Modeling Head-Related Transfer Functions Based on Pinna Anthropometry

Navarun Gupta, PhD
Electrical and Computer Engineering Department,
Florida International University
Miami, FL 33174

Armando Barreto, PhD
Electrical and Computer Engineering Department,
Florida International University
Miami, FL 33174

Maroof Choudhury
Electrical and Computer Engineering Department,
Florida International University
Miami, FL 33174

Abstract

The use of Head-Related Transfer Functions (HRTFs) in creating 3D sounds is gaining wide acceptance in multimedia applications. This paper presents a new method of modeling HRTFs based on the shape and size of the outer ear. Using signal processing tools, such as Prony's signal modeling method, an appropriate set of time delays and a resonant frequency were used to approximate the measured Head-Related Impulse Responses (HRIRs). HRIRs represent HRTFs in time domain. Statistical analysis was used to find out empirical equations describing how the reflections and resonances are determined by the shape and size of the pinna features obtained from 3D images of 15 experimental subjects modeled in the project. These equations were used to yield "Model HRTFs" that can create elevation effects.

Listening tests conducted on 10 subjects showed that these model HRTFs were 5% more effective than generic HRTFs in the frontal plane. This model is a simple, yet effective method of creating customizable HRTFs. It reduces the computational and storage demands, while preserving a sufficient number of perceptually relevant spectral cues.

Keywords

HRTF, HRIR, Pinna, Prony's Method, Spatialization

Introduction

3D sound systems are gaining increasing relevance due to their continuous expansion applications in the entertainment industry, computer gaming, and also in the broader fields of Virtual Reality (VR) and Human-Computer Interaction (HCI). HRTFs are becoming a popular means of recreating binaural sounds because of significant advances in signal processing hardware. Once limited by speed and costs, now digital signal processors are economical and fast.

As the sound waves leave the source, they are reflected and diffracted by the listener's torso, shoulders, head, and outer ears, before entering the auditory canal. This transformation in sound is captured by the HRTFs and it is unique for every direction around the listener. The HRTFs are also unique to the listener because of wide variations in the shapes and size of head and outer ears.

This implies that the HRTF pairs should be defined individually for the prospective listener, since the physical elements cited as their defining factors (torso, head and outer ears) are clearly individual and must match the localization clues that each individual learns to extract form real world sounds. Currently, some sound spatialization systems will make use of HRTFs that are empirically measured for each prospective user. These "custom" HRTFs are "anthropometrically correct" for each user, but the equipment, facilities and expertise required to obtain these "measured HRTF pairs", constrain their application to high-end, purpose-specific sound spatialization systems only. The most common alternative is the use of "generic" HRTFs, originally recorded from a manikin shaped according to "average" anthropometric data, expected to be representative of a pool of prospective listeners. While this solution has proved practical and useful for consumer-grade sound spatialization applications, it should be recognized that it is fundamentally limited by its disregard for the individual aspect of the anthropomorphic nature of the signal processing challenge being addressed.

This papers reports on our work to advance an alternative approach to sound spatialization, based on the postulation of anthropometrically-related "structural models" [Brown and Duda, 1998] that will transform a single-channel audio signal into a Left/Right binaural spatialized pair, according to the sound source simulation. Specifically, the work reported here proposes linkages between the parameters of the HRTF model and key anthropometric features of the intended listener, so that the model, and consequently the resulting HRTF's are easily "customizable" according to a small set of anthropometric measurements.

Measurement and Implementation of HRTFs

A speaker is placed at known relative positions with respect to the subject for whom the HRTFs are being determined, and a known, broad-band audio signal is used as excitation. In our laboratory, we use the Ausim3D's HeadZap HRTF Measurement System [AuSIM, Inc., 2000]. This system measures a 256-point impulse response for both the left and the right ear using a sampling frequency of 96 KHz. Golay codes are used to generate a broad-spectrum stimulus signal delivered through a Bose Acoustimass speaker. The response is measured using miniature blocked meatus microphones placed at the entrance to the ear canal on each side of the head. Under control of the system, the excitation sound is issued and both responses (left and right ear) are captured. Since the Golay code sequences played are meant to represent a broad-band excitation equivalent to an impulse, the sequences captured in each ear are the impulse responses corresponding to the HRTFs. Therefore these responses are called Head-Related Impulse Responses (HRIRs). The system provides these measured HRIRs as a pair of 256-point minimum-phase vectors, and an additional delay value that represents the Interaural Time Difference (ITD), i.e., the additional delay observed before the onset of the response collected from the ear that is farthest from the speaker position. In addition to the longer onset delay of the response from the "far" or "contralateral" ear (with respect to the sound source), this response will typically be smaller in amplitude than the response collected in the "near" or "ipsilateral" ear. The difference in amplitude between HRIRs in a pair is referred to as the Interaural Intensity Difference (IID).

