AN EFFICIENT HRTF MODEL FOR 3-D SOUND

C. Phillip Brown

Department of Electrical Engineering
University of Maryland
College Park, MD 20742
email: cpbrown@isr.umd.edu

Richard O. Duda

Department of Electrical Engineering
San Jose State University
San Jose, CA 95192
email: rod@duda.org

ABSTRACT

A simple model is presented for synthesizing binaural sound from a monaural source. The model produces vertical as well as horizontal and externalization effects. The simplicity of the model permits efficient implementation, allowing for real-time multisource operation. Additionally, the parameters in the model can be adjusted to fit a particular individual's characteristics.

1. INTRODUCTION

Three-dimensional sound is of central importance for virtual-reality systems, and is becoming increasingly important for auditory displays and for computer-human interaction in general. Current methods are either limited in their capabilities or computationally expensive. We present a modeling approach that promises to be both effective and efficient.

It is well known that the physical effects of the diffraction of sound waves by the human torso, shoulders, head and pinnae modify the spectrum of the sound that reaches the ear drums [1]. These changes are captured by the Head-Related Transfer Function (HRTF), which not only varies in a complex way with azimuth, elevation, range, and frequency, but also varies significantly from person to person [7].

Psychoacoustic studies of sound localization cues have focused on the spectral behavior of the HRTF [3], [7]. The information in the HRTF is also contained in the temporal behavior of the equivalent Head-Related Impulse Response (HRIR). In real-time synthesis, if the sound source moves relative to the head, the HRIR must be modified accordingly. This is typically done by computationally expensive interpolation. To produce convincing elevation effects, the HRIR must be measured separately for each listener, which is inconvenient and limits applications. Finally, sounds synthesized in this fashion are usually poorly externalized unless a room model is added [5].

Our approach, which was inspired by the work of Genuit [6], replaces experimentally measured HRIR's by a simple signal processing model. The model has separate modules to account for azimuth, elevation and range. Its structure leads to an efficient real-time filter whose parameters can be adjusted to account for person-to-person differences. Although a review of HRTF data and models is beyond the scope of this paper, we arrived at this structure after studying both the characteristics of experimental data and other HRTF models [2]-[4], [6]-[8].

2. DESCRIPTION OF THE MODEL

2.1 Overview

The overall structure of our model is shown in Fig. 1. The components correspond to major structural parts of the body and the external environment. It can be viewed as a highly simplified representation of some very complex phenomena. However, our goal is not to faithfully simulate physical processes, but rather to exploit our knowledge about these physical processes to provide the simplest customizable system that is capable of producing strong impressions of all three spatial dimensions — azimuth, elevation and range.

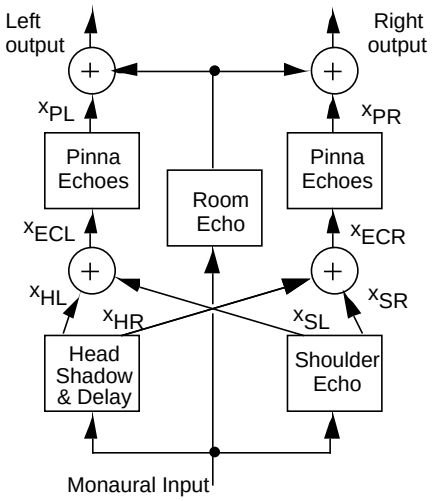

Figure 1: Components of the model.

The monaural input goes to a head model, a shoulder model and a room model. The outputs of the head and shoulder models are summed to give the pressures at the entrance of the ear canals as if the pinnae were absent. These signals are then processed by the pinna model to produce elevation effects. Finally, the outputs of a room model are added to provide range effects.

Each component of the model affects at least one of the three spatial dimensions. The head model, which depends only on azimuth, provides the well-known Interaural Time Difference (ITD) and Interaural Level Difference (ILD) cues. The shoulder model, which depends on both azimuth and elevation, provides

relatively weak elevation cues; while it is discussed further in [2], it is omitted from the remainder of this paper. The pinna model, which is weakly azimuth dependent and strongly elevation dependent, is the most novel part of the model. The room model introduces early reflections to provide externalization, and is noteworthy mostly for its extreme simplicity.

In the formulas that follow, we use interaural-polar coordinates to specify the location of a source relative to the head. Where elevation φ is measured from the horizontal plane in the usual vertical-polar system, the azimuth θ is measured from the median plane in the interaural-polar system. Thus, a surface of constant azimuth is a cone of essentially constant ITD.

