Modeling of Head Related Transfer Functions for Immersive Audio Using a State-Space Approach

Panayiotis Georgiou and Chris Kyriakakis
Immersive Audio Laboratory – Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-2564
georgiou@sipi.usc.edu, ckyriak@imsc.usc.edu

Abstract

Accurate localization of sound in 3-D space is based on variations in the spectrum of sound sources. These variations arise mainly from reflection and diffraction effects caused by the pinnae and are described through a set of Head-Related Transfer Functions *(HRTF's) that are unique for each azimuth and elevation angle. A virtual sound source can be rendered in the desired location by filtering with the corresponding HRTF for each ear. Previous work on HRTF modeling has mainly focused on methods that attempt to model each transfer function individually. These methods are generally computationally-complex and cannot be used for real-time spatial rendering of multiple moving sources. In this work we provide an alternative approach, which uses a multiple-input single-output state-space system to create a combined model of the HRTF's for all directions. This method exploits the similarities among the different HRTF's to achieve a significant reduction in the model size with a minimum loss of accuracy.*

1. Introduction

Applications for 3-D sound rendering include teleimmersion; augmented and virtual reality for manufacturing and entertainment; teleconferencing and telepresence; air-traffic control; pilot warning and guidance systems; displays for the visually impaired; distance learning; and professional sound and picture editing for television and film. Work on sound localization finds its roots as early as the beginning of the twentieth century when Lord Rayleigh [10] first presented the Duplex Theory that emphasized the importance of *Interaural Time Differences* (ITD) and *Interaural Amplitude Differences* (IAD) in source localization. It is notable that human listeners can detect ITD's as small as $7\mu s$ [9], which makes it an important cue for localization. Nevertheless, ITD's and IAD's alone are not sufficient to explain localization of sounds in the median plane, in which ITD's and IAD's are both zero.

Variations in the spectrum as a function of azimuth and elevation angles also play a key role in sound localization. These variations arise mainly from reflection and diffraction effects caused by the outer ear (pinna) that give rise to amplitude and phase changes for each angle. These effects are described by a set of functions known as the *Head-Related Transfer Functions* (HRTF's).

One of the key drawbacks of 3-D audio rendering systems arises from the fact that each listener has HRTF's that are unique for each angle. Measurement of HRTF's is a tedious process that is impractical to perform for every possible angle around the listener. Typically, a relatively small number of angles are measured and various methods are used to generate the HRTF's for an arbitrary angle. Previous work in this area includes modeling using principal component analysis [7], as well as spatial feature extraction and regularization [2].

In this paper, we present a two-layer method of modeling HRTF's for immersive audio rendering systems. This method allows for two degrees of control over the accuracy of the model. For example, increasing the number of measured HRTF's improves the spatial resolution of the system. On the other hand, increasing the order of the model extracted from each measured HRTF improves the accuracy of the response for each measured direction. Kung's method [8] was used to convert the time-domain representation of HRTF's in state-space form. The models were compared both in their *Finite Impulse Response* (FIR) filter form and their state-space form. It is clear that the state-space method can achieve greater accuracy with lower order filters. This was also shown using a balanced model truncation method [11]. Although an *Infinite Impulse Response* (IIR) equivalent of the state-space filter could be used without any theoretical loss of accuracy, it can often lead to numerical er-

0-7803-5700-0/99/$10.00

rors causing an unstable system, due to the large number of poles in the filter. State-space filters do not suffer as much from the instability problems of IIR filters, but require a larger number of parameters for a filter of the same order. However, considering that there are similarities among the impulse responses for different azimuths and elevations, a combined single system model for all directions can provide, as we will show, a significant reduction.

Previous work on HRTF modeling has mainly focused on methods that attempt to model each direction-specific transformation as a separate transfer function. In this paper we present a method that attempts to provide a single model for the entire 3-D space. The model builds on a generalization of work by Haneda *et al.* [6], in which the authors proposed a model that shares common poles (but not zeros) for all directions. Our model uses a multiple-input single-output state-space system to create a combined model of the HRTF's for all directions simultaneously. It exploits the similarities among the different HRTF's to achieve a significant reduction in the model size with a minimum loss of accuracy.

