State-Space Architectures for Binaural Environment Modeling

Norman H Adams

March 16, 2007

Abstract

Binaural displays for immersive listening must model reverberant environments, multiple sound sources, and accommodate source and head motion. Many displays model such phenomena as a collection of spatially distributed point sources; source signals are filtered with multiple head-related transfer functions (HRTFs) and then combined into a single binaural signal for display. The computational load of such systems scales linearly with the number of HRTFs modeled by the display. 'Realistic' auditory scenes often require a large number of HRTFs, hence this framework is computationally untenable. We propose a method that significantly eases this load by formulating the HRTF filter array as a state-space system. Three state-space architectures are explored; MIMO, SIMO and MISO. The relative merits of each are found to depend on the specific application. If only a single source signal is displayed, a SIMO architecture yields the lowest computational cost for a fixed approximation quality. However, if multiple signals are displayed a pair of MISO systems yield the lowest cost. For the MIMO architecture, the interaural time delay degrades performance substantially and a hybrid technique is described to mitigate this problem. Hankel methods are found to be a good choice for model reduction, and yield displays with superior approximation quality relative to conventional FIR filter arrays of equal computational complexity.

1 Introduction

Binaural displays seek to immerse a listener in a 3D virtual auditory scene (VAS) using only a pair of conventional headphones [1]. The synthesis of perceptually convincing VAS requires the modeling of complex physical phenomena. Typical physical acoustic scenes contain multiple sound sources, sources both in the far- and near-fields, and interactions with the environment such as reflections and diffusion, as well as source and listener motion [2]. Methods have been proposed for modeling such physical phenomena in binaural displays, but always with a formidable computational burden. Numerous emerging applications have shown interest in binaural display technology, from video games and virtual reality to sonification and musical composition. However, the computational efficiency of flexible and perceptually satisfying binaural displays must be improved in order for the technology be broadly accepted.

The present work considers a flexible framework for modeling a wide variety of auditory scenes, and proposes an efficient implementation for this framework. The framework filters source signals with numerous *head-related transfer functions* (HRTFs) simultaneously. A naive implementation of this framework is untenable, as the computational cost scales linearly with the number of HRTFs included, and the number of HRTFs required for a 'realistic' auditory scene is often large. To address this problem, we explore reduced-order *multiple-input multiple-output* (MIMO) state-space systems for the HRTF filter array, and find that state-space systems offer substantial computational improvements relative to the conventional FIR filter arrays. Previous research has considered state-space approaches for low-order filter design, however we are unaware of any studies that compare the net computational cost of a state-space implementation to that of a conventional filter array. We show that a large array of HRTFs can be efficiently approximated using a single low-order state-space system.

The remainder of this section gives background on binaural displays, as well as reports on previous studies that employ state-space methods for binaural displays. Section 2 formulates the HRTF filter array in the state-space, and describes three state-space architectures. Multiple architectures are considered because the performance of the simplest architecture suffers due to the *interaural time delay* (ITD) between the ipsilateral and contralateral HRTFs. For this reason, two alternative architectures are considered that circumvent the ITD problem. Furthermore, a hybrid method for the first

architecture is proposed to mediate this problem. A classical Hankel-optimal technique is used to reduce the order of the state-space systems. System performance is then characterized using an 'auditory' $\mathcal{L}^2$ error metric in Section 3.

1.1 Virtual Auditory Scene Models

An auditory scene that consists of a single stationary sound source in the far-field of a listener in an anechoic environment can be modeled using a single pair of *head related transfer functions* (HRTFs) [3]. The HRTF represents the acoustic filtering of a plane wave enroute to a listener's two ears due to the head, pinna and torso of the listener, and hence is unique to the listener. Such transfer functions can be implemented using appropriately measured *head related impulse responses* (HRIRs), which, empirically, require approximately 200 FIR filter coefficients at a sampling rate of 44.1 kHz. Even a single HRTF pair presents a substantial computational load, hence considerable effort has been made to find low-order approximations to measured HRIRs [4].

However, it is well known that binaural displays, as described above, are perceptually unsatisfying [5]. Virtual sound sources are often not externalized, with the perceived sound object located inside the listener's head. Virtual auditory scenes presented over headphones often lack *presence*. Localization errors, and in particular front-back errors, are common [6]. These deficiencies are often attributed in the literature to the assumptions that underly conventional binaural displays; that sound sources are stationary, located in the far-field, and in free space. Given that we rarely experience such primitive auditory scenes, it is not surprising that virtual auditory scenes based on these assumptions are perceptually inadequate for an immersive experience.

Recently, binaural displays have been designed that account for reflective environments [7–12], source and listener motion [7–10, 13–15], and spatially-extended sources [16]. All of these examples share a similar framework: monaural source signals are filtered with multiple HRTF pairs instead of a single HRTF pair. That is, a monaural source is *auralized* at D directions simultaneously, and then the D binaural signals are combined so as to model the desired auditory scene. This framework provides intuitive models for many auditory scenes: reflections can be modeled using image sources [17] or ray-tracing [18], source or listener motion can be included using dynamic amplitude panning [19], and spatially-extended sources can be decomposed

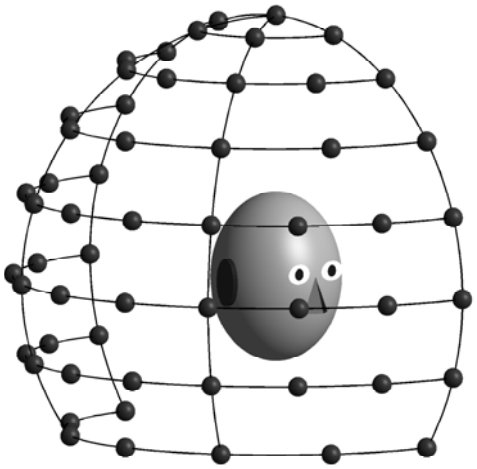

Figure 1: Virtual source locations in the median plane and to the right of a listener.

into a collection of point sources. An example arrangement of $D = 97$ directions is shown in Figure 1 with small spheres representing each direction[1]. For clarity, only the directions in the median plane and to the right of the listener are shown.

