EFFICIENT BINAURAL DISPLAY USING MIMO STATE-SPACE SYSTEMS

Norman Adams and Gregory Wakefield

University of Michigan, 2260 Hayward St., Ann Arbor, MI 48109
{nhadams,ghw}@umich.edu

ABSTRACT

Binaural displays for immersive listening must model reverberant acoustic environments, multiple sound sources, and compensate for head motion. Many displays accomplish this by convolving collections of spatially distributed sources with head-related transfer functions (HRTFs). In the absence of reflections, the computational load scales linearly with the number of sources; however, when reflections are present, the load scales exponentially such that the binaural display reaches the limit of available processing power for even relatively sparse acoustic scenes. We propose a method that significantly eases this exponential growth by formulating the HRTF filter array as a MIMO state-space system. Two MIMO architectures are explored; the relative merits of each are found to depend on the specific application. Hankel-optimal 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.

Index Terms— Acoustic signal processing, headphones, MIMO systems, reduced order systems, Hankel matrices

1. INTRODUCTION

In the ideal case, an acoustic source is a motionless point in an anechoic environment which lies in the far field of the listener. In such cases, the relationship between the source signal and the signals at the listener's ears is completely determined by a pair of *head related transfer functions* (HRTFs). Such transfer functions can be implemented using appropriately measured *head related impulse responses* (HRIRs), which, empirically, require approximately 200 taps at a cost of 17.6 MIPS for a sampling rate of 44.1 kHz. Because this is a substantial computational load, considerable effort has been made to find low-order approximations to measured HRIRs [1, 2]. In the more practical case where the environment is reverberant [3] and spatially-extended sources [4] undergo motion relative to the listener [5], the savings gained by such low-order approximations is rapidly spent on the exponential rise in computational cost when each source or reflection is convolved with direction-appropriate HRIRs. The challenge in *binaural displays* concerns the best system architecture to move from the synthesis of ideal cases to realistic acoustic scenes.

The present work explores reduced-order *multiple-input multiple-output* (MIMO) state-space systems in which one or more monaural signals are filtered with numerous HRTFs simultaneously. We argue that MIMO state-space architectures offer substantial computational improvements relative to the standard FIR filter arrays. Section 2 describes two state-space architectures and two order-reduction techniques for the HRTF filter array problem. For one architecture, the *interaural time difference* (ITD) is found to degrade the approximation, and a hybrid method is proposed to mediate this problem. System performance is characterized in the Section 3.

1.1. Background

HRTFs measured at different directions are correlated [2]. Numerous studies have found that collections of HRTFs can be reasonably represented in low-dimensional spaces. However, such representations do not yield low cost filters for individual HRTFs. A system that models HRTFs at many directions simultaneously may be able to utilize the correlation properties of HRTFs to reduce the net cost of the system.

Two recent studies propose state-space systems that model HRTFs at multiple directions. In [6] MISO systems are designed that model multiple HRTFs for each ear. HRTF correlation is not utilized in this work however, as separate systems are designed for each HRTF individually, and then merged into one large system. In contrast, [7] considers a MIMO state-space design that directly models many HRTFs. Both studies employed a *balanced model truncation* (BMT) technique to design low-order systems [8], and found that with sufficiently high order, state-space systems well approximated the measured HRTFs. However, neither study considered the computational advantages of state-space implementations.

2. METHODS

Consider a discrete-time MIMO state-space system,

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

The authors thank Dr. Thomas Santoro for his assistance in measuring HRTFs. This work was supported in part by a scholarship from the AFCEA, and ONR N0001406WX20041.

where $\mathbf{x}[n]$ is the state vector of size N, $\mathbf{u}[n]$ is the input vector of size P, and $\mathbf{y}[n]$ is the output vector of size M. The matrix impulse response of this system is

$$\begin{aligned}\mathbf{h}[n] &= \begin{bmatrix} h_{11}[n] & \dots & h_{1P}[n] \\ \vdots & \ddots & \vdots \\ h_{M1}[n] & \dots & h_{MP}[n] \end{bmatrix} \\ &= \begin{cases} \mathbf{C}\mathbf{A}^{n-1}\mathbf{B} & n > 0 \\ \mathbf{0} & n \le 0 \end{cases}\end{aligned} \tag{2}$$

For state-space systems of this form, the time-domain behavior of the system can be represented by the Hankel matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}[1] & \mathbf{h}[2] & \mathbf{h}[3] & \dots \\ \mathbf{h}[2] & \mathbf{h}[3] & & \\ \mathbf{h}[3] & & & \\ \vdots & \vdots & & \ddots \end{bmatrix} \tag{3}$$

We seek matrices $\big(\mathbf{A}, \mathbf{B}, \mathbf{C}\big)$ such that the matrix impulse response is a convenient arrangement of the HRIRs. Below we describe two arrangements of the HRIRs in the matrix impulse response.