2. Spatial Audio Rendering

One way to spatially render 3-D sound is to filter a monaural (non-directional) signal with the HRTF's of the desired direction. This involves a single filter per ear for each direction and a selection of the correct filter taps through a lookup table. The main disadvantage of this process is that only one direction can be rendered at a time and interpolation can be problematic. In our work we extract and model the important cues of ITD and IAD as a separate layer, thus avoiding the problem of dual half-impulse responses created by interpolation. The second layer of the interpolation deals with the angle-dependent spectrum variations (Fig. 1). This is a multiple-input single-output system (for each channel), which we created in state-space form.

The signal for any angle θ can be fed to the input corresponding to that angle, or if there is no input corresponding to θ then the signal can be split into the two adjacent inputs (or more in the case of both azimuth and elevation variations). In order to proceed with the two-layered model described above, we first extract the delay from the measured impulse responses. Fig. 2 shows the delay extracted from the measurements and fitted with a sixth order polynomial.

It should be noted that here the azimuth is measured from the center of the head relative to the midcoronal and towards the face as shown in Fig. 3 and not relative to the midsagittal and clockwise as is common practice. For example, the azimuth of 270° relative to the midsagittal corresponds to 180° for the right ear but to 0° for the left ear measured with this proposed convention. This method of representation was chosen because it allows us to use a common delay

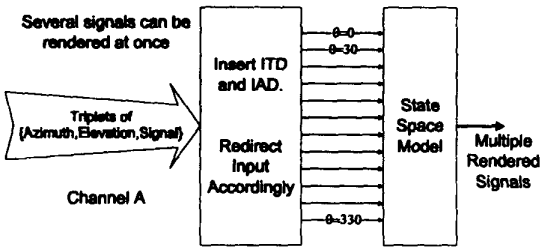

Figure 1. The unprocessed signals are passed to the algorithm along with the desired azimuth and elevation angles of projection.

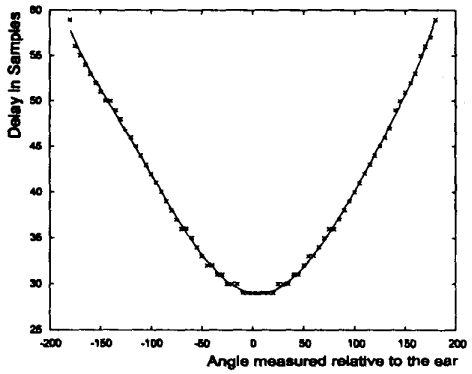

Figure 2. Extracted delay and sixth order polynomial fit.

function for both ears.

Similarly, we can approximate the gain with a 14th order polynomial as in Fig. 4. The advantages of polynomial fitting are not so obvious when only one elevation is considered, but become more evident when the entire 3-D space is taken into consideration.

3. FIR Filter Reduction

The measurements used in this paper consist of impulse responses taken using a KEMAR dummy head [5]. These 512-point impulse responses can be used as an FIR model against which our comparisons will be based. In order to reduce these impulse responses we used the method first proposed by Kung [8] at the 12th Asilomar Conference on Circuits, Systems and Computers. The one input-one output case is briefly described below.

Note that alternative methods can be used (see Mackenzie *et al.* [11]). For this and other methods, the reader can refer to the original paper by Kung [8], as well as Beliczynski *et al.* [1] and references therein.

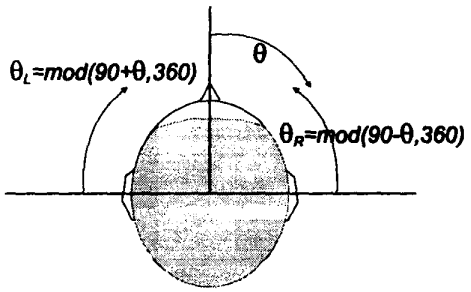

Figure 3. Proposed convention of measuring azimuth in order to have a single delay and gain function for both ears.

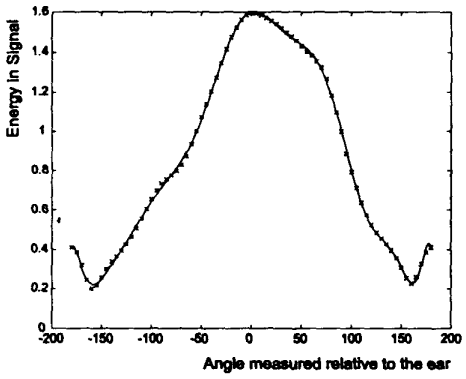

Figure 4. Extracted energy and a twelfth order polynomial fit.