This framework compounds the computational burden however, as the overall computational cost scales linearly with D if individual filters are employed for each HRTF. A typical architecture for such binaural displays is shown in Figure 2. Acoustic reflections, in particular, are problematic, as enclosures often result in a large number of acoustic reflections from different directions impinging on the listener before the sound field becomes *diffuse* [20, 21]. Hence D must be large. It is unclear how accurately the *early reflections* must be modeled however [22, 23]. The auditory system gives *precedence* to the first (i.e. direct) sound wave in perceiving the location of a sound source [24]. Some studies, for example [11], propose filter arrays that model reflected waves less accurately than the direct wave. However, incorporating source or listener motion into such a display would be difficult, as the arrangement of directions that are accurately modeled versus those that are not must be dynamically updated. Furthermore, localization

[1]The collection of directions shown in Figure 1 is a subset of the directions measured for the HRTF datasets used in the present study, see Section 3.1.

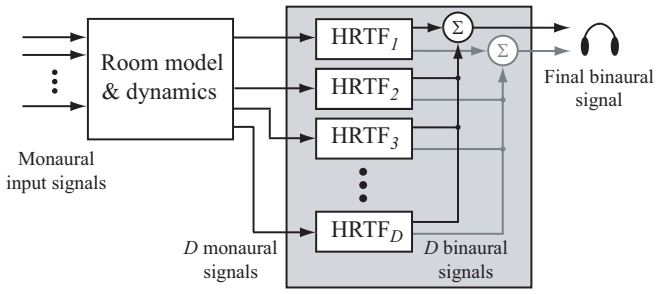

Figure 2: A naive implementation of a multi-direction binaural display.

performance alone may not be a suitable criteria for designing a binaural display, as a listener may be able to correctly 'localize' a sound source without being perceptually immersed in the auditoryscene[2]. The present work describes the design of state-space systems that model D pairs of HRTFs with uniform weight. This approach yields a flexible binaural display, as both acoustic reflections and motion can be implemented with simple scale-and-delay filter banks [7–9]. Nonetheless, it is possible to use non-uniform weights so as to more accurately model some directions.

1.2 State-Space HRTF Models

HRTFs measured for different directions, and for different listeners, are locally redundant. Even if the *common transfer function* (CTF) is removed from a collection of HRTFs, leaving only the *directional transfer functions* (DTFs), the transfer functions are still similar in shape, especially for nearby directions [3]. Numerous studies have found that collections of HRTFs can be reasonably represented in low dimensional spaces. In [25] it is shown that most of the HRTF variance can be accounted for with the first five principal components of a measured dataset. However, such representations do not necessarily yield low cost filters for individual HRTFs. A system that models HRTFs at many directions simultaneously may be able to utilize the redundancy of HRTF datasets to reduce the net cost of the system. Indeed, it has been shown that HRTFs can be accurately approximated using IIR

[2]For example, studies report 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, or on, the head [1, 7].

filters with common poles [26]. 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 state-space system share the same denominator polynomial [27].

Two recent studies propose state-space systems that model HRTFs at multiple directions simultaneously. In [28] MISO systems are designed that model multiple HRTFs for each ear. HRTF redundancy is not fully exploited in this work however, as separate state-space systems are designed for each HRTF individually, and then combined into one large system. In contrast, [29] considers a MIMO state-space design that directly models multiple HRTFs in the horizontal plane. Both studies employed BMT to design low-order systems. It was shown that for sufficiently large system order, the localization performance of a listener using a state-space system was similar to the performance if a high-order FIR array was used [29]. However, neither study considered in detail the computational advantages of state-space implementations. The specific aim of the present work is to demonstrate that a substantial computational savings can be achieved for binaural environment modeling using reduced-order state-space systems.

2 Methods

Consider a stable, causal, discrete-time MIMO state-space system[3]

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

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 $\mathbf{\Sigma} = \big(\mathbf{A}, \mathbf{B}, \mathbf{C}\big)$

[3]For convenience, the systems considered here have no feed-through term (the $\mathbf{D}\mathbf{u}[n]$ term), similar to [28, 29]. The Hankel operator is not influenced by the $\mathbf{D}$ matrix, hence the choice of $\mathbf{D}$ is somewhat arbitrary for Hankel-optimal model reduction. In the present work we simply set $\mathbf{D} = \mathbf{0}$. The interested reader is referred to [30] for a detailed discussion of this term.

represent the state-space system. The matrix impulse response of $\mathbf{\Sigma}$ is

$$\begin{aligned} \mathbf{h}[n] &= \begin{bmatrix} h_{11}[n] & \dots & h_{1M}[n] \\ \vdots & \ddots & \vdots \\ h_{P1}[n] & \dots & h_{PM}[n] \end{bmatrix} \\ &= \begin{cases} \mathbf{C}\mathbf{A}^{n-1}\mathbf{B} & n > 0 \\ \mathbf{0} & n \leq 0 \end{cases} \end{aligned} \tag{2}$$

For the binaural display application we seek a state-space system $\mathbf{\Sigma}$ that models a convenient arrangement of the HRTFs. Three possible arrangements are described below. After formulating the HRTF filter array as a state-space system, Hankel-norm Optimal Approximation (HOA) is applied to the system to reduce the order to $N < N_0$, yielding a low-order system $\widehat{\mathbf{\Sigma}} = \left(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}\right)$. The HOA method of model reduction was first published by Glover [30], and has since found use in numerous model reduction applications. In [31], HOA and BMT are described and compared in detail for HRTF modeling.

2.1 State-Space Architectures

Let $h_d^L[n]$ and $h_d^R[n]$ be the HRIRs for the left and right ears for direction d. For a binaural display that filters a source signal at D directions simultaneously, we consider three state-space architectures that implement the $2D$ transfer functions. The three architectures are shown in Figure 3. For each architecture we focus on the design of the state-space system, shown in gray. Two of the architectures employ a single state-space system, whereas the third employs two state-space systems. All three architectures can readily accommodate acoustic reflections and motion by placing a scale-and-delay filter array either before or after the state-space system. There are significant differences between the three architectures however, and it is unclear a priori which architecture yields the greatest computational savings.