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, there are two obvious choices of system architecture for the $2D$ transfer functions. The state-space system can be implemented using either a SIMO architecture with one input, $2D$ outputs and matrix impulse response

$$\mathbf{h}[n] = \begin{bmatrix} h_1^L[n] & h_1^R[n] & h_2^L[n] & h_2^R[n] & \dots & h_D^R[n] \end{bmatrix}^T \tag{4}$$

or a MIMO architecture with D inputs, 2 outputs and matrix impulse response

$$\mathbf{h}[n] = \begin{bmatrix} h_1^L[n] & h_2^L[n] & \dots & h_D^L[n] \\ h_1^R[n] & h_2^R[n] & \dots & h_D^R[n] \end{bmatrix} \tag{5}$$

The two architectures have relative advantages and disadvantages. Both architectures can readily accommodate acoustic reflections and motion by placing a scale-and-delay filter array either after the state-space system for the SIMO architecture, or before the state-space system for the MIMO architecture. The SIMO architecture has the disadvantage that multiple sources at different locations cannot be presented simultaneously. However, the MIMO architecture has the disadvantage that the ITD must be included in the state-space system, which decorrelates $\mathbf{h}[n]$. For the SIMO architecture, the ITD can be implemented externally.

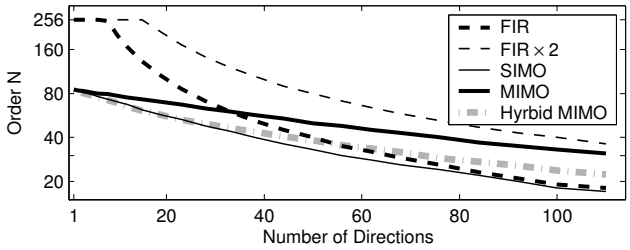

Fig. 1. System order as a function of D for four systems (FIR, SIMO, MIMO, and Hybrid MIMO) with cost $C = 4000$, and one system (FIR$\times 2$) with cost $C = 8000$.

2.2. Model-order reduction

Using, for example, the controller canonical form, construction of state-space systems that implement the measured HRIRs exactly is straightforward. Such state-space systems are high order and far more computationally expensive than conventional FIR filter arrays. Accordingly, we consider two popular techniques to reduce the system order: BMT and *Hankel-norm optimal approximation* (HOA) [9]. We find that systems with as few as $N \simeq 20$ states can reasonably approximate measured HRTFs, even for systems with many directions, $D \simeq 100$.

BMT technique. For a given set of D directions, the HRIRs are arranged into either a SIMO or MIMO matrix impulse response, with an extra zero prepended to each HRIR, as the state-space system in (1) has no feed-through term. For MIMO systems the ITD is included, but is removed for the SIMO systems. Using BMT, an order N system can be designed by: 1. constructing the Hankel matrix $\mathbf{H}$, 2. computing the SVD of $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, 3. conformally partitioning $\big(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}\big)$ along the N^{th} row of $\mathbf{\Sigma}$, and 4. constructing $\big(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}\big)$ from the partitioned SVD [8].

Metric for determining optimality. While the BMT method of model reduction is convenient, it is not optimal in any specific sense. One metric for which optimal solutions are known for MIMO systems is the *Hankel-norm*. Let $(\sigma_1 \ge \sigma_2 \ge \sigma_3 \cdots)$ be the singular values of $\mathbf{H}$, the main diagonal of $\mathbf{\Sigma}$. The largest singular value, σ_1, is known as the Hankel-norm of the system $\big(\mathbf{A}, \mathbf{B}, \mathbf{C}\big)$. For low-order approximation $\big(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}\big)$ the Hankel error of the system is $\sigma_1\big(\mathbf{H} - \widehat{\mathbf{H}}\big)$, where $\widehat{\mathbf{H}}$ is the Hankel matrix of the low-order system. Interpreting the Hankel error is facilitated by comparing it to the L_∞ spectral error, which is common in acoustic applications. The Hankel error is a lower bound for the L_∞ spectral error

$$\sigma_1\big(\mathbf{H} - \widehat{\mathbf{H}}\big) \le \max_{\omega}\ \sigma_1\big(H(\omega) - \widehat{H}(\omega)\big) \tag{6}$$

where $H(\omega)$ and $\widehat{H}(\omega)$ are the measured and approximated matrix frequency responses, respectively, and have the same dimensions as $\mathbf{h}[n]$. For state-space systems, the Hankel error is often found to be a tight lower bound.