The measurement of HRIRs is carried out in 12 azimuths $\theta = \{0^o, 30^o, 60^o, 90^o, 120^o, 150^o, 180^o, -150^o, -120^o, -90^o, -60^o, -30^o\}$, and 6 elevations $\phi = \{-36^o, -18^o, 0^o, 18^o, 36^o, 54^o\}$. The left (L) and right (R) HRIRs collected for a source location at azimuth θ and elevation ϕ will be symbolized by $h_{L,\theta,\phi}$ and $h_{R,\theta,\phi}$, respectively. The corresponding HRTFs would be $H_{L,\theta,\phi}$ and $H_{R,\theta,\phi}$. The creation of a spatialized binaural sound (left and right channels) involves convolving the single-channel digital sound to be spatialized, s(n), with the HRIR pair corresponding to the azimuth and elevation of the intended virtual source location:

$$y_{L,\theta,\phi}(n) = \sum_{k=-\infty}^{\infty} h_{L,\theta,\phi}(k) \cdot s(n-k) \quad , \text{ and } \quad y_{R,\theta,\phi}(n) = \sum_{k=-\infty}^{\infty} h_{R,\theta,\phi}(k) \cdot s(n-k) \tag{1}$$

Structural Models

Structural models of HRTF are based on the premise that each anthropometric feature of the listener affects the HRTF in a way that can be described mathematically. Because such a model has its origin in the physical characteristics of the entities involved in the phenomenon, it should be possible to derive the value of its parameters (for a given source location), from the sizes of those entities, i.e., the anthropometric features of the intended listener. Proper identification of such parameters, and adequate association of their numerical values with the anthropometric features of the intended listener may provide a mechanism to interactively adjust a generic base model to the specific characteristics of an individual. One of the most practical models has been proposed by Brown and Duda [Brown and Duda, 1998]. Their model is illustrated in Figure 1:

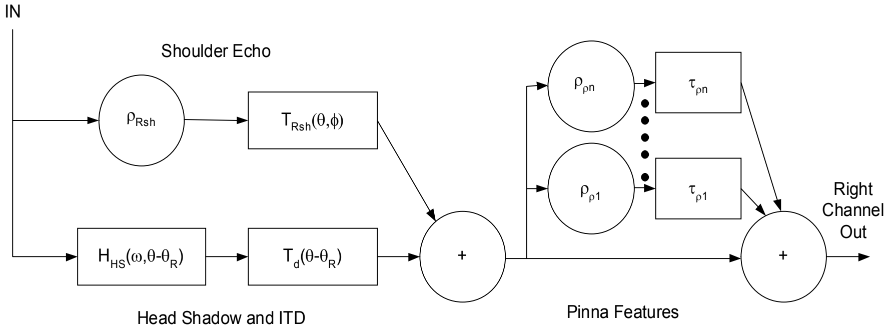

Figure1: Right Channel half for Brown & Duda's Structural HRTF Model. The model comprises a symmetric left half (not shown). From [Brown and Duda, 1998]

Customizable Pinna Model

The definition and anthropometric characterization of the PINNA MODEL has remained an open question, so far, and it is the objective of our work. Carlile [Carlile, 1996] divides pinna models according to the main phenomenon that they address: Resonating, diffractive and reflective. From these, reflective models have attracted the most attention in the literature.

The intent of our work is to define a functional pinna sub-model that has anthropometric plausibility and then associate its parameters to anthropometric features of the listener's pinna.

Taking into account the information available about the existence of a resonant effect implemented by the ear's concha [Shaw and Teranishi, 1967], and according to the reflective pinna models discussed previously, we propose that the pinna may, in turn, be modeled as the series connection of an equivalent second-order resonator and a series of characteristic echoes, representing the delayed and attenuated secondary paths taken by the incoming sound, in addition to a "direct path". A block diagram representation of this model is shown in Figure 2.

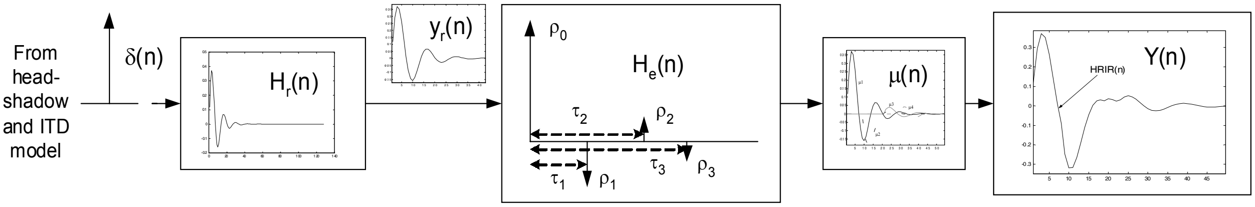

Figure 2. Block Diagram of the proposed Pinna Model.