Perhaps the most obvious architecture is a MIMO system with D inputs and 2 outputs, analogous to the HRTF filter array in Figure 2. The matrix impulse response for this system is

$$\mathbf{h}[n] = \begin{bmatrix} h_1^L[n] & h_2^L[n] & \dots & h_D^L[n] \\ h_1^R[n] & h_2^R[n] & \dots & h_D^R[n] \end{bmatrix} \tag{3}$$

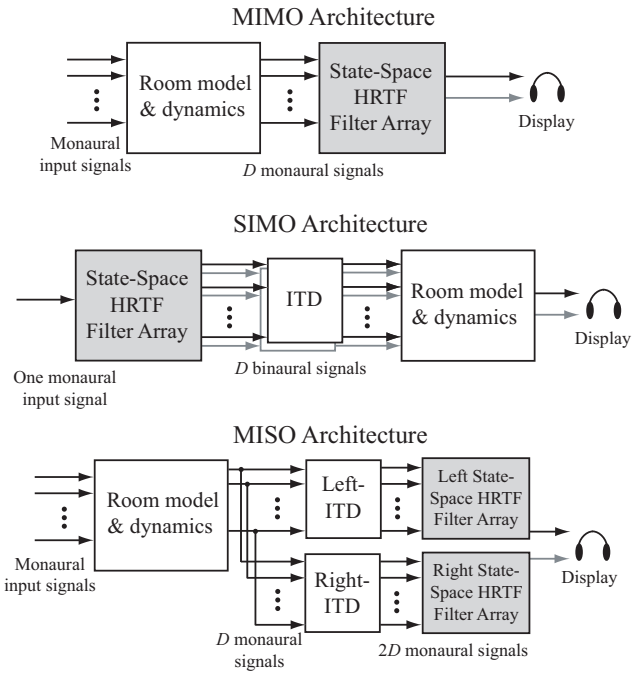

Figure 3: Three architectures for a binaural display.

Clearly, the interaural time delay (ITD) and interaural level difference (ILD) must be included in the transfer functions that the MIMO state-space system models. For the model reduction methods considered here, the ITD is problematic, whereas the ILD is not. The ILD effectively weights the transfer functions such that, after model reduction, the ipsilateral HRTFs are more accurately approximated than the contralateral HRTFs. However, it may be desirable to model the ipsilateral HRTFs more precisely than the contralateral HRTFs [32]. Furthermore, while the contralateral HRTFs are approximated less accurately, the desired ILD is retained by the model reduction method procedure. In contrast, the ITD is not retained by the model reduction procedure in all cases. Often the phase responses of the contralateral HRTFs are smeared such that the ITD is lost. This issue is described in detail below, but the problem can be circumvented by using an architecture that allows the ITD to be modeled externally.

The two alternative architectures employ state space systems with either one input or one output. In this case the ITD can be modeled outside the

state-space systems[4]. The HRTFs can be modeled either using a SIMO system with one input, $2D$ outputs, and matrix impulse response

$$\mathbf{h}[n] = \left[\begin{array}{cccccc} h_1^L[n] & h_1^R[n] & h_2^L[n] & h_2^R[n] & \ldots & h_D^R[n] \end{array} \right]^T \tag{4}$$

or as two MISO systems, one for each ear, with D inputs each. In this case the matrix impulse responses are

$$\begin{array}{rcl} \mathbf{h}_L[n] & = & \left[\begin{array}{cccc} h_1^L[n] & h_2^L[n] & \ldots & h_D^L[n] \end{array} \right] \\ \mathbf{h}_R[n] & = & \left[\begin{array}{cccc} h_1^R[n] & h_2^R[n] & \ldots & h_D^R[n] \end{array} \right] \end{array} \tag{5}$$

Clearly, the SIMO architecture has the disadvantage that multiple source signals at different locations cannot be presented simultaneously. 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.

For all three architectures it is straightforward to design a state-space system that implements the collection of $2D$ HRIRs exactly. Such a state-space system is high order and computationally prohibitive. As such, we employ HOA to reduce to computational load of the state-space system. However, the modeling ITD for the MIMO architecture is problematic. This issue is discussed below, and a hybrid method is proposed to mitigate it.

2.2 ITD and Singular Values

The precise phase response of a system is often found to be perceptually unimportant for many audio applications. For binaural displays, listeners appear to be insensitive to many types of distortion of the phase response of HRTFs, so long as the magnitude of the distortion is not too large [34]. Measured HRTFs are nearly minimum-phase, with the addition of a linear-phase term in the contralateral HRTFs. While the frequency-dependent part of the phase response is relatively unimportant, the linear-phase term is perceptually critical for spatial hearing. Indeed, human listener's are sensitive to perturbations in the ITD of a virtual source as small as several microseconds [1]. This perceptual sensitivity presents a difficulty in the design of low-order MIMO state-space systems.

[4]The ILD is still modeled inside the state-space systems.

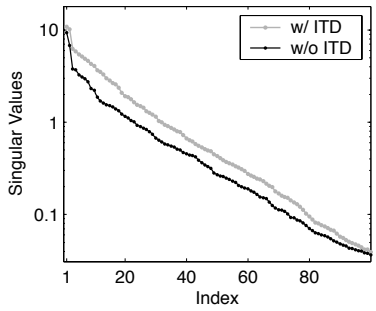

Figure 4: The first 100 Hankel singular values for two full-order SIMO systems. The two systems model the same 19 contralateral HRTFs. One system includes the ITD (gray line) and other does not (black line). Both linear and log scales are shown.

Hankel methods are known to well approximate transfer functions that are minimum phase, or nearly so [33]. However, including linear phase terms in the transfer functions moves many of the zeros of the filter outside the unit circle. Canceling these zeros with the state-space poles during model reduction distorts the phase response of the approximant. For MIMO systems, the delay terms of the contralateral HRTFs, which are unique for each direction, places zeros at unique locations outside the unit circle. In order to accurately retain the time-delays in the state-space approximant, it is necessary that groups of (common) poles be included for each unique linear-phase term. There are few analytic tools for understanding the time-domain distortion if fewer poles are retained. In leu of a simple analytic theory an example makes clear the practical influence of ITD on state-space model 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. 4. The first 100 singular values of the system with ITD are larger than those of the system without ITD, hence a low-order system with $N < 100$ will have larger Hankel error, and likely larger $\mathcal{L}^\infty$ error, if the ITD is modeled [31]. This performance degradation does not affect all 19 transfer functions equally, however.

