bmatrix} \qquad \underline{1} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{B.6}$$

[5] If $M > P$, swap inputs with outputs, perform HOA, and swap back when done: $\widehat{\mathbf{A}} = \mathbf{A}^T$, $\widehat{\mathbf{B}} = \mathbf{C}^T$, $\widehat{\mathbf{C}} = \mathbf{B}^T$.

$\mathbf{I}'$ has dimension $(N_0+1)\times(N_0+1)$, and $\underline{1}$ has dimension $(N_0+1)\times 1$. The $\mathbf{C}_0$ matrix is $P \times M(N_0+1)$ and is constructed from the measured HRIRs

$$\begin{aligned}
\mathbf{C}_0 \;=\; & \left[\begin{matrix} h_{11}[1] & h_{11}[2] & \ldots & h_{11}[N_0+1] \\ h_{21}[1] & h_{21}[2] & \ldots & h_{21}[N_0+1] \\ \vdots & \vdots & \ddots & \vdots \\ h_{P1}[1] & h_{P1}[2] & \ldots & h_{P1}[N_0+1] \end{matrix}\right. \quad \ldots \\
& \qquad \left.\begin{matrix} h_{12}[1] & h_{12}[2] & \ldots & h_{1M}[N_0+1] \\ h_{22}[1] & h_{22}[2] & \ldots & h_{2M}[N_0+1] \\ \vdots & \vdots & \ddots & \vdots \\ h_{P2}[1] & h_{P2}[2] & \ldots & h_{PM}[N_0+1] \end{matrix}\right]
\end{aligned} \tag{B.7}$$

where $h_{pm}[n]$ is the impulse response between input m and output p.

3. Convert the discrete-time system above to continuous-time. The HOA algorithm is simpler in continuous-time, and it is common even when designing low-order discrete-time systems to convert the original system to continuous-time using a bilinear transform, and then convert back to discrete-time after performing HOA (117; 119; 156). A discrete-time HOA algorithm is given in (154), although it is more dense than the algorithm below. The bilinear transform to continuous-time is

$$\begin{aligned}
\mathbf{A}_c &= \left(\mathbf{I}+\mathbf{A}_0\right)^{-1}\left(\mathbf{A}_0-\mathbf{I}\right) \\
\mathbf{B}_c &= \sqrt{2}\left(\mathbf{I}+\mathbf{A}_0\right)^{-1}\mathbf{B}_0 \\
\mathbf{C}_c &= \sqrt{2}\,\mathbf{C}_0\left(\mathbf{I}+\mathbf{A}_0\right)^{-1}
\end{aligned} \tag{B.8}$$

4. The controllability and observability Gramians, $\mathcal{P}$ and $\mathcal{Q}$, of the system $(\mathbf{A_c}, \mathbf{B_c}, \mathbf{C_c})$

are defined as

$$\begin{aligned} \mathcal{P} &\triangleq \int_0^\infty e^{\mathbf{A}_c t}\, \mathbf{B}_c\, \mathbf{B}_c^*\, e^{\mathbf{A}_c^* t}\, dt \\ \mathcal{Q} &\triangleq \int_0^\infty e^{\mathbf{A}_c^* t}\, \mathbf{C}_c^*\, \mathbf{C}_c\, e^{\mathbf{A}_c t}\, dt \end{aligned} \tag{B.9}$$

Both Gramians are size $M(N_0{+}1)\times M(N_0{+}1)$. The system $(\mathbf{A}_c,\mathbf{B}_c,\mathbf{C}_c)$ is stable, and the eigenvalues of $\mathbf{A}_c$ are strictly in the left half of the complex plane. Hence the integrals above converge. However, this is a numerically prohibitive integral to evaluate. The Gramians are typically computed by considering the corresponding matrix differential equations, which yield the following linear equations, known as the Lyapunov equations

$$\begin{aligned} \mathbf{A}_c\mathcal{P} + \mathcal{P}\mathbf{A}_c^* + \mathbf{B}_c\mathbf{B}_c^* &= 0 \\ \mathbf{A}_c^*\mathcal{Q} + \mathcal{Q}\mathbf{A}_c + \mathbf{C}_c^*\mathbf{C}_c &= 0 \end{aligned} \tag{B.10}$$

An efficient algorithm for solving matrix equations of this form is given in (157). Because the system $(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)$ is in controller canonical form, the controllability Gramian is simply the identity matrix, and this simplifies the computation of a balancing transform somewhat (117). The observability Gramian $\mathcal{Q}$ must be computed by solving the Lyapunov equation.

5. Find a balancing transform $\mathbf{T}$ from the observability Gramian $\mathcal{Q}$. There are many balancing transforms, however care must be taken in choosing a transform, as they may be ill-conditioned (65). The transform below is well-conditioned for all of the HRTF data used in the present study (126).