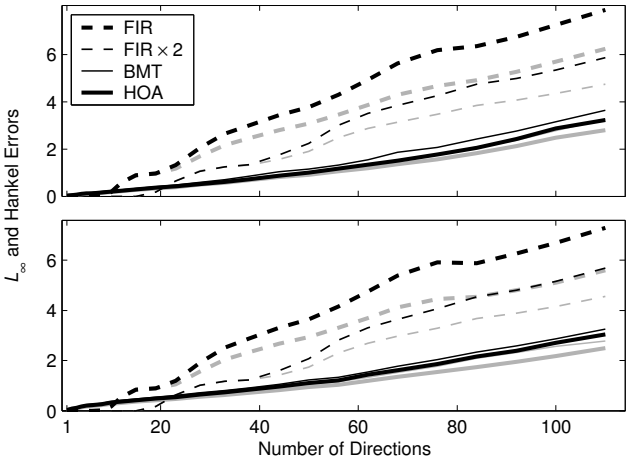

Fig. 2. L_∞ and Hankel error for four SIMO systems (top panel) and four MIMO systems (bottom panel). The L_∞ error is shown by blacks lines, and the Hankel error by gray lines.

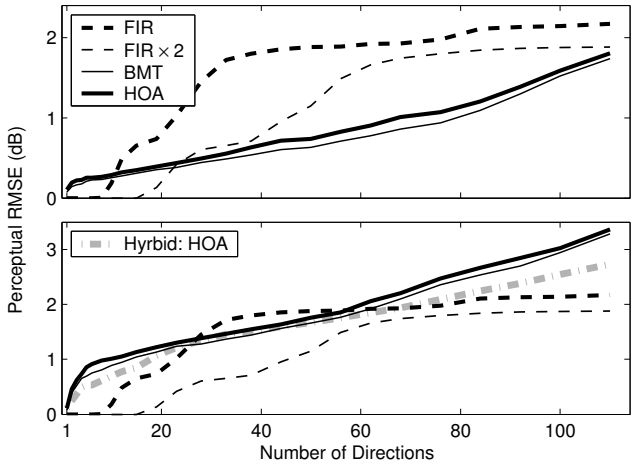

Fig. 3. Average 'perceptual' RMSE for SIMO (top panel) and MIMO (bottom panel) architectures.

HOA technique. The HOA method is somewhat involved, and described in [10]. An order N system designed using HOA minimizes the Hankel error, which can be shown to equal the $(N+1)^{\text{th}}$ singular value of the original system,

$$\sigma_1\big(\mathbf{H} - \widehat{\mathbf{H}}\big) = \sigma_{N+1}\big(\mathbf{H}\big) \tag{7}$$

2.3. Hybrid MIMO system

BMT and HOA excel at approximating systems of transfer functions with similar phase behavior, such as HRTFs, if the linear-phase term of the contralateral HRTFs is removed [6]. We have confirmed that including the ITD in $\mathbf{h}[n]$ increases the singular values of $\mathbf{H}$ and degrades the approximation. In particular, the contralateral impulse responses are 'smeared,' such that there is no longer a clear ITD. For orders $N < 50$, this phase distortion is audible for directions with large ITD.

To address the problems associated with the contralateral HRTFs, we propose a hybrid MIMO system that uses a state-space system for the HRTFs with little or no time delay, and FIR filters for the HRTFs with large delay. The design procedure is similar to that described above, except the HRIRs with large delay ($> 300\mu s$) are zeroed-out for the design of the state-space system, and FIR filters are connected between the corresponding input/output pairs. The FIR and state-space system orders are chosen such that the FIR filters account for no more than a third of the total computational cost. In this way, the desired phase response is preserved even for small N.

3. PERFORMANCE CHARACTERIZATION

Truncated minimum-phase FIR filters, as used in many contemporary displays, provide a baseline for comparison with the state-space systems. Performance is presented as a function of D, the number of directions, for a fixed computational cost. We define the computational cost as the total number of multiplies per sample period, as used in [1, 4]. FIR filter arrays and state-space systems are designed from an HRTF dataset measured from 8 listeners at 253 spatial directions [2]. The measured HRIRs have length 256 at $f_s = 44.1$kHz. For every system designed, D of the 253 directions are chosen such that the density around the listener is approximately uniform. For $1 \leq D \leq 110$, separate systems are designed for each listener, and results are averaged across listeners.