According to this model, the impulse input will project the underdamped oscillatory impulse response of the resonator, $H_r(n)$, as the intermediate output, $y_r(n)$, which will then be convolved with the impulse response of the block of echoes, $H_e(n)$, to yield the overall output, Y(n). Since $H_e(n)$ is expected to consist only of a few non-zero values, the output sequence can be interpreted as the superposition of several copies of the resonant response, $y_r(n)$, re-sized and delayed according to the several non-zero values of $H_e(n)$, as shown in the figure. The re-sizing factors and latencies for each "echo" have been assigned the variable names ρ_k and τ_k, where k is the number of the echo ($\tau_0 = 0$). In turn, the second-order resonant block can be defined by the frequency, f, and damping factor, σ, for its impulse response.

Thus, the instantiation of this proposed model will require the identification of f, σ, and the several ρ and τ values, to characterize the parameters of the model that successfully approximates an HRIR collected for a given azimuth and elevation, through the Y(n) provided by the pinna model. The main challenge in this operation is the fact that the several replicas of the damped oscillation are irreversibly mixed together, partially overlapping in time, in the measured HRIR. This problem was addressed by the sequential application of Prony's modeling algorithm [Osborne and Smyth, 1995] to partial segments of the response. Prony's method approximates a given signal μ(t) as the superposition of p damped sinusoidals:

$$\mu(t) = \sum_{j=1}^{p} \rho_j e^{(\sigma_j t)} Sin(2\pi f_j t + \xi_j) \qquad (2)$$

Measurement of Anthropometric Features

Key anthropometric features of the ears of the 15 experimental subjects in the study (same 15 subjects for whom the HRIRs were empirically measured with the Ausim 3D system) were captured by means of digital photography (including a distance reference), and laser 3-D scanning, using a Polhemus FastScan handheld scanner. Figure 6 shows a sample digital photograph, a sample 3-D reconstruction from a 3-D scanned file, and a schematic drawing identifying some of the key anthropometric features estimated for each of the 30 ears involved in this study. The features measured are: Ear length (E_L), Ear width (E_W), Concha width (C_W), Concha height (C_H), Helix length (H_L), Concha area (C_A), Concha volume (C_V) and Concha depth (C_D).

Association between Model Parameters and Anthropometric Measurements

Following the procedure described in the two preceding sections two independent sets of data were available for each pinna of each on of the 15 subjects in the study:

Estimated Model Parameters:

$r_{0\phi}$, $\alpha_{0\,\phi}$, $\rho_{0\,\phi}$, $\rho_{1\,\phi}$, $\tau_{1\,\phi}$, $\rho_{2\,\phi}$, $\tau_{2\,\phi}$, $\rho_{3\,\phi}$ and $\tau_{3\,\phi}$, for $\phi = -36^\circ, -18^\circ, 0^\circ, 18^\circ, 36^\circ$, and 54°
(Note, here r_0 and α_0 are the magnitude and angle of the poles of the resonator, which define the resonator response $\mu_0(n)$, in terms of its frequency f_0 and its damping factor σ_0.

Measured Anthropometric Features:
$E_{L}, E_{W}, C_{H}, C_{W}, C_{A}, C_{V}, C_{D}$ and H_L

Under the assumption that the model parameters depend of the anthropometric features, a general dependency equation may be set, for each model parameter. For example, for the amplitude of the first reflection in the pinna model, ρ_0, at $\phi = 54^{o}$, the following equation may be set up:

$$\rho_{0\ \phi=54} = KEL(E_L) + KEW(E_W) + KCH(C_H) + KCW(C_W) + KCA(C_A) + KCV(C_V) + KCD(C_D) + KHL(H_L) + B \quad (3)$$

Coalescing the data from both ears, at the same elevation, (under the assumption of symmetry), 30 equations like the one above can be set up, for each model parameter, at each elevation. Each group of 30 equations can then be analyzed through multiple regression to estimate the values of the constants (KEL, KEW, …KHL, B). The multiple regression analysis was carried out using the Statistical Package for the Social Sciences (SPSS).