   Compute the SVD of $\mathcal{Q}$

$$\mathcal{Q} = \mathbf{V}\mathbf{S}\mathbf{V}^T \tag{B.11}$$

where $\mathbf{V}$ is a unitary matrix and the singular values of $\mathcal{Q}$, $(\sigma_1 > \sigma_2 > \cdots > \sigma_{M(N_0+1)})$, are arranged along the diagonal of $\mathbf{S}$. The symmetry of the SVD in this case is due to the symmetry of the $\mathcal{Q}$. Let

$$\mathbf{U} = \mathbf{V}\mathbf{S}^{1/4} \tag{B.12}$$

The matrix $\mathbf{U}^T$ is itself a balancing transform. However, we seek to isolate the state that corresponds to the $(N+1)^{\text{th}}$ largest singular value. We move this state to the end of the state-vector by permuting the columns of the transform matrix

$$\widetilde{\mathbf{U}} = \big[\underline{u}_1 \ \cdots \ \underline{u}_N, \ \underline{u}_{N+2} \ \cdots \ \underline{u}_{M(N_0+1)}, \ \underline{u}_{N+1}\big] \tag{B.13}$$

where $\underline{u}_k$ is the k^{th} column of $\mathbf{U}$. The balancing transform is then given by $\mathbf{T} = \widetilde{\mathbf{U}}^T$.

6. Balance the system using similarity transform $\mathbf{T}$

$$\begin{aligned} \mathbf{A}_b &= \mathbf{T}\,\mathbf{A}_c\,\mathbf{T}^{-1} \\ \mathbf{B}_b &= \mathbf{T}\,\mathbf{B}_c \\ \mathbf{C}_b &= \mathbf{C}_c\,\mathbf{T}^{-1} \end{aligned} \tag{B.14}$$

7. Partition the matrices so as to isolate the state corresponding to the $(N+1)^{\text{th}}$ singular value

$$\mathbf{A}_b = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad \mathbf{B}_b = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \quad \mathbf{C}_b^T = \begin{bmatrix} \mathbf{C}_1^T \\ \mathbf{C}_2^T \end{bmatrix} \tag{B.15}$$

where $\mathbf{A}_{11}$ has dimensions $\big(M(N_0+1)-1\big) \times \big(M(N_0+1)-1\big)$, $\mathbf{A}_{12}$ and $\mathbf{A}_{21}$ are vectors, and $\mathbf{A}_{22}$ is a scalar. Matrix $\mathbf{B}_1$ has dimensions $\big(M(N_0+1)-1\big) \times M$, matrix $\mathbf{C}_1$ has dimensions $P \times \big(M(N_0+1)-1\big)$, and $\mathbf{B}_2$ and $\mathbf{C}_2$ are vectors.

8. Let $\mathbf{\Gamma} = \mathbf{\Sigma}_1 - \sigma_{N+1}\mathbf{I}$, where

$$\mathbf{\Sigma}_1 = \operatorname{diag}\big(\sigma_1 \ \cdots \ \sigma_N, \ \sigma_{N+2} \ \cdots \ \sigma_{P(N_0+1)}\big) \tag{B.16}$$

   and σ_k is the k^{th} singular value of $\mathcal{Q}$. Also let $\mathbf{W} = (\underline{C}_2^T)^\dagger \underline{B}_2$, where † represents the Moore-Penrose pseudoinverse[6].

9. Construct the following order $M(N_0+1)-1$ system

$$\begin{aligned}
\widetilde{\mathbf{A}} &= \mathbf{\Gamma}^{-1}\big(\mathbf{A}_{11}\sigma_{N+1} + \mathbf{\Sigma}_1\mathbf{A}_{11}\mathbf{\Sigma}_1 + \sigma_{N+1}\mathbf{C}_1^T\mathbf{W}\mathbf{B}_1^T\big) \\
\widetilde{\mathbf{B}} &= \mathbf{\Gamma}^{-1}\big(\mathbf{\Sigma}_1\mathbf{B}_1 - \sqrt{\sigma_{N+1}}\mathbf{C}_1^T\mathbf{W}\big) \\
\widetilde{\mathbf{C}} &= \mathbf{C}_1\mathbf{\Sigma}_1 - \sqrt{\sigma_{N+1}}\mathbf{W}\mathbf{B}_1^T
\end{aligned} \tag{B.17}$$

   which is referred to as an *all pass dilation* of $\big(\mathbf{A}_b, \mathbf{B}_b, \mathbf{C}_b\big)$, as the spectral error between the two systems is constant

$$\sigma_{\max}\big(\mathbf{T}_b(j\omega) - \widetilde{\mathbf{T}}(j\omega)\big) = \sqrt{\sigma_{N+1}} \tag{B.18}$$

   where $\mathbf{T}_b(j\omega)$ and $\widetilde{\mathbf{T}}(j\omega)$ are the matrix transfer functions for the two systems.

10. The system $\big(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{C}}\big)$ is not stable. Of the $M(N_0+1)-1$ eigenvalues in this system, exactly N are stable. It is necessary to extract the stable order N subsystem from $\big(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{C}}\big)$. There are several methods for accomplishing this. The most direct method is to compute the modal decomposition of $\widetilde{\mathbf{A}}$ (134), and constructing a diagonal $\widehat{\mathbf{A}}_c$ from the N eigenvalues with negative real part. The N corresponding eigenvectors similarly transform $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{C}}$. A more involved method that yields real system matrices is described below (117; 158).