An order N FIR filter array with P inputs and M outputs requires $C = PM(N+1)$ multiples per sample period. Any MIMO state-space system of the form (1) can be converted to Schur form, in which case the A matrix is block triangular and the computational cost of the system is $C = N^2/2 + (P + M + 1)N$. Fig. 1 shows the filter order of the four systems discussed above for equal cost. Filter order N is chosen such that the total cost of the system is $C \leq 4000$. $C = 4000$ is approximately equal to the cost of implementing eight full-order binaural HRIRs. This cost bound applies to all results presented below, unless otherwise noted. To better gauge the relative performance of the state-space systems, an FIR filter array of twice the computational cost is also shown.

We first consider the approximation quality in terms of the L_∞ and Hankel errors, as described in § 2.2. Fig. 2 shows these two errors for four SIMO architectures (top panel) and four MIMO architectures (bottom panel). Four implementations of each architecture are shown, three of equal cost (one FIR, and two state-space, using BMT and HOA) and one of double cost (FIR $\times 2$). For all systems, the L_∞ error is bounded from below by the Hankel error, but for the state-space systems, the Hankel error provides a tight bound on the L_∞ error. For the FIR filter array, the error is zero for $D \leq 8$. However, for large D the error of the FIR systems

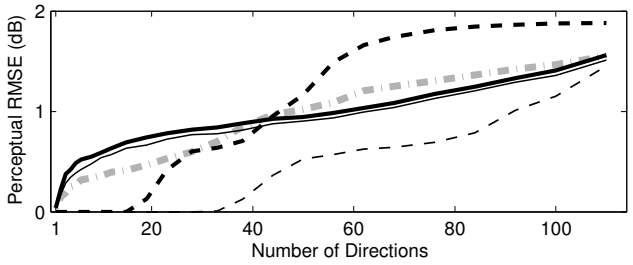

Fig. 4. Average 'perceptual' RMSE for four MIMO systems with cost $C = 8000$, and one MIMO system with cost $C = 16000$. See Fig. 3 legend.

grows more rapidly than the state-space systems such that for $D > 20$, the state-space systems achieve lower error than even the FIR $\times 2$ filter array. The state-space systems yield similar error, although the HOA method achieves slightly lower error.

A common perceptual-error metric for audio applications is the L_2 norm applied to the log-magnitude log-frequency error spectrum after critical-band smoothing [1, 4]. For the MIMO systems considered here, the error is measured for each input/output pair and averaged. Fig. 3 shows the average error for four SIMO architectures (top panel) and five MIMO architectures (bottom panel). The same systems are shown as in Fig. 2, with the addition of one hybrid MIMO system.

For the SIMO architecture, the state-space systems also yield significantly lower error for $D > 20$. As D increases beyond 100 directions, the order of the state-space systems fall below $N = 20$ and performance rapidly deteriorates. The BMT method yields slightly lower perceptual RMSE than the HOA method, even though HOA yields lower L_∞ error. This is due in part to the difference between the L_∞ and L_2 norms. The HOA method distributes the spectral error uniformly across frequency, a necessary condition for reducing the L_∞ error. In contrast, BMT is often found to yield peaks in the error at or near spectral notches in the original transfer functions [9], which may be undesirable given the perceptual importance of HRTF notches.

While MIMO state-space systems yield promising performance in Figs. 2 and 3, the phase distortion due to ITD is problematic for binaural displays. The hybrid MIMO system described in § 2.3 retains the desired phase response. The performance of the hybrid MIMO system is similar for BMT and HOA. Fig. 3 shows performance of a hybrid MIMO system designed using HOA. The hybrid MIMO system yields a modest improvement over the FIR filter array for $30 < D < 60$. The relative performance of the MIMO state-space systems improves if a higher computational cost bound is used. Fig. 4 shows the average perceptual RMSE for the same MIMO systems if the cost bounds are doubled.

Formal listening tests to validate these methods are in preparation, but informal listening tests confirm the numerical results presented above. Furthermore, we found the state-space implementations to be robust to coefficient quantization error, unlike some IIR filter designs.

4. CONCLUSION

We have shown that state-space systems designed using Hankel methods offer a substantial computational savings over conventional filter arrays for binaural displays that filter many directions simultaneously. BMT and HOA yield similar performance, although subtle differences are demonstrated. The SIMO architecture offers a direct improvement over conventional methods, whereas the MIMO architecture requires a hybrid approach to preserve the ITD, and only offers an advantage for relatively high cost systems.

5. REFERENCES

[1] 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.

[2] 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.

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

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

[6] P. Georgiou and C. Kyriakakis, "Modeling of Head Related Transfer Functions for Immersive Audio Using a State-Space Approach," in *Proc. IEEE Asil. Conf. on Sig., Sys. and Comp.*, 1999, vol. 1, pp. 720–724.

[7] 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.

[8] 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.

[9] 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 ICASSP*, 1994.

[10] 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.