Accelerated Method for the Reduced-Parameter Modeling of Head-Related Transfer Functions for Customizable Spatial Audio

KENNETH JOHN FALLER II[1], ARMANDO BARRETO[1], NAVARUN GUPTA[2] and NAPHTALI RISHE[3]

Electrical and Computer Engineering Department[1] and School of Computing and Information Science [3]
Florida International University
Miami, FL 33174
USA
kfall001@fiu.edu http://dsplab.eng.fiu.edu/

Department of Electrical and Computer Enginering[2]
University of Bridgeport
Bridgeport, CT 06604
USA

Abstract: - Many spatial audio ("3D sound") systems are based on the use of Head-Related Transfer Functions (HRTFs). Since the measurement of these HRTFs for each prospective listener is impractical in many applications, developers frequently use "generic" HRTFs (e.g., from a mannequin), sacrificing the superior spatialization that could be provided by "individual" HRTFs. This paper presents an improved method for decomposing the impulse response of measured HRTFs, known as Head-Related Impulse Responses (HRIRs), into multiple delayed and scaled damped sinusoids, as a means to obtain the parameters that will instantiate a general model in order to produce the same impulse response as the original HRTF. This is the first step in developing alternative functional models of HRTFs that would contain only a few parameters related to anatomical features of the intended listener, which could be estimated for each user, i.e., "customizable" HRTF models. The new decomposition algorithm is based on the use of higher-order (higher than second order) Steiglitz-McBride functional approximations supplemented by frequency-domain selection of the most appropriate damped sinusoid to represent each of the components of the original, measured HRIR under analysis.

Key-Words: - Head-Related Impulse Responses (HRIR), Prony modeling method, Steiglitz-McBride iterative approximation method, customizable spatial audio.

1 Introduction

The proliferation of computer applications in many aspects of our daily lives has brought about an expanded field of applications for digital audio. In addition to its high fidelity, digital audio may be added with a sense of sound source location, managed independently as one more attribute of the digital sound. This constitutes the basis of three dimensional (3-D) spatial audio and its applications have become increasingly popular in scientific, and commercial systems. The spatial audio effect can be achieved through different methods. The multi-channel approach requires that speakers be physically positioned around the listener (e.g., Dolby® 5.1 array). However, this approach is not portable and, therefore, it is impractical for several applications. The two-channel approach applies digital signal processing (DSP) techniques to an original digital sound to create a pair of signals (left channel and right channel) that can be delivered to a listener through headphones. In consequence, this approach is much more portable. The DSP-based creation of the left and right channel signals (i.e., a "binaural sound"), involves the use of special filters, characterized by their impulse responses, which are known as Head-Related Impulse Responses (HRIRs). The transfer functions of these filters are known as Head-Related Transfer Functions (HRTFs), and are meant to emulate the effect of anatomical (torso, head, external ear, etc.) and environmental (walls, floor, etc.) factors which cause modification of a sound as it propagates from its source to each of the listener's eardrums. Therefore, every position and each ear will have a specific HRIR. Convolving a sound signal with the two HRIRs corresponding to a specific source position results in a binaural sound (left channel, right channel) that, when played to a listener through stereo headphones will cause a perception similar to that of a sound emanating from the source location in question, specified by azimuth, elevation and distance (Fig. 1).

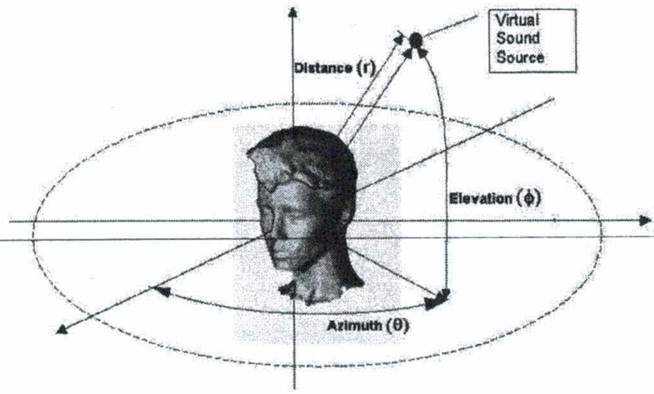

Fig. 1. Diagram of spherical coordinate system.