[6] This step is commonly stated as: find a unitary matrix $\mathbf{W}$ that satisfies $\mathbf{C}_2^T\mathbf{W} = \mathbf{B}_2$. The pseudoinverse provides one possible solution (117; 158).

(a) Compute the complex Schur decomposition of $\widetilde{\mathbf{A}}$. Find a unitary matrix $\mathbf{U}$ such that $\mathbf{U}^*\widetilde{\mathbf{A}}\mathbf{U} = \widetilde{\mathbf{A}}_t$ is a triangular matrix with diagonal elements given by the eigenvalues of $\widetilde{\mathbf{A}}$.

(b) It is necessary to transform $\widetilde{\mathbf{A}}_t$ such that the N stable eigenvalues appear as the first N diagonal elements. This can be accomplished by applying a sequence of *Givens rotations* to the $\widetilde{\mathbf{A}}_t$ and $\mathbf{U}$ matrices (135), yielding a new unitary transform matrix $\widetilde{\mathbf{U}}$

$$\widetilde{\mathbf{U}}^*\widetilde{\mathbf{A}}\widetilde{\mathbf{U}} = \widetilde{\mathbf{A}}_p = \begin{bmatrix} \widetilde{\mathbf{A}}_{11} & \widetilde{\mathbf{A}}_{12} \\ \mathbf{0} & \widetilde{\mathbf{A}}_{22} \end{bmatrix} \tag{B.19}$$

where $\widetilde{\mathbf{A}}_p$ is real and $\widetilde{\mathbf{A}}_{11}$ is $N \times N$.

(c) Find a matrix $\mathbf{X}$ that satisfies

$$\widetilde{\mathbf{A}}_{11}\mathbf{X} - \mathbf{X}\widetilde{\mathbf{A}}_{22} + \widetilde{\mathbf{A}}_{12} = \mathbf{0} \tag{B.20}$$

This is similar to solving the continuous-time Lyapunov equations, and the algorithm in (157) can be used.

(d) The system matrices of an order N HOA continuous-time system are given by

$$\begin{aligned} \widehat{\mathbf{A}}_c &= \widetilde{\mathbf{A}}_{11} \\ \widehat{\mathbf{B}}_c &= [\mathbf{I},\ -\mathbf{X}]\,\widetilde{\mathbf{U}}^*\,\widetilde{\mathbf{B}} \\ \widehat{\mathbf{C}}_c &= \widetilde{\mathbf{C}}\,\widetilde{\mathbf{U}} \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \end{aligned} \tag{B.21}$$

where $\mathbf{I}$ is $N \times N$.

11. The system matrices of the final order N discrete-time system are given by the

bilinear transformation of the continuous-time solution

$$\begin{aligned}
\widehat{\mathbf{A}} &= \left(\mathbf{I} + \widehat{\mathbf{A}}_c\right)\left(\mathbf{I} - \widehat{\mathbf{A}}_c\right)^{-1} \\
\widehat{\mathbf{B}} &= \sqrt{2}\left(\mathbf{I} - \widehat{\mathbf{A}}_c\right)^{-1}\widehat{\mathbf{B}}_c \\
\widehat{\mathbf{C}} &= \sqrt{2}\,\widehat{\mathbf{C}}_c\left(\mathbf{I} - \widehat{\mathbf{A}}_c\right)^{-1} \qquad \text{(B.22)}
\end{aligned}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization.* Cambridge, MA: MIT Press, 1997.

[2] D. Begault and E. Wenzel, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on Spatial Perception of a Virtual Speech Source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[3] D. Zotkin, R. Duraiswami, and L. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.

[4] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, Sept. 1999.

[5] V. Algazi, R. Duda, and D. Thompson, "Motion Tracked Binaural Sound," *J. Audio Eng. Soc.*, vol. 52, no. 11, pp. 1142–1156, Nov. 2004.

[6] S. Takane, Y. Suzuki, T. Miyajime, and T. Sone, "ADISE: A new method for high definition virtual acoustic display," in *Proc. Int. Conf. on Auditory Display*, 2002, Kyoto, Japan.

[7] N. Adams and G. Wakefield, "The binaural display of clouds of point sources," *Proc. IEEE Workshop on App. of Signal Processing to Audio and Acoust.*, October 2005, New Paltz, NY.

[8] P. Georgiou and C. Kyriakakis, "Modeling of Head Related Transfer Functions for Immersive Audio Using a State-Space Approach," in *Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, vol. 1, 1999, pp. 720–724.

[9] D. Grantham, J. Willhite, K. Frampton, and D. Ashmead, "Reduced order modeling of head related impulse responses for virtual acoustic displays," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3116–3125, May 2005.

[10] P. Chanda, S. Park, and T. Kang, "A Binaural Synthesis with Multiple Sound Sources Based on Spatial Features of Head-related Transfer Functions," in *Proc. IEEE Int. Joint Conf. on Neural Networks*, 2006.