Fig. 5 shows the block impulse response of eight state-space systems constructed from the 19 HRIRs. Time indexes the horizontal axis, and azimuth indexes the vertical axis[5]. Because the systems have SIMO architecture, the impulse responses, oriented in this way, are equivalent to the top 19 rows of the Hankel matrix $\mathcal{H}$ of each system [35][6]. The leftmost column of Fig. 5 shows the ideal matrix impulse responses for the systems with and without ITD. The remaining three columns show the resulting 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, retaining only the largest singular value ($N = 1$) is sufficient to retain most of the energy of all impulse responses. In contrast, for the system with ITD, only the impulse responses at 0° to 180° retain most of their energy. For all other azimuths, the impulse responses are virtually zero. If the six largest singular values are retained, the system without ITD is reasonably well approximated for all azimuths. For the system with ITD however, the impulse responses far from the median are still virtually zero, and even the responses close to the median exhibit much phase distortion. Of course, as N increases, both systems are well approximated at all azimuths, as shown in the $N = 30$ case.

Modeling HRTFs with ITD using low-order state-space systems also presents a perceptual problem: the contralateral impulse responses are often smeared, such that there is no longer a clear ITD. A typical example is shown in Fig. 6. Measured left and right HRIRs for a position on the left, behind the listener, are shown by thin black lines. An order $N = 40$ MIMO system is designed from measured HRIRs for $D = 68$ directions, including the direction shown in Fig. 6, using the HOA method. The resulting impulse responses for the same direction are shown by thick gray lines. The left ear response is accurately approximated, whereas the right ear response is not. The time delay of the right ear response is partially filled-in, an unacceptable distortion for binaural display applications. The extent of this phase distortion varies with N, but is usually negligible for $N > 70$ even for large D. However, we seek MIMO state-space systems with lower order. In this case the phase distortion is problematic for some directions: impulse responses with small time delay, less than 400 ms, experience little phase distortion, but responses with

[5]A grayscale colormap is used in which zero is mapped to gray, positive values to black and negative values to white. The colormap is warped somewhat to make the black-to-white fluctuations more apparent, although the warping is modest enough that the interaural level difference (ILD) is still visible for azimuths far from the median plane

[6]The impulse responses shown in Fig. 5 have been shifted to show $t = 0$ more clearly.

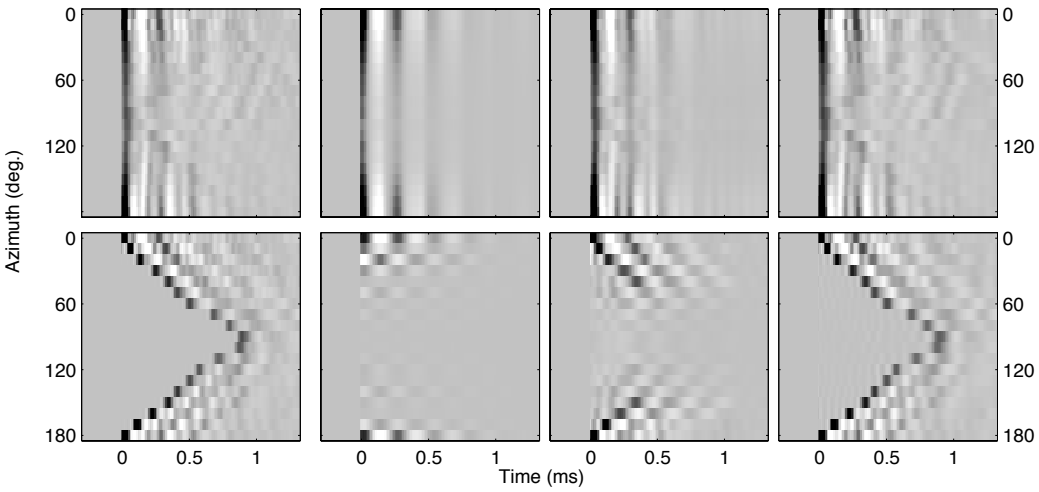

Figure 5: Matrix impulse responses for eight SIMO systems; four without ITD (top row) and four with ITD (bottom row). This is equivalent to the first 19 rows of the Hankel matrix of each system. The leftmost column shows the ideal matrices for this system (constructed from 19 contralateral HRIRs). The remaining three columns show response matrices for $N = (1, 6, 30)$ approximants.

large time delay experience 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.

This observation would seem to agree with results presented in [29]. In this study, MIMO state-space system that model HRTFs with ITD were evaluated by human listeners. Localization errors were found to be small for systems with order $N > 80$. For systems with lower order, localization performance was poor for locations far from the median plane. It is possible that the poor localization performance is a result of this phase distortion.

Classical state-space theory provides few tools for addressing diversity in the phase response of transfer function matrices. Recently, time-delay systems have been the subject of several studies [36]. Unfortunately, attempts to model the present HRTF filter array as a single time-delay system have been unsuccessful, as the HRTF filter array is a multi-time-delay system.

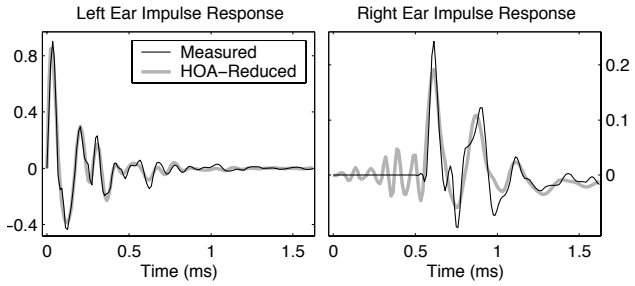

Figure 6: Measured HRIRs and HOA-reduced impulse responses for azimuth $\theta = -120^\circ$ and elevation $\phi = 0^\circ$.