Since each person's head and torso are different, the creation of highly convincing binaural sounds requires the convolution of digital sounds with the pair of HRIRs estimated from each individual listener. However, the determination of the "individual" HRIR pairs corresponding to varied positions around a specific subject requires the use of specialized and expensive equipment and the involvement of trained personnel, which makes it unaffordable to most users of spatial audio systems. Instead, many applications of spatial audio systems make use of "generic" HRIR pairs obtained from a mannequin of "average anatomical dimensions" (e.g., MIT's measurements of a KEMAR Dummy-Head Microphone) or using a limited number of subjects to represent the general population (e.g., the CIPIC Database). This type of "generic" HRIRs provides an approximate sense of source locations in many users, but does not have as high spatialization fidelity as individual HRIRs [1].

Our goal is the development of "customizable" HRIRs obtained from a generic dynamic model that could be instantiated differently for each particular listener, by taking into account the relevant physical measurements of the intended listener, in order to still provide a high-fidelity spatialization. Unfortunately, the currently used representation of HRIRs as long (e.g., 128, 256, 512) collections of values obtained as the response to impulse-equivalent functions, such as Golay codes, cannot be altered in any simple way that would factor in the geometrical characteristics of the intended listener. Therefore, we believe that the first step towards customizable HRTFs is the substitution of their current representation in terms of large sets of HRIR sample values with an equivalent functional model requiring the instantiation of a much smaller number of parameters related to the geometry of each intended listener.

2 Methodology

2.1 Reduced-Parameter Pinna Model

Brown and Duda [2] have proposed that a "structural" model for binaural sound synthesis should "cascade" the effects (e.g., diffraction, inter-aural delay, etc.) of the listener's head with the local monaural effects of the geometry of the pinna or outer ear. Previous work by Algazi et al. [3] has already yielded a functional model for the listener's head that can be customized according to 3 simple anatomical measurements. Therefore, the objective of our work is to establish a reduced-parameter pinna model which could be instantiated on the basis of geometrical measurements from the intended listener.

We have previously proposed a pinna model in which the transformation of sound traveling to the eardrum takes place by superposition of a number of reflections of the incoming sound in the ear, which are also affected by the effect of the pinna cavities, such as the concha, acting as resonators [4]. The basic formulation of this model is shown as a block diagram in Figure 2.

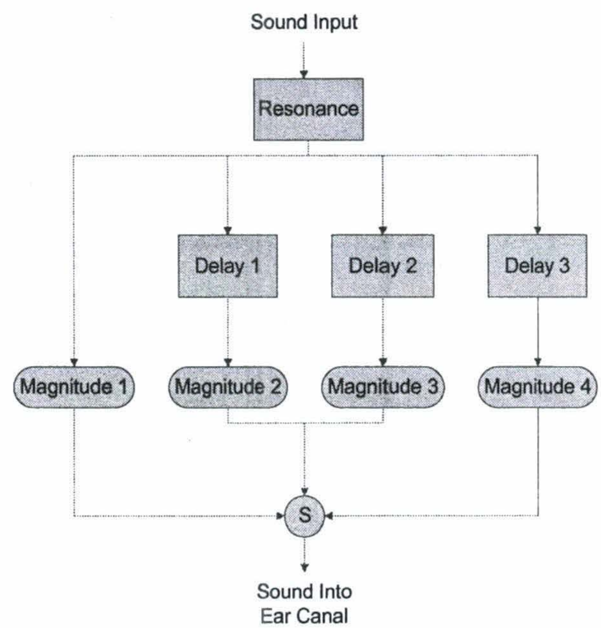

Fig. 2. Block diagram of pinna model (for 4 components).

In this model the parallel paths represent the multiple bounces of the incoming sound wave on the geometrical structures that constitute the pinna. Since the different trajectories represented have, in effect, different lengths they are affected by different delays, τ_i. Similarly, the loss of energy in each reflection of sound is modeled by a magnitude factor, ρ_i, in each of the paths of the block diagram. If the resonator block in the diagram can be represented by two parameters, such as the angle and the radius of