Testing the Model and Results

Using the predictive equations found above, for each subject tested, a Model HRTF was created. Ultimately, the efficiency of the modeled sequences obtained by predicting the model parameters from the anthropometric measurements of the subjects was gauged in listening tests. In these tests, white noise bursts were spatialized using the modeled sequences, M(n), that had been obtained based on the ear measurements of the subject under test, for the six elevations under study. The order in which these elevations where used for the spatialization was randomized. Each elevation was simulated four times (i.e., there were 24 trials for each side of the head.) In each trial the subject would listen to each spatialized sound and then use a graphic user interface to indicate the perceived elevation. Since the spatialization was performed to emulate six specific locations, the absolute value of the angular difference between the perceived elevation and the emulated one would be considered as the elevation error for the trial. The subjects listened to the original, modeled and generic HRTFs.

Figure 3 illustrates the average angular error (across all 10 subjects) experienced in the perception of the different emulated elevations for Original, Generic (B&K) and Model HRIRs. The global average error (across all subjects and all elevations) with the original HRTFs, was 23.7°. The corresponding global average error with modeled HRIRs was 29.9°. Finally, the global average error when the subjects used the generic HRIRs, collected from the B&K manikin, was 31.4°.

It should be noted, however, that near the horizontal plane (e.g., between $\phi = -18^{o}$ and $\phi = 36^{o}$), the performance of the modeled HRIRs was close to or better than, that of the individually measured HRIRs.

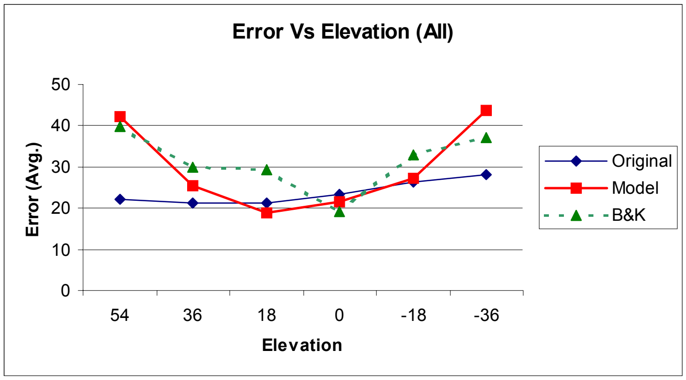

Figure 3. Localization performance using 3 different types of HRIRs.

Conclusions

This paper has presented a proposed functional model of the pinna, to be used as the output block in a structural HRTF model. The definition of the model, containing a second order resonance and non-recursive filter representing a number of "echoes" at varied latencies, was introduced and justified in terms of the sequential Prony deconvolution of the damped oscillatory response of the resonator, from the measured HRIRs of 15 subjects, at 6 different elevations. The listening tests confirmed that the modeled HRIRs obtained from the proposed model helped the volunteers perform better (29.9^o average error) than "generic" HRIRs from a manikin (31.4^o average error), in localizing broad-band sounds spatialized to 6 different elevations on the frontal plane. However, the performance of the volunteers was better with their own individual (measured) HRIRs (23.7^o average error).

Although this study resorted to the sue of a relatively expensive 3-D laser scanner and specialized software to determine some of the anthropometric features of our subjects, which is a prerequisite to the use of the predictive equations developed in this research, it is likely that empirical relationships can be found to obtain these feature values from two-dimensional high-resolution photographs (commonly available) and a few direct physical measurements in the subject.

Acknowledgements
This research was supported by NSF grants EIA-9906600, HRD-0317692, and IIS-0308155.

References:

AuSIM, Inc., "HeadZap: AuSIM3D HRTF Measurement System Manual". AuSIM, Inc., 4962 El Camino Real, Suite 101, Los Altos, CA 94022, 2000.

Brown, C. P. and Duda, R. O., "A Structural Model for Binaural Sound Synthesis", IEEE Trans. Speech and Audio Processing, Vol. 6, No. 5, pp.476-488, September 1998.

Carlile, S., "The Physical Basis and Psychophysical Basis of Sound Localization", in Virtual Auditory Space: Generation and Applications, S. Carlile, Ed., pp. 27-28, R. G. Landes, Austin TX, 1996.

Mills, A. W., "Auditory Localization," in Foundations of Modern Auditory Theory, Vol. II (J, V. Tobias, Ed.), pp. 303-348, Academic Press, New York, 1972.

Osborne, M. R., and Smyth, G. K., "A modified Prony algorithm for fitting sums of exponential functions", SIAM J. Sci. Statist. Comput., vol. 16, pp. 119-138, 1995.

Shaw, E. A. G., Teranishi, R., "Sound pressure generated in an external-ear replica and real human ears by a nearby point source", J. Acoust. Soc. Am., vol. 44, No. 1, pp. 240-249, 1967.

Watkins, A. J., "Psychoacoustical Aspects of Synthesized Vertical Locale Cues", J. Acoust. Soc. Am., vol. 63, no. 4, pp. 1152-1165, 1978.