Consider an impulse response model of a causal, stable, multivariable and linear time-invariant system. If the system state space model is

$$x(n+1) = Ax(n) + Bu(n) \quad (1)$$
$$y(n) = Cx(n) + Du(n) \quad (2)$$

and an impulse is applied to the system then (assuming that $u_0 = 1$, without loss of generality):

$$\begin{array}{lllllll} y_0 & = & D & & & & \\ x_1 & = & B & y_1 & = & CB & \\ x_2 & = & AB & y_2 & = & CAB & \\ x_3 & = & A^2B & y_2 & = & CA^2B & (3) \\ & \cdots & & & \cdots & & \\ & \cdots & & & \cdots & & \\ x_N & = & A^NB & y_N & = & CA^NB & \end{array}$$

Forming the above into a matrix:

$$\begin{bmatrix} y(n) \\ y(n+1) \\ y(n+2) \\ \vdots \end{bmatrix} = \begin{bmatrix} CB & CAB & CA^2B & \cdots \\ CAB & CA^2B & CA^3B & \cdots \\ CA^2B & CA^3 & \cdots & \cdots \\ CA^3B & CA^4B & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} u(n) \\ 0 \\ \vdots \\ \vdots \end{bmatrix}$$

Separating the Hankel matrix (i.e., the matrix that in position (i,j) is $CA^{i+j-1}B$) and expressing it in its Singular Value Decomposition (SVD) components:

$$H = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} B & AB & A^2B \ldots \end{bmatrix} = \Omega\Gamma = U\Sigma V^T \quad (4)$$

where U, V are unitary matrices and Σ contains the singular values along its diagonal in decreasing magnitude, i.e.,

$$\Sigma = \mathrm{Diag}[\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_r, \sigma_{r+1}, \ldots \sigma_{N+1}] \quad (5)$$

and Ω and Γ are the extended observability and reachability matrices that can be expressed in terms of the SVD components of H as:

$$\Omega = U\Sigma^{\frac{1}{2}} \text{ and } \Gamma = \Sigma^{\frac{1}{2}}V^T \quad (6)$$

One way to reduce the model is to use

$$H = \begin{bmatrix} U_n & \overline{U_n} \end{bmatrix} \cdot \begin{bmatrix} \Sigma_n & 0 \\ 0 & \overline{\Sigma_n} \end{bmatrix} \begin{bmatrix} V_n \\ \overline{V_n}^T \end{bmatrix} \quad (7)$$

and reduce Ω and Γ to:

$$\Omega_n = U_n\Sigma_n^{\frac{1}{2}} \text{ and } \Gamma_n = \Sigma_n^{\frac{1}{2}}V_n^T \quad (8)$$

This will give:

$$\begin{array}{lllll} A & = & \Sigma^{-\frac{1}{2}}U_n^TU_n^{\uparrow}\Sigma^{\frac{1}{2}} & C & = & U_n^1\Sigma^{\frac{1}{2}} \\ B & = & \Sigma^{\frac{1}{2}}\left(V_n^1\right)^T & D & = & y_0 \end{array} \quad (9)$$

While there are several definitions for U_n and $U_n^{\uparrow}$, one that also guarantees stability is

$$U_n = \begin{bmatrix} U_n^1 \\ \vdots \\ U_n^{N-1} \\ U_n^N \end{bmatrix} \text{ and } U_n^{\uparrow} = \begin{bmatrix} U_n^2 \\ \vdots \\ U_n^N \\ 0 \end{bmatrix} \quad (10)$$

To achieve higher speeds in model creation and the ability to handle any model size, Kung's method is performed on each impulse response separately. This avoids the dimension increase of the Hankel matrix and consequently drops the computational cost of the SVD significantly since SVD is an O(3) operation. The individual state-space models are combined in a single model to form the final model. Further reduction can be achieved on the resulting model if desired.