We have explored several ad hoc state-space systems to remedy the ITD problem for the MIMO architecture, however only a simple hybrid system was found to both retain the necessary phase responses and be computationally efficient. The hybrid system consists of a MIMO state-space system, as described above, with individual FIR filters connected to the input-output pairs that correspond to HRIRs 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. 6, 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. In this way, a low-order hybrid MIMO system is built that retains the necessary phase responses.

2.3 Computational Cost

In the next section, state-space systems are compared to FIR filter arrays of *equal computational cost*. A measure of computational cost that is consistent for both system architectures must be defined. We define the cost C as the number of multiplications operations required per sample period, or equiva-

lently, 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 [4, 16, 37]. 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 models $D = 10$ HRIRs 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 an infinite number of 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 ($\mathbf{A}$,$\mathbf{B}$,$\mathbf{C}$) such that the $\mathbf{A}$ matrix becomes more sparse. We employ a *Schur decomposition* to triangularize the $\mathbf{A}$ matrix [38]. Because we seek system matrices that are strictly real, the new $\mathbf{A}$ matrix is only quasitriangular. In this case the cost of the final state-space system is not greater than $C = N^2/2 + (P+M+1)N$.

This definition of computational cost, C, neglects memory requirements. A MIMO FIR filter array of order N with M inputs and P outputs requires MN memory cells[7], whereas a state-space system of order N requires only N memory cells. There is no universal standard for weighting computational cost, in terms of arithmetic operations, and memory requirements however. In the experiments presented below, systems are compared that have equal computational cost, and the memory requirements of each system are not considered.

3 Performance Characterization

To characterize the performance of the state-space systems described above, a numerical experiment is conducted in which multiple HRTF systems of varying size D, but fixed cost C, are constructed. The number of directions D that are required for a binaural display depends on the application. For simple virtual auditory scenes, a small number of directions may suffice.

[7]In this case the number of 'memory cells' is the number of 'numbers' that must be retained between time steps. The number of *bits* required depends on the data type used by the system, for example 16-bit integers versus 64-bit floats. Data types for state-space systems are discussed in Section 4.

However, an 'ecologically realistic' virtual auditory scene would likely require a large number of directions surrounding the listener. As such, we view D as an independent variable and choose the largest system order $N \in \mathbb{Z}_+$ such that the total computational cost of the system, as defined in Section 2.3, does exceed some bound $C_{\max}$.

An FIR filter array is used as a baseline for comparison with the state-space systems. For any given cost bound, two FIR arrays are constructed, one with cost bound by $C_{\max}$, and the other with cost bound by $2C_{\max}$. We include the 'double-cost' FIR array to gauge the relative 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. This would seem to be a large enough margin of improvement to warrant the use of state-space systems in practical binaural displays. FIR filters of order N are constructed by truncating all but the first $N+1$ samples of minimum-phase HRIRs. Hence the FIR filters are optimal FIR approximations in terms of $\mathcal{L}^2$ error [37].

A warped $\mathcal{L}^2$ error is reported as the primary performance metric. This metric has been previously used in HRTF approximation studies. Hankel and $\mathcal{L}^\infty$ errors have been reported elsewhere for the present study [31, 39]. For MIMO state-space systems, a measure of the ITD distortion is also reported. In addition to average error metrics, the frequency responses of several example systems are shown.

3.1 HRTF Measurements

The HRTFs used to design the low-order systems were measured at the Naval Submarine Medical Research Laboratory in Groton, CT. The HRTFs of eight individuals were measured. Using a vertical-polar coordinate system, HRTFs were measured in 10° increments in azimuth around the listener, and in 18° increments in elevation from $-36°$ to $+90°$, yielding a total of 253 pairs of HRTFs for each listener. For the experiment below, systems are designed that model D of the 253 measured directions. For every system, D directions are chosen randomly subject to a constraint that the D directions be approximately uniformly distributed around the listener.

Golay codes are employed to minimize bias in the measurement and identification process. The HRTF measurement process is described in detail in [40]. At a sampling rate of 44.1 kHz the measured HRIRs have length

256, order $N = 255$. The measured HRIRs are nearly minimum-phase. To simplify the analysis, and to guarantee the performance of the truncated FIR filters, all HRIRs are transformed so as to be strictly minimum-phase for this experiment. We also applied the state-space methods to the nearly minimum-phase HRIRs and observed that performance did not degrade, however in this case the performance of truncated FIR filters cannot be viewed as optimal in any sense.

3.2 An Example

Before we present the results of the main experiment, we consider an example. Consider a SIMO system that models $D = 44$ HRTF pairs. This system has $M = 1$ input and $P = 88$ outputs. One direction included in this system is to the right, behind the listener, at azimuth $\theta = 120°$ and elevation $\phi = 0°$. The HRTF magnitude responses for this direction are shown with a thin black line in Figure 7. 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. Both systems are designed to not exceed a cost bound of $C_{\max} = 3000$. This bound is approximately equivalent to the cost of six full-order HRIR pairs. 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 response of these two systems at the aforementioned direction is also shown in Figure 7.

Overall, the state-space response appears to more accurately approximate the measured HRTFs than the FIR response. At low-frequencies, the response of the FIR filters diverge from the desired response. This is a trend seen throughout the results. For the example shown in figure 7, the FIR approximation is especially poor below 300 Hz. It is unclear however that this low-frequency degradation is relevant for binaural display applications, as binaural phase differences appear to be perceptually dominant for localization in this frequency domain [1].