2.2 The head model

Diffraction of the incident sound wave by the head leads to the sound being delayed and "shadowed" at the far ear. If the head is approximated by a sphere of radius *a*, Woodworth's formulas [1] provide an accurate estimate of the time delay. Let $T_L(\theta)$ be the difference between the time that the incident wave strikes the head and the time that it reaches the left ear, let $T_R(\theta)$ be the corresponding time difference for the right ear, and let c be the speed of sound. Then

$$T_L(\theta) = \frac{a + a\theta}{c} \quad \text{and} \quad T_R(\theta) = \frac{a - a\sin\theta}{c}$$

if $0° \le \theta \le 90°$. (If $-90° \le \theta \le 0°$, the expressions for $T_L(\theta)$ and $T_R(\theta)$ are interchanged.) With interaural-polar coordinates, these formulas are valid for any elevation angle.

The effects of head shadow are introduced by the following simple one-pole/one-zero transfer function:

$$H(s,\theta) = \frac{\alpha(\theta)\, s + \beta}{s + \beta}, \text{ where } \beta = \frac{2c}{a}.$$

Here the pole is fixed at $s = -\beta$, and the coefficient $\alpha(\theta)$, which varies from 0 to 2, shifts the position of the zero as the azimuth changes. The case $\alpha = 0$ results in maximum head shadow, and corresponds to sound arriving directly opposite the ear, while the case $\alpha = 2$ produces a 6-dB boost at high frequencies, and corresponds to sound directly incident on the ear. We note in passing that the ITD from Woodworth's formulas is frequency independent, but the head-shadow filter introduces the required additional low-frequency group delay at low frequencies [3]. For ears placed diagonally across the head, we use $\alpha_L(\theta) = 1 - \sin(\theta)$ for the left ear and $\alpha_R(\theta) = 1 + \sin(\theta)$ for the right ear.[1] This leads to the head shadow and delay module shown in Fig. 2.

[1] These formulas produce the maximum head shadow at the point opposite to the point of incidence. For an ideal sphere, the maximum head shadow occurs closer to ±150° from the point of incidence [1]. Furthermore, human ears are not across a diameter, but are set back about 10°, so that the far ear is more or less maximally shadowed when the near ear receives direct incidence. If desired, the formulas for $\alpha(\theta)$ can easily be refined to incorporate these facts [2].

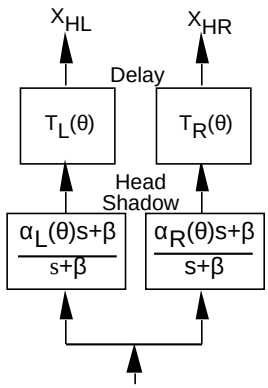

Figure 2: The head model

2.3 The pinna model

Carlile [3] divides pinna models into three classes: resonator, reflective, and diffractive. Because the physical interpretation of pinna elevation cues is somewhat controversial, we shall refer to the peaks and troughs that can be seen in the impulse response as "events." We account for these pinna events by the multipath model shown in Fig. 3. Here ρ_k is the amplitude and τ_k is the time delay associated with the k^{th} of n events

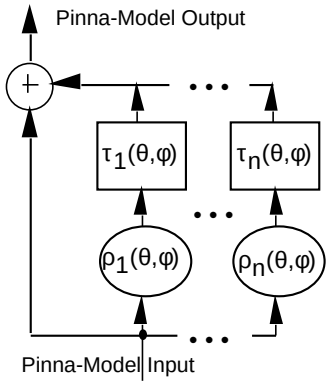

Figure 3: The pinna model

Informal listening tests revealed that there was little to be gained by using more than $n = 5$ events. In addition, it appeared to be adequate to use constant values for the amplitudes ρ_k. When we examined images of the HRIR data from three human subjects, we observed that the time delays could be well approximated by an equation of the form:

$$\tau_k(\theta,\varphi) = A_k \cos(\theta / 2)\sin(D_k(90° - \varphi)) + B_k$$

In the cases we examined, only D_k had to be adapted to an individual listener. (In the table below, D_{k1} is for two of the subjects, and D_{k2} is for the third.) By symmetry, the time delay for the left ear at azimuth θ should be the time delay for the right ear at azimuth $-\theta$. However, in the experiments that we describe ahead, we found that $\tau_k(\theta,\varphi)$ was basically an even

function of θ, and thus that the pinna delays for the two ears were roughly the same. The following table lists numerical values for the parameters of the pinna model; A_k and B_k are given in samples at a 44.1-kHz sampling rate, so that one sample is 22.7 μsec.