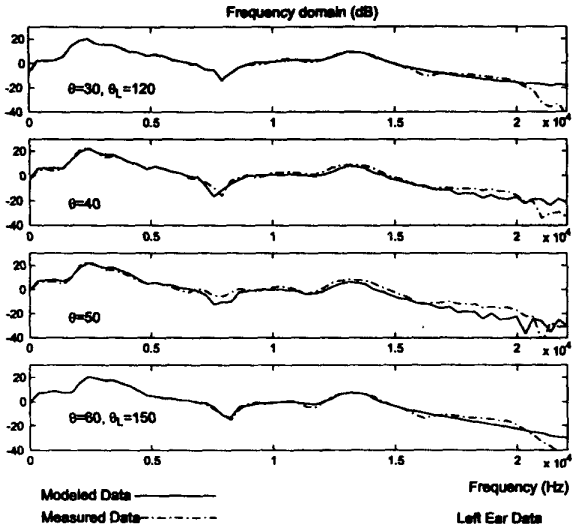

Figure 5. Frequency domain of measured and simulated impulse responses for a model created with a 30° resolution. $\theta = 40$, and $\theta = 50$ were not used for the creation of the model

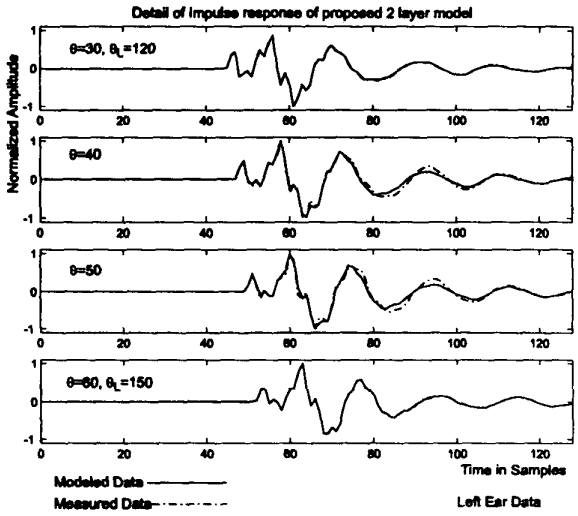

Figure 6. Detail of the time domain of Fig. 5

4. Results

The advantages of the two-layer HRTF model can better be observed by examining a few representative impulse responses. Figs. 5 and 6 show the measured data with a dashed line and the simulated data with a solid line. The model was created with data measured every 30°, and therefore only data from the first and last plot of each figure were used for the creation of the model. The other two simulated responses in the plot correspond to data synthesized from the 30° and 60° inputs of the state-space model. For example, angle 40° corresponds to $\frac{2}{3}$ of the input signal being fed through the 30° input, while the remaining $\frac{1}{3}$ is input to the 60° direction. As expected, the two main cues of delay and gain were preserved in the impulse response since they are generated from a separate, very accurate layer. The second layer can then be reduced according to the desired accuracy.

Fig. 7 shows the performance of a further reduced state space model. The model was reduced to less than a third its initial size (down to 191 states from 600). The reduction was performed using techniques as described in [3] and [12]. As can be seen from the figures, there was some minor loss of accuracy. Fig. 8 displays the performance of an equivalent model size that was created by reducing each individual HRTF to a 16 state model. These models correspond to a combined model of 192 states that is of equivalent size to the previous combined model but that performs very poorly. The advantage of performing the reduction to the combined model is clearly evident.

5. Conclusions and Future Research Directions

Although the state-space model is computationally expensive compared to an FIR filter, it provides several advantages over the latter while avoiding some of the disadvantages of IIR filters. Recent advances in FPGA technology allow large matrix multiplications at very high speeds that would make construction of a large size state-space model possible. Dandalis *et al.* [4] consider $N \times N$ times $N \times N$ matrix multiplication, which can be extended to $N \times N$ times $N \times 1$ multiplication (the most expensive operation in the state space representation). N can be given by [4]

$$\frac{N^2}{p \times f_{FPGA}} < \frac{1}{44.1kHz} \tag{11}$$

for a signal sampled at 44.1kHz, where f_{FPGA} is the FPGA clock frequency and p is the number of parallel multipliers.

Today's FPGA's with speeds exceeding 150MHz and $p > 100$ can easily handle state-space models of more than 500 states built on a single FPGA. As technology in this field is advancing with the System On a Chip model rapidly gaining ground, it will not be long before state-space models of more than a thousand states can be calculated in real time.

Another advantage that comes with the use of a state-space model is memory, which eliminates the audible "clicking" noise heard when changing from filter to filter. In fact, a model with many states eliminates the need for

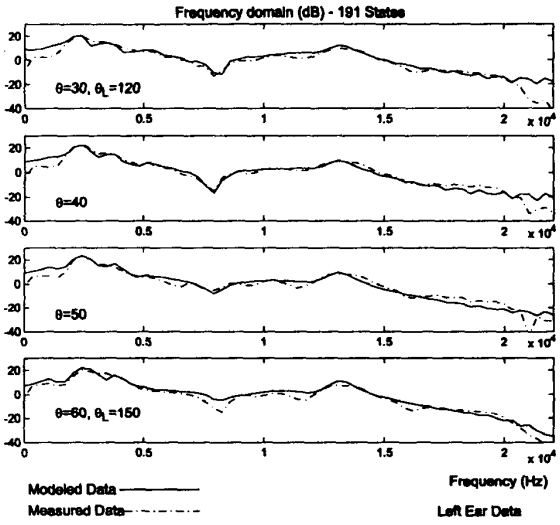

Figure 7. Model used is reduced down to 191 states from an original size of 600 states. Accuracy has not decreased significantly.

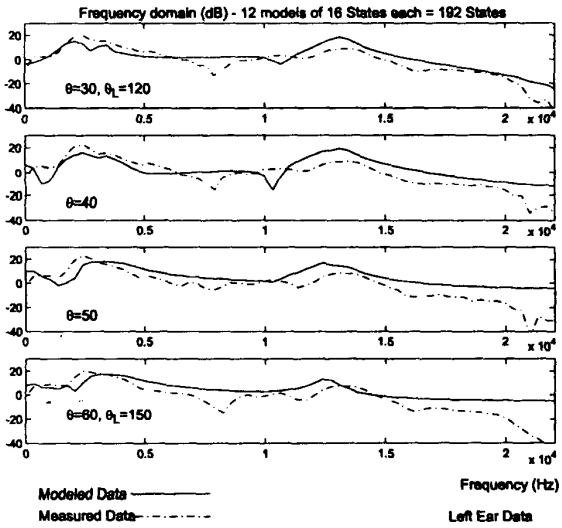

Figure 8. 12 Models of total 192 states. Accuracy has dropped significantly in comparison with Fig. 7 although model size is the same.

interpolation due to the memory. Interpolation, by passing a signal to two inputs at once, is however desirable to avoid sudden jumps in space of the virtual source.

Finally, we have demonstrated that while a single model for the whole space can achieve spatial rendering of multiple sources at once, it can also result in a smaller size than the individual models for all directions combined.

Acknowledgement: This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center with additional support from the Annenberg Center for Communication at USC and the California Trade and Commerce Agency. Equipment grants were provided by IBM and Intel Corporation.

References

[1] 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. *IEEE Trans. Acoust., Speech, and Signal Process.*, 3:593–596, April 1994.

[2] J. Chen, B. D. Van Veen, and K. E. Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *Journal of the Acoustical Society of America*, 97(1):439–52, January 1995.

[3] R. Y. Chiang and M. G. Safonov. *Robust Control Toolbox User's Guide*. The MathWorks, Inc., January 1998. Ver. 2.

[4] A. Dandalis and V. K. Prasanna. Mapping homogeneous computations onto dynamically configurable coarse-grained architectures. *IEEE Symposium on Field-Programming Custom Computing Machines*, April 1998. (extended abstract).

[5] B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. Technical Report 280, MIT Media Lab Perceptual Computing, May 1994. http://sound.media.mit.edu/KEMAR.html.

[6] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki. Common-acoustical-pole and zero modeling of head-related transfer functions. *IEEE Transactions on Speech and Audio Processing*, 7(2):188–96, March 1999.

[7] D. Kistler and F. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, 91(3):1637–47, March 1992.

[8] S. Kung. A new identification and model reduction algorithm via singular value decompositions. *Conference Record of the Twelfth Asilomar Conference on Circuits, Systems and Computers*, pages 705–14, November 1978.

[9] C. Kyriakakis. Fundamental and technological limitations of immersive audio systems. In *Proceedings of the IEEE*, volume 86, pages 941–51, USA, May 1998. IEEE.

[10] Lord Rayleigh (J. W. Strutt). On our perception of sound direction. *Phil. Mag.*, 13:214–232, 1907.

[11] J. Mackenzie, J. Huopaniemi, V. Valimaki, and I. Kale. Low-order modeling of head-related transfer functions using balanced model truncation. *IEEE Signal Processing Letters*, 4(2):39–41, February 1997.

[12] The MathWorks, Inc. *Control System Toolbox User's Guide*. The MathWorks, Inc., January 1999. Fourth Printing.