From the right column of Fig. 7 it can be seen that the spectral notches in 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 modeled by the state-space system, but is shifted to a slightly higher frequency by the FIR array. A more significant difference is seen in the right ear response at 8.5 kHz, where the measured HRTF exhibits a sharp, lopsided notch. The state-space system also exhibits a lopsided

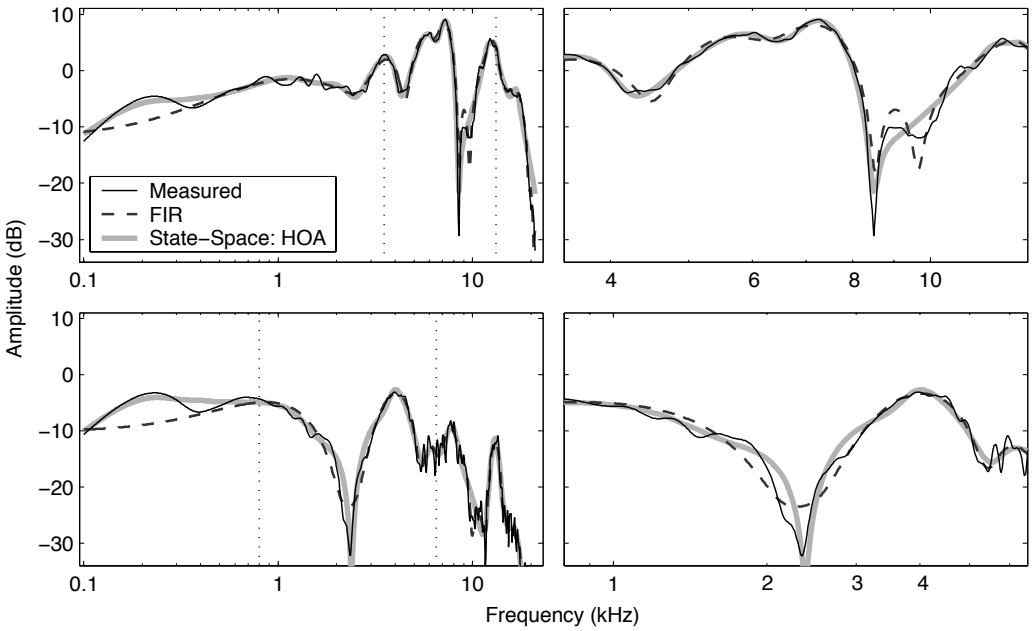

Figure 7: Magnitude responses for direction $\theta = 120^\circ$ and $\phi = 0^\circ$, for the right ear (top) and left ear (bottom). The vertical dotted lines in the left column indicate the frequency bounds of the right column.

notch at the same frequency, whereas the FIR array exhibits two notches with the same depth (-18dB), 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 systems, whereas the FIR array exhibits only a shallow dip at this frequency. This is another trend seen throughout the results: spectral notches are more accurately modeled by the state-space systems than the FIR arrays, especially for notches below 5 kHz.

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

For the main experiment, we select a cost bound of $C_{\max} = 4000$, which is approximately the cost of eight full-order HRIR pairs. We then design state-space systems that meet this bound for a varying number of directions $1 \leq D \leq 110$. Three system architectures are considered: MIMO, SIMO and MISO, as described in Section 2.1. Recall that the ITD is modeled by the

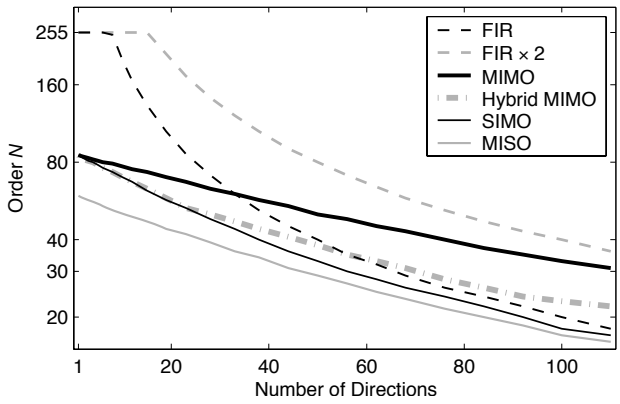

Figure 8: System order for six systems as a function of the number of directions D. The system labelled 'FIR $\times 2$' has a cost bound of 8000, and the remaining five have a cost bound of 4000.

MIMO system, but not by the SIMO or MISO systems. Two FIR arrays are designed for each configuration. For each configuration, separate systems are designed for each of the eight HRTF datasets described above. The results are then averaged across the eight individuals. We found that the approximation error varied little across individual, so we report only mean values without standard deviation bars.

Figure 8 shows the order N for six systems: MIMO, SIMO and MISO state-space systems with cost $C \leq C_{\max}$, a hybrid MIMO system 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. For the hybrid MIMO architecture, N is the order of the state-space component. Note that for the FIR filter arrays, it is not necessary to truncate the measured HRIRs if $D \leq 8$ in order to satisfy the cost constraint. And for the 'double-cost' array no truncation in required for $D \leq 16$. Hence the approximation error for the two FIR arrays will be zero for $D \leq (8, 16)$.

The 'auditory' $\mathcal{L}^2$ error that we report has been previously employed in HRTF approximation studies [4]. This error measure is computed by warping the log-magnitude response of both the ideal and low-order systems to a log-frequency scale. A fifth-octave smoothing filter is then applied to both responses in order to model the critical bands of the auditory system. The

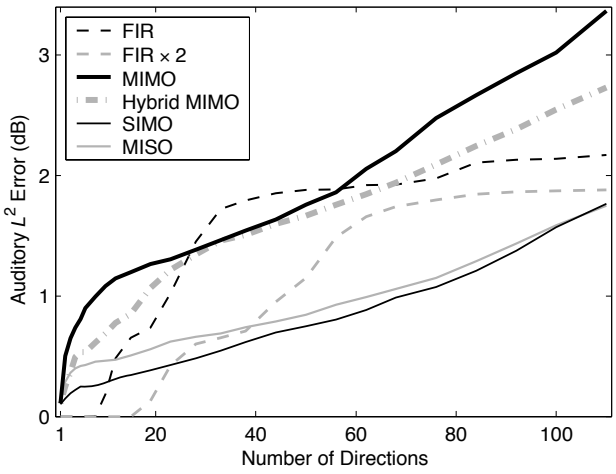

Figure 9: Auditory $\mathcal{L}^2$ error as a function of D.

$\mathcal{L}^2$ error (RMSE) between the two modified responses is then computed over the range 300 Hz to 16 kHz. For systems with multiple inputs or outputs, the auditory $\mathcal{L}^2$ error is computed for each input-output pair, and averaged.

Note that the lower bound of the $\mathcal{L}^2$ integration is lower than in [4]. It is generally known that, in free-field conditions, ILD cues below 1kHz do not greatly affect the localization of sound sources [1]. However, it has also been shown that low-pass sources can be localized in elevation [41]. Furthermore, studies have shown that sensitivity to ILD cues extends below 1kHz in the presence of reflecting surfaces [42], and that the sense of 'externalization' is affected by ILD cues below 1 kHz [43]. 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 [44]. 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 9 shows the auditory $\mathcal{L}^2$ results for the same arrangement of systems as Figure 8. 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 as D approaches 100. The approximation error of

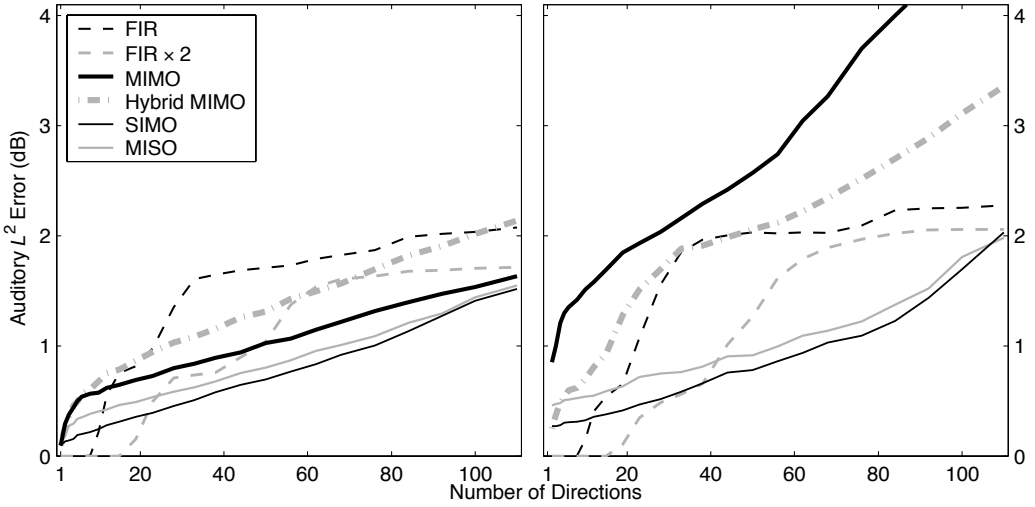

Figure 10: Auditory $\mathcal{L}^2$ error for MIMO systems as a function of D for ipsilateral and median directions (left) and contralateral directions (right).

the two FIR arrays exhibits a pronounced knee at $D = 30$ and $D = 60$. This knee is due in part to the rate at which the order of the FIR arrays decreases as D increases (note that the vertical axis in Figure 8 is logarithmic).

In contrast to the SIMO and MISO architectures, the two MIMO architectures do not perform well. For $30 < D < 60$, the state-space systems yield slightly lower error than an FIR array of equal cost. For no D does either MIMO system yield lower error than the 'double-cost' FIR array. The mediocre performance can be elucidated by examining the performance of ipsilateral and contralateral responses separately, as shown in Figure 10. The MIMO systems perform relatively well for the ipsilateral responses, but poorly for the contralateral responses. This performance disparity is apparent because the auditory $\mathcal{L}^2$ error is a log-magnitude metric; the ILD lessens the influence of the contralateral approximation with linear-magnitude metrics. The poor performance of the MIMO systems is not due to the ILD, however. As described in Section 2.2, the poor approximation of the contralateral HRTFs is due to the ITD. Indeed, we found that if the ITD is factored out of the MIMO state-space system, then the performance is similar to that of the SIMO and MISO systems. For the SIMO and MISO architectures, the approximation error for the ipsilateral and contralateral

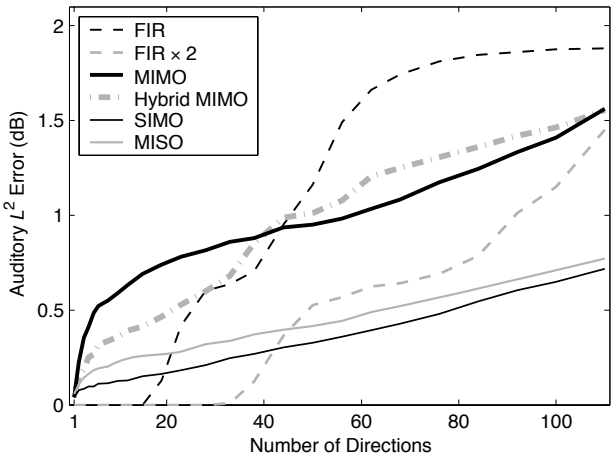

Figure 11: Auditory $\mathcal{L}^2$ error as a function of D. The same results are shown as Fig. 9, with double the cost bound.

responses is similar.

The relative performance of the state-space systems improves if we consider systems with higher cost bounds. Figure 11 shows that same results as Figure 9 if the experiment is repeated with $C_{\max} = 8000$, approximately the cost of 16 full-order HRIR pairs. The SIMO and MISO state-space systems again outperform the MIMO state-space systems. However, in this case the MIMO state-space systems outperform the FIR array for $D > 45$. Overall we found the relative performance of the state-space systems to improve in the 'high-cost' domain. However, for cost bounds as low $C_{\max} = 2000$ we found state-space systems exhibited lower approximation error than equal cost FIR arrays for $D > 40$.

In terms of auditory $\mathcal{L}^2$ error, the hybrid MIMO system yield performance that is comparable to that of the state-space MIMO system. However, this error measure reflects only the error between the magnitude responses and does not reflect phase distortion. We report the ITD distortion next.

3.4 ITD Error Results

We report an intuitive measure of ITD distortion below, rather than a conventional analytic measures of phase distortion. Numerous studies have shown that so long as the ITD is properly modeled, and the HRTF filters are nearly

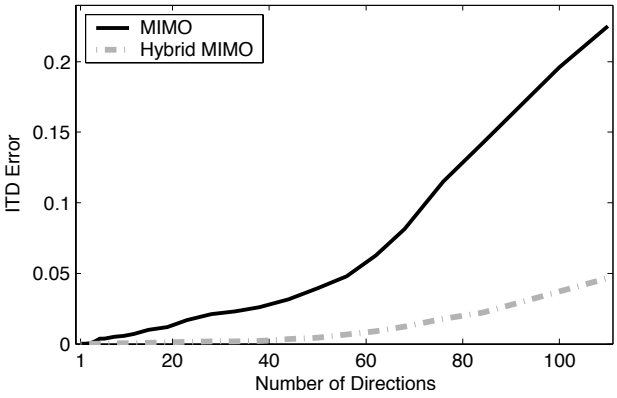

Figure 12: ITD error as a function of D for four MIMO systems.

minimum-phase, the precise phase response of the system is not perceptually salient [34,43]. Hence, we consider a measure of phase error that reflects only the distortion in the time-delay of the contralateral responses. For a single contralateral impulse response, the ITD error is defined as the fraction of the impulse response energy that appears before the time-delay prescribed by the ITD. Hence this error lies between zero and one. The ITD error is computed for each contralateral response in the system, and averaged.

Only the MIMO architectures exhibit ITD error, as the ITD was modeled externally by the other architectures. Hence, we examine the ITD distortion of only the state-space and hybrid MIMO systems. Figure 12 shows the ITD error for two MIMO systems with a cost bound of $C_{\max} = 4000$. The hybrid system exhibits much less ITD error than the conventional state-space systems. It is unclear however if the ITD error of the hybrid systems is small enough for binaural applications.

4 Discussion

Informal listening tests confirm the numerical results presented above: for a fixed C we are able to construct state-space systems that are less discriminable from measured HRIRs than FIR arrays of equal cost C. A formal experiment is currently being conducted. The audible artifacts that sometimes appear in the design of low-order IIR filters for audio applications [16,28] are not apparent in any of the state-space HRTF systems.

Comparing systems with different architecture is difficult. As mentioned in Section 2.1, the SIMO architecture has the obvious disadvantage of only being able to display a single monaural source signal. While this architecture exhibits the best performance, the single-input constraint is too restrictive for many applications. The MIMO and MISO architectures do not have this limitation. The performance of the MISO architecture is promising, and somewhat surprising. Apparently the reduced system order for the MISO architecture does not degrade performance unduly. Overall, the MISO architecture, originally proposed by Georgiou and Kyriakakis appears to be the best choice for binaural environment modeling [28]. We would like to find a more satisfying solution for the ITD distortion in the MIMO architecture: we found that MIMO systems perform very well if the ITD is neglected. Furthermore, the MIMO architecture may be expanded to model the HRTFs for multiple listeners simultaneously (with multiple pairs of outputs) without only a modest increase in computational cost.

The computational cost C that we defined in Section 2.3 is independent of the choice of data type: fixed-point versus floating-point. Fixed-point computations are simpler to implement, whereas floating-point computations are more robust. In practice, for FIR filters this choice is of little significance, as coefficient quantization error is typically benign for systems without feedback. For IIR filters and state-space systems, the issue requires more caution [45,46]. For the state-space systems considered here, we have performed a simple analysis of the state values and found that they yield first- and second-order statistics that are very similar to those of the input signal[8]. We have also examined the affect of small perturbations on the feedback matrix, $\widehat{\mathbf{A}}$, and found the affects on the resulting transfer functions to be small in most cases. While by no means sufficient, these two experiments imply that fixed-point data types may be sufficient for state-space binaural displays. Additionally, a recent study concluded for audio IIR filter design that fixed-point data types may be preferable to floating-point data types [47].

5 Conclusion

The present work explores low-order state-space models of HRTFs. Many contemporary binaural environment models are implemented with an array

[8]Indeed, one can listen to any one state value, and it will 'sound' quite similar to the monaural input signal

of HRTF filters. If more than 20 directions are included in the array, we found that the array can be replaced with a state-space system of lower computational cost. Three state-space architectures are considered: MIMO, SIMO and MISO. The MIMO architecture is perhaps the most general and intuitive, but performance suffers due to the ITD. We propose an ad hoc hybrid system to mediate this problem; the ITD distortion is greatly reduced, but the 'compensation' filters limit the computational savings of this architecture. Because the ITD can be modeled externally with the SIMO and MISO architectures, the performance of these systems is substantially better. Overall, the MISO architecture appears to be best suited for the binaural display application.

Binaural displays are an ideal candidate for reduced-order state-space models, as there is a clear need to model multiple similar transfer functions simultaneously. We have demonstrated that for this application and state-space systems can achieve a significant computational savings relative to conventional filter arrays.

References

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

[2] 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.

[3] C. Cheng and G. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *J. Audio Eng. Soc.*, vol. 9, no. 4, pp. 231–249, 2001.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

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

[22] S. Begault, "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.

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

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] K. Glover, "All optimal Hankel-norm approximations of linear multivariable systems and their $\updownarrow^{\infty}$-error bounds," *Int'l. J. Control*, vol. 39, pp. 1115–1193, 1984.

[31] N. Adams, "Model reduction of head-related transfer-function arrays in the state-space," 2007, Technical Report available online at http://www.eecs.umich.edu/systems/TechReportList.html.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge, UK: Cambridge University Press, 1990.

[39] N. Adams and G. Wakefield, "Efficient binaural display using MIMO state-space systems," *Proc. IEEE Int. Conf. on Acout. Speech and Sig. Proc.*, April 2007, Honolulu, HI.

[40] 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.

[41] 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.

[42] 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.

[43] W. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3678–3688, June 1996.

[44] 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.

[45] S. Hwang, "Roundoff Noise in State-Space Digital Filtering: A General Analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 256–262, June 1976.

[46] J. Whidborne, D.-W. Gu, J. Wu, and S. Chen, "Optimal controller and filter realizations using finite-precision, floating-point arithmetic," *Int. J. of Systems Science*, vol. 36, no. 7, pp. 405–413, June 2005.

[47] J. Moorer, "48-Bit Integer Processing Beats 32-Bit Floating-Point for Professional Audio Applications," in *Proc. 107th AES Convention*, Sept. 1999, preprint 5038.