k	ρ_k	A_k	B_k	D_{k1}	D_{k2}
1	.5	1	2	1	.85
2	-1	5	4	.5	.35
3	.5	5	7	.5	.35
4	-.25	5	11	.5	.35
5	.25	5	13	.5	.35

Table 1: Pinna model coefficients

Shoulder reflection can be included by the same kind of model [2]. However, since informal listening tests indicated that the shoulder echo had very little effect on perceived elevation, we do not consider it further in this paper.

2.4 The room model

It is well known that simulated room reverberation produces an externalization effect [5]. The range module shown in Fig. 4 also accounts for the inverse distance effect. The monaural input is delayed by an amount τ_E and mixed with the outputs of the head model. The ratio of direct to reverberant energy is adjusted by the ratio of the channel gains K_L and K_R to the "echo gain" K_E. Unless the source is close to being directly ahead or directly behind, it usually seems externalized when τ_E was around 15 ms and the echo gain was 15 dB below the channel gains.

Clearly, this is an extremely elementary "room model." The single early reflection is not even processed by the HRIR model. One can substitute much more realistic room models, and can then use the HRIR model to localize the echoes and reverberation. However, even a model this simple can greatly improve the sense of externalization for headphone listening.

3. LISTENING TESTS

To evaluate the degree to which the model matched the measured HRIR's, listening tests were performed on three subjects at a fixed azimuth angle of +60° and with no shoulder-echo or range effects. These tests do not provide a good measure of the accuracy of the localization effects produced by the model, but they do provide a measure of how well the model substitutes for the measured HRIR's.

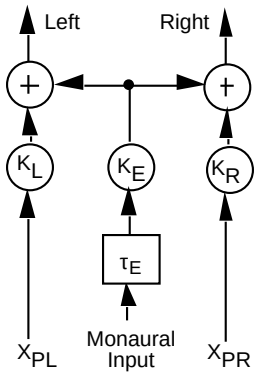

Figure 4: The room model

All listening tests were performed in identical fashion. The sound source was a 500-ms burst of Gaussian white noise that was filtered by either a measured or a modeled HRIR and played over Etymotic model ER-2 in-ear phones. Each subject was first presented with a noise burst filtered with his or her own experimentally measured HRIR for some randomly selected elevation. We call this the target sound. The system also had stored a set of 35 reference sounds formed by filtering the noise burst with the model HRIR at elevations in 5° increments from -85° to +90°. The subject was then asked to ignore any timbre differences and to match the perceived elevation of the target to one of the 35 reference sounds. Thus, the subject was asked to change the elevation angle for the model to best match the subject's own measured HRIR data. No restrictions were placed on the number of times the subject could listen to either the target sound or the reference sounds before making the final choice. The process was repeated 50 times.

To establish a performance baseline, in an initial task the 35 reference sounds actually came from using the subject's own HRIR, so that an exact match was always possible. The results for all three subjects are shown in Fig. 5. The data points for each subject are represented by o, + and *. The dashed line represents an ideal match. The solid line is a best fit straight line for the data points for all three subjects.

Several trends are observable from this data. The first trend was a slight tendency to place the sound somewhat "higher" than the correct elevation. This bias is revealed by the best-fit line being above the ideal line. The second trend was reduced accuracy near the elevation end-points and a tendency to compress the range. The subjects perceived the sound as "higher" when the elevation was near φ = -85° and "lower" near φ = +90°. The mean absolute error (mean deviation from ideal) for all three subjects was 12.0°.

The same subjects were then asked to perform the same task using the reference sounds produced by the model. The results for all three subjects are shown in Fig. 6. Again, a reduced accuracy near the extreme elevations (φ = -85° and φ = +90°) was apparent. Additionally, the overall accuracy was worse than the baseline test. The mean absolute error for all three subjects was 23.4° for the case of the model.