[11] J. W. Strutt (Baron Rayleigh), *The Theory of Sound, v.1 & 2.* New York: Dover, 1945 (*1st ed.* 1878).

[12] S. Carlile, "The Physical and Psychophysical Basis of Sound Localization," in *Virtual Auditory Space: Generation and Applications*, S. Carlile, Ed. Berlin, Germany: Springer-Verlag, 1996.

[13] R. Woodworth and H. Schlosberg, *Experimental Psychology.* New York: Holt, Rinehart and Winston, 1962.

[14] G. Plenge, "On the differences between localization and lateralization," *J. Acoust. Soc. Am.*, vol. 56, no. 3, pp. 944–951, Sept. 1974.

[15] S. Carlile, "Auditory Space," in *Virtual Auditory Space: Generation and Applications*, S. Carlile, Ed. Berlin, Germany: Springer-Verlag, 1996.

[16] R. Litovsky, H. Colburn, W. Yost, and S. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, October 1999.

[17] C. Cheng, "Visualiation, Measurement, and Interpolation of Head-Related Transfer Functions (HRTF'S) with Applications in Electro-Acoustic Music," Ph.D. dissertation, Univ. of Michigan, Ann Arbor, 2001.

[18] R. Duda and W. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, vol. 104, pp. 3048–3058, 1998.

[19] F. Wightman and D. Kistler, "Headphone simulation of free-field listening. i: Stimulus synthesis," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 858–867, Feb. 1989.

[20] V. Algazi, C. Avendano, and R. Duda, "Elevation localization and head-related transfer functions analysis at low frequencies," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1110–1122, Mar. 2001.

[21] J. Middlebrooks, J. Makous, and D. Green, "Directional sensitivity of sound-pressure levels in the human ear canal," *J. Acoust. Soc. Am.*, vol. 86, pp. 89–108, 1989.

[22] D. Hammershøi and H. Møller, "Sound transmission to and within the human ear canal," *J. Acoust. Soc. Am.*, vol. 100, pp. 408–427, 1996.

[23] V. Algazi, R. Duda, R. Duraiswami, N. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053–2064, Nov. 2002.

[24] W. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," *Technical Report, MIT*, 1994, http://sound.media.mit.edu/KEMAR.html.

[25] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, “Localization using nonindividualized head-related transfer functions,” *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, July 1993.

[26] F. Wightman and D. Kistler, “Headphone simulation of free-field listening. ii: Psychophysical validation,” *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 868–878, Feb. 1989.

[27] H. Møller, M. Sørensen, C. Jensen, and D. Hammershøi, “Binaural Technique: Do We Need Individual Recordings?” *J. Audio Eng. Soc.*, vol. 44, pp. 451–469, June 1996.

[28] W. Martens, “Perceptual evaluation of filters controlling source direction: Customized and generalized HRTFs for binaural synthesis,” *Acoust. Sci. & Tech.*, vol. 24, no. 5, pp. 220–232, October 2003.

[29] J. Middlebrooks, “Individual differences in external-ear transfer functions reduced by scaling in frequency,” *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1480–1492, Sept. 1999.

[30] N. Durlach, B. Shinn-Cunningham, and R. Held, “Supernormal Auditory Localization,” *Presence*, vol. 2, no. 2, pp. 89–103, 1993.

[31] W. Rabinowitz, J. Maxwell, Y. Shao, and M. Wei, “Sound Localization Cues for a Magnified Head: Implications from Sound Diffration about a Rigid Sphere,” *Presence*, vol. 2, no. 2, pp. 125–129, 1993.

[32] S. Shimada, N. Hayashi, and S. Hayashi, “A Clustering Method for Sound Localization Transfer Functions,” *J. Audio Eng. Soc.*, vol. 42, no. 7/8, pp. 577–583, 1994.

[33] N.-M. Cheung, S. Trautman, and A. Horner, “Head-Related Transfer Function Modeling in 3-D Sound Systems with Genetic Algorithms,” *J. Audio Eng. Soc.*, vol. 46, no. 6, pp. 531–539, June 1998.

[34] J. Middlebrooks, “Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency,” *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, Sept. 1999.

[35] J. Middlebrooks, E. Macpherson, and Z. Onsan, “Psychophysical customization of directional transfer functions for virtual sound localization,” *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3088–3091, Dec. 2000.

[36] C. Jin, A. van Schaik, V. Best, and S. Carlile, “Individualization in spatial-audio coding,” in *Proc. of the IEEE Workshop on App. of Signal Processing to Audio and Acoust.*, 2003.

[37] V. Algazi, P. Divenyi, V. Martinez, and R. Duda, "Subject Dependent Transfer Functions in Spatial Hearing," in *Proc. of IEEE Midwest Sym. on Circuits and Systems*, 1997.

[38] R. Duraiswami, R. Duda, and V. A. et al., "Creating virtual spatial audio via scientific computing and computer vision," in *Proc. of 140th ASA meeting*, 2000, http://www.acoustics.org/press/140th/duraiswami.htm.

[39] V. Algazi, C. Avendano, and R. Duda, "Estimation of a spherical-head model from anthropometry," *J. Audio Eng. Soc.*, vol. 49, pp. 472–479, 2001.

[40] B. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *J. Acoust. Soc. Am.*, vol. 107, no. 3, pp. 1627–1636, Mar. 2000.

[41] C. Brown and R. Duda, "A Structural Model for Binaural Sound Synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sept. 1998.

[42] E. Shaw, "Acoustical features of the human external ear," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. Gilkey and T. Anderson, Eds. Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 25–47.

[43] G. Kuhn and R. Guerney, "Sound pressure distribution about the human head and torso," *J. Acoust. Soc. Am.*, vol. 73, no. 1, pp. 95–105, 1983.

[44] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 1567–1576, Feb. 1977.

[45] M. Otani and S. Ise, "A fast calculation method of the head-related transfer functions for multiple source points based on the boundary element method," *Acoust. Sci. & Tech.*, vol. 24, no. 5, pp. 259–266, Oct. 2003.

[46] I. Jolliffe, *Principal Component Analysis.* New York, NY: Springer, 2002.

[47] W. Martens, "Principal components analysis and resynthesis of spectral cues to perceived direction," in *Proc. of the Int. Comp. Music Conf.*, 1987, pp. 274–281, san Francisco, CA.

[48] D. Kistler and F. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, Mar. 1992.

[49] J. Middlebrooks and D. Green, "Observations on a principal components analysis of head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 92, no. 1, pp. 597–599, July 1992.

[50] M. Blommer, "Pole-Zero Modeling and Principal Component Analysis of Head-Related Transfer Function," Ph.D. dissertation, Univ. of Michigan, Ann Arbor, 1996.

[51] Z. Wu, F. Chan, F. Lam, and J. Chan, "A time domain binaural model based on spatial feature extraction for the head-related transfer function," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2211–2218, Oct. 1997.

[52] R. Duraiswami and V. Raykar, "The manifolds of spatial hearing," in *Proc. IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, Mar. 2005.

[53] R. Duda, "Modeling Head Related Transfer Functions," in *Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, vol. 2, 1993, pp. 996–1000.

[54] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-Acoustical-Pole and Zero Modeling of Head-Related Transfer Functions," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 188–196, Mar. 1999.

[55] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing.* Englewood Cliffds, NJ: Prentice Hall, 1989.

[56] J. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Proc. 98th Audio Engr. Soc. Conv., preprint 3980*, Paris, France, 1995.

[57] A. Kulkarni, S. Isabelle, and H. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2821–2840, May 1999.

[58] F. Wightman and D. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, pp. 1648–1661, 1992.

[59] A. Kulkarni, S. Isabelle, and H. Colburn, "On the minimum-phase approximation of the head-related transfer functions," *Proc. IEEE Workshop on App. of Signal Processing to Audio and Acoust.*, Oct. 1995, new Paltz, NY.

[60] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design," *J. Audio Eng. Soc.*, vol. 47, no. 4, pp. 218–239, 1999.

[61] S.-P. Wu and W. Putnam, "Minimum perceptual spectral distance FIR filter design," in *Proc. IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, Mar. 1997.

[62] J. Torres, M. Petraglia, and R. Tenenbaum, "HRTF Modeling for Efficient Auralization," in *Proc. IEEE Int. Sym. on Industrial Electronics*, June 2003.

[63] M. Blommer and G. Wakefield, "Pole-Zero Approximations for Head-Related Transfer Functions Using a Logarithmic Error Criterion," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 278–287, May 1997.

[64] A. Kulkarni and H. Colburn, "Infinite-impulse-response models of the head-related trasnfer function," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1714–1728, April 2004.

[65] B. Beliczynski, I. Kale, and G. Cain, "Approximation of FIR by IIR Digital Filters: An Algorithm Based on Balanced Model Truncation," *IEEE Trans. Signal Processing*, vol. 40, no. 3, pp. 532–542, Mar. 1992.

[66] J. Mackenzie, J. Huopaniemi, V. Välimäki, and I. Kale, "Low-Order Modeling of Head-Related Transfer Functions Using Balanced Model Truncation," *IEEE Signal Processing Letters*, vol. 4, no. 2, pp. 39–41, Feb. 1997.

[67] F. Freeland, L. Biscainho, and P. Diniz, "Interpositional Transfer Function for 3D-Sound Generation," *J. Audio Eng. Soc.*, vol. 52, no. 9, pp. 915–930, Sept. 2004.

[68] D. Pralong and S. Carlile, "Generation and Validation of Virtual Auditory Space," in *Virtual Auditory Space: Generation and Applications*, S. Carlile, Ed. Berlin, Germany: Springer-Verlag, 1996.

[69] E. Langendijk and A. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 528–537, January 2000.

[70] S. Carlile, C. Jin, and V. Harvey, "The Generation and Validation of High Fidelity Virtual Auditory Space," in *Proc. 20th Annual Conf. IEEE Eng. in Medicine and Biology Soc.*, 1998.