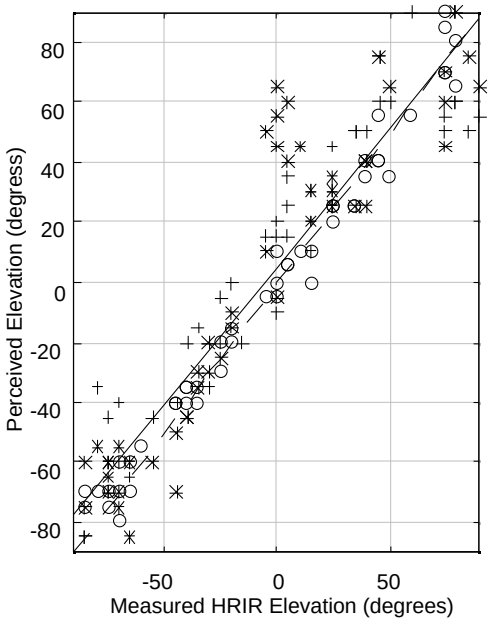

Figure 5: Measured HRIR localization accuracy

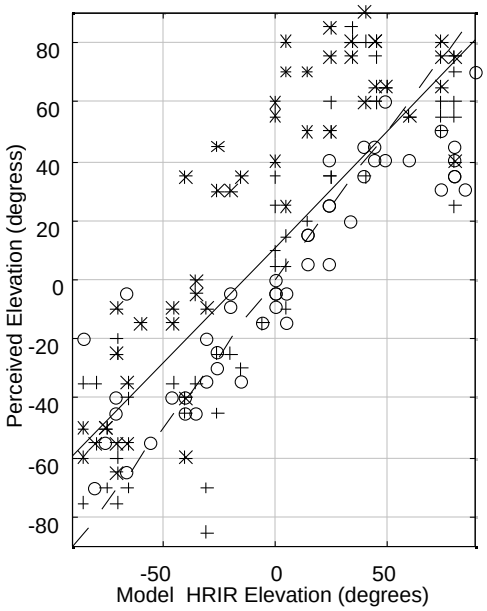

Figure 6: Model location accuracy

4. CONCLUSIONS

We have presented a simple but promising signal processing model for synthesizing binaural sound. The model has separate components for azimuth (head shadow and ITD), elevation (pinna and shoulder echoes) and range (room echo). The simplicity of the IIR head-shadow filter and FIR shoulder and pinna filters enables inexpensive real-time implementation.

No critical customization is required for azimuth or range effects. To obtain convincing elevation effects, the filter coefficients must be adjusted to match the individual listener. Fortunately, the number of variable parameters is small.

Much work remains to be done. Neither the measured HRTF nor the modeled HRTF are convincing at low elevations, and this needs to be better understood. The matching task used in the listening test needs to be replaced by a much better psychoacoustic localization test. A formal procedure for deriving filter coefficients from HRTF data needs to be created. Range estimation needs to be systematically investigated. However, we believe that physically-based models are a very promising approach for the real-time synthesis of 3-D sound.

Acknowledgments

This work stems from an M.S. thesis by the first author [2]. The material is based upon work supported by the National Science Foundation under Grant No. IRI-9214233. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are also indebted to Richard Lyon for his encouragement and advice.

REFERENCES

1. Blauert, J. , *Spatial Hearing (Revised Edition)*, MIT Press, Cambridge, MA, 1997.

2. Brown, C. P. “Modeling the Elevation Characteristics of the Head-Related Impulse Response,” M.S. thesis, Dept. of Elec. Engr., San Jose State Univ., May 1996.

3. Carlile, S., “The physical and psychophysical basis of sound localization,” in Carlile, S. (ed.), *Virtual Auditory Space: Generation and Applications*, 27-78, R. G. Landes, Austin, TX, 1996.

4. Duda, R. O. “Modeling head related transfer functions,” in *Proc. Twenty-Seventh Annual Asilomar Conference on Signals, Systems and Computers* (Asilomar CA, November 1993).

5. Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., and Wenzel, E. M. “On the externalization of auditory images,” *Presence*, 1: 251-257, Spring 1992.

6. Genuit, K., “Ein Modell zur Beschreibung von Aussen-ohrubertragungseigenschaften,” Ph.D. dissertation, Rheinisch-Westfalischen Technichen Hochschule Aachen, Aachen, Germany, December 1984.

7. Kistler, D. J., and Wightman, F. L. “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction,” *J. Acoust. Soc. Am.*, 91, March 1992, 1637-1647.

8. Shinn-Cunningham, B. and Kulkarni, A., “Recent developments in virtual auditory space,” in Carlile, S. (ed), *Virtual Auditory Space: Generation and Applications*, 185-243, R. G. Landes, Austin, TX, 1996.