[71] J. Sandvad and D. Hammershøi, "Binaural auralization. Comparison of FIR and IIR Filter representations of HIRs," in *Proc. 96th Audio Engr. Soc. Con.*, Amsterdam, preprint 3862, 1994.

[72] H. Hacihabiboglu, "A fixed-cost variable-length auralization filter model utilizing the precedence effect," in *Proc. IEEE Workshop of App. of Signal Processing to Audio and Acoust.*, 2003, New Paltz, NY.

[73] C. Cheng and G. Wakefield, "Moving Sound Source Synthesis for Binaural Electroacoustic Music Using Interpolated Head-Related Transfer Functions (HRTFs)," *Computer Music Journal*, vol. 25, no. 4, pp. 57–80, 2001.

[74] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *J. Acoust. Soc. Am.*, vol. 100, pp. 3785–3793, 1996.

[75] H. Mano, T. Nakamura, and W. Martens, "Perceptual Evaluation of an Earphone Correction Filter for Spatial Sound Reproduction," *J. Three Dimensional Images*, vol. 16, no. 4, pp. 48–55, 2002.

[76] K. McAnally and R. Martin, "Variability in the Headphone-to-Ear-Canal Transfer Function," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 263–266, April 2002.

[77] J. Russotti, T. Santoro, and G. Haskell, "Proposed Technique for Earphone Calibration," *J. Audio Eng. Soc.*, vol. 36, no. 9, pp. 643–649, Sept. 1988.

[78] H. Møller, C. Jensen, D. Hammershøi, and M. Sørensen, "Transfer Characteristics of Headphones Measured on Human Ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217, April 1995.

[79] W. Martens, "Individualized and generalized earphone correction filters for spatial sound reproduction," in *Proc. Int. Conf. on Auditory Display*, 2003, boston, MA.

[80] A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Am., letter to the editor*, vol. 107, no. 2, pp. 1071–1074, February 2000.

[81] D. Begault, *3D sound for virtual reality and multimedia.* Boston, MA: Academic Press, Inc., 1994.

[82] E. Wenzel and S. Foster, "Perceptual consequences of interpolating Head-Related Transfer Functions during spatial syntehsis," *Proc. IEEE Workshop on App. of Signal Processing to Audio and Acoust.*, Oct. 1993, new Paltz, NY.

[83] K. Hartung, J. Braasch, and S. Sterbing, "Comparison of Different Methods for the Interpolations of Head-Related Transfer Functions," in *Proc. 16th Audio Engr. Soc. Conf., paper number 16-028*, Mar. 1999.

[84] M. Matsumoto, M. Tohyama, and H. Yanagawa, "A method of interpolating binaural impulse responses for moving sound images," *Acoust. Sci. & Tech.*, vol. 24, no. 5, pp. 284–292, Oct. 2003.

[85] P. Runkle, M. Blommer, and G. Wakefield, "A comparison of Head Related Transfer Function interpolation methods," *Proc. IEEE Workshop on App. of Signal Processing to Audio and Acoust.*, Oct. 1995, new Paltz, NY.

[86] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, June 1997.

[87] D. Begault, "Challenges to the Successful Implementation of 3-D Sound," *J. Audio Eng. Soc.*, vol. 39, no. 11, pp. 864–870, Nov. 1991.

[88] B. Shinn-Cunningham and A. Kulkarni, "Recent Developments in Virtual Auditory Space," in *Virtual Auditory Space: Generation and Applications*, S. Carlile, Ed. Berlin, Germany: Springer-Verlag, 1996.

[89] K. Maki, S. Furukawa, and T. Hirahara, "Acoustical cues for localization by gerbils in an ecologically realistic environment," in *Assoc. for Research in Otolaryncology*, 2003, abstract 26, Poster 352.

[90] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *J. Exp. Psychol.*, vol. 27, pp. 339–368, 1940.

[91] F. Wightman and D. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2841–2853, May 1999.

[92] S. Perrett and W. Noble, "The effect of head rotations on vertical plane sound localization," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2325–2332, Oct. 1997.

[93] M. Kato, H. Uematsu, M. Kashino, and T. Hirahara, "The effect of head motion on the accuracy of sound localization," *Acoust. Sci. & Tech.*, vol. 24, no. 5, pp. 315–317, Oct. 2003.

[94] G. Reid and E. Milios, "Active stereo sound localization," *J. Acoust. Soc. Am.*, vol. 113, no. 1, pp. 185–193, 2003.

[95] D. Begault, "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems," *J. Audio Eng. Soc.*, vol. 40, no. 11, pp. 895–904, Nov. 1992.

[96] B. Shinn-Cunningham, "The perceptual consequences of creating a realistic, reverberant 3-d audio display," in *International Congress on Acoustics*, Kyoto, Japan, 2004.

[97] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization - An Overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, Nov. 1993.

[98] H. Kuttruff, *Room Acoustics.* London: Applied Science Publishers Ltd., 1979.

[99] L. Savioja and T. Lokki, "Digital Waveguide Mesh for Room Acoustic Modeling," in *Proc. ACM SIGGRAPH*, 2001, snowbird, UT.

[100] J. Smith and D. Rocchesso, "Aspects of Digital Waveguide Networks for Acoustic Modeling Applications," *Web Publications, Stanford University*, Dec. 1997, http://ccrma.stanford.edu/ jos/wgj/.

[101] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.

[102] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[103] A. Krokstad, S. Strøm, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. Sound Vib.*, vol. 8, no. 1, pp. 118–125, 1968.

[104] B.-I. Dalenbäck, "Room acoustic prediction based on a unified treatment of diffuse and specular reflectance," *J. Acoust. Soc. Am.*, vol. 100, Aug. 1996.

[105] W. Gardner, "Reverberation Algorithms," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer Academic Publishers, 1998.

[106] R. Heinz, "Binaural Room Simulation Based on an Image Source Model with Addition of Statistical Methods to Include the Diffuse Sound Scattering of Walls and to Predict the Reverberant Tail," *Applied Acoustics*, vol. 38, pp. 145–159, 1993.

[107] R. Duraiswami, N. Gumerov, D. Zotkin, and L. Davis, "Efficient evaluation of reverberant sound fields," *Proc. IEEE Workshop on App. of Signal Processing to Audio and Acoust.*, Oct. 2001, new Paltz, NY.

[108] A. Sarti and S. Tubaro, "Low-cost geometry-based acoustic rendering," in *Proc. Conf. on Digital Audio Effects (DAFX)*, Dec. 2001.

[109] U. Kristiansen, A. Krokstad, and T. Follestad, "Extending the Image Method to Higher-Order Reflections," *Applied Acoustics*, vol. 38, pp. 195–206, 1993.

[110] L. Beranek, "Concert hall acoustics - 1992," *J. Acoust. Soc. Am.*, vol. 92, no. 1, pp. 1–38, July 1992.

[111] D. Begault, "Audible and inaudible early reflections: thresholds for auralization system design," *Proc.* 100*th* *Conv. of Audio Eng. Soc. - Preprint 4244*, 1996.

[112] S. Bech, "Perception of reproduced sound: Audibility of Individual reflections in a complete sound field, II," *Proc.* 99*th* *Conv. of Audio Eng. Soc. - Preprint 4093*, 1995.

[113] M. Ebata, T. Sone, and T. Nimura, "On the Perception of Direction of Echo," *J. Acoust. Soc. Am.*, vol. 44, no. 2, pp. 542–547, 1968.

[114] B. Rakerd and W. Hartmann, "Localization of sound in rooms II: The effect of a single reflecting surface," *J. Acoust. Soc. Am.*, vol. 78, no. 2, pp. 524–533, Aug. 1985.

[115] P. Zurek, "Measurements of binaural echo suppression," *J. Acoust. Soc. Am.*, vol. 66, no. 6, pp. 1750–1757, Dec. 1979.

[116] S. Kung, "A new identification and model reduction algorithm via singular value decompositions," in *Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, 1978, pp. 705–714.

[117] K. Glover, "All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds," *Int'l. J. Control*, vol. 39, pp. 1115–1193, 1984.

[118] J. Bay, *Fundamentals of linear state space systems.* Boston, MA: McGraw-Hill, 1999.

[119] A. Antoulas, "Approximation of linear dynamical systems," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, J. Webster, Ed. New York: John Wiley and Sons, Inc., 1999, vol. 11, pp. 403–422.

[120] T. Hartley, R. Veillette, J. Abreu-Garcia, A. Chicatelli, and R. Hartmann, "To Err is Normable: The Computation of Frequency-Domain Error Bounds From Time-Domain Data," *NASA Contractor Report NASA/CR-1998-208516*, 1998, available online: http://gltrs.grc.nasa.gov/reports/1998/CR-1998-208516.pdf.

[121] K. Glover, R. Curtain, and J. Partington, "Realization and approximation of linear finite-dimensional systems with error bounds," *SIAM J. Control and Optimization*, vol. 26, no. 4, pp. 863–898, July 1988.

[122] V. Adamjan, D. Arov, and M. Krein, "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem," *Math. USSR Sbornik*, vol. 15, pp. 31–73, 1971.

[123] B. Chen, S. Peng, and B. Chiou, "Iir filter design via optimal Hankel-norm approximation," *IEE Proc. G., Circuits, Devices and Systems*, vol. 139, no. 5, pp. 586–590, Oct. 1992.

[124] I. Kale, J. Gryka, G. Cain, and B. Beliczynski, "FIR filter order reduction: balanced model truncation and Hankel-norm optimal approximation," *IEE Proc.-Vis. Image Signal Process.*, vol. 141, no. 3, pp. 168–174, June 1994.

[125] B. Beliczynski, J. Gryka, and I. Kale, "Critical comparison of hankel-norm optimal approximation and balanced model truncation algorithms as vehicles for FIR-to-IIR filter order reduction," in *Proc. IEEE Conf. on Acoust., Speech, and Signal Processing (ICASSP)*, 1994.

[126] B. Moore, "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," *IEEE Trans. Automatic Control*, vol. AC-26, no. 1, pp. 17–32, Feb. 1981.

[127] A. Antoulas, D. Sorensen, and S. Gugergin, "A survey of model reduction methods for large-scale systems," *Contemporary Mathematics*, vol. 280, pp. 193–219, 2001.

[128] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control.* Chichester, UK: John Wiley and Sons, 1996.

[129] H. Gao, J. Lam, C. Wang, and S. Xu, "H^∞ model reduction for discrete time-delay systems: delay-independent and delay-dependent approaches," *Int. J. Control*, vol. 77, no. 4, pp. 321–335, Mar. 2004.

[130] E. Macpherson and J. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, p. 2219, 2002.

[131] E. Macpherson and A. Sabin, "Binaural weighting of monaural spectral cues for sound localization," *J. Acoust. Soc. Am.*, vol. 121, p. 3677, 2007.

[132] C. Avendano, R. Duda, and V. Algazi, "Modeling the Contralateral HRTF," in *Proc. AES 16th International Conference on Spatial Sound Reproduction*, 1999, pp. 313–318, Rovaniemi, Finland.

[133] F. Wightman and D. Kistler, "Sound localization with unilaterally degraded spectral cues (A)," *J. Acoust. Soc. Am.*, vol. 105, no. 2, p. 1162, Feb. 1999.

[134] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge, UK: Cambridge University Press, 1990.

[135] G. Golub and C. V. Loan, *Matrix Computations.* Baltimore, MD: John Hopkins University Press, 1996.

[136] E. Wenzel, "The effect of increasing system latency on localization of virtual sounds with short and long duration," in *Proc. Int. Conf. on Auditory Display*, July 2001.

[137] G. Robel, "On computing the infinite norm," *IEEE Trans. on Auto. Control*, vol. 34, no. 8, pp. 882–884, August 1989.

[138] W. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3678–3688, June 1996.

[139] P. Bloom, "Creating Source Elevation Illusions by Spectral Manipulation," *J. Audio Eng. Soc.*, vol. 25, no. 9, pp. 560–565, Sept. 1977.

[140] C. Lim and R. Duda, "Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model," in *Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, vol. 1, 1994, pp. 399–403.

[141] U. Al-Saggaf and G. Franklin, "Model Reduction Via Balanced Realizations: An Extension and Frequency Weighting Techniques," *IEEE Trans. Automatic Control*, vol. 33, no. 7, pp. 687–692, 1988.

[142] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *Proc. IEEE Workshop on App. of Signal Processing to Audio and Acoust.*, 2001.

[143] R. Hartung and A. Raab, "Efficient modeling of head-related transfer functions," *Acta Acoustica*, vol. 82, no. S88, 1996.

[144] J. Moorer, "48-Bit Integer Processing Beats 32-Bit Floating-Point for Professional Audio Applications," in *Proc. 107th AES Convention*, Sept. 1999, preprint 5038.

[145] D. Green, "Psychoacoustics and Detection Theory," in *Signal Detection and Recognition by Human Observers*, J. Swets, Ed. New York: John Wiley and Sons, Inc., 1964.

[146] H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, no. 2, pp. 467–537, 1971.

[147] M. Leek, "Adaptive procedures in psychophysical research," *Perception & Psychophysics*, vol. 63, no. 8, pp. 1279–1292, 2001.

[148] W. Drennan and C. Watson, "Sources of variation in profile analysis. I. Individual differences and extended training," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2491–2497, Nov. 2001.

[149] Bang and Olufsen, "Music for Archimedes," 1992, B&O 101.

[150] CBS, "New CBS audio-file sound effects," 1977, A2 14062/A 14064.

[151] A. Kulkarni and H. Colburn, "Role of spectral detail in sound-source localisation," *Nature*, vol. 396, pp. 747–749, Dec. 1998.

[152] D. Green and G. Kidd, "Further studies of auditory profile analysis," *J. Acoust. Soc. Am.*, vol. 73, no. 4, pp. 1260–1265, April 1983.

[153] Beyer-Dynamic, "Headzone PRO," 2006, http://www.beyerdynamic.com.

[154] C. K. Chui and G. Chen, *Discrete H^∞ Optimization.* Berlin: Springer-Verlag, 1997.

[155] D. Luenberger, "Canonical Forms for Linear Multivariable Systems," *IEEE Trans. Auto. Control*, pp. 290–293, June 1967.

[156] J. Huang and G. Gu, "A Direct Approach to the Design of QMF Banks via Frequency Domain Optimization," *IEEE Trans. on Signal Processing*, vol. 46, no. 8, pp. 2131–2138, August 1998.

[157] R. Bartels and G. Stewart, "Solution of the Matrix Equation AX + XB = C," *Comm. of the ACM*, vol. 15, no. 9, pp. 820–826, Sept. 1972.

[158] L. Fortuna, G. Nunnari, and A. Gallo, *Model Order Reduction Techniques with Applications in Electrical Engineering.* London: Springer-Verlag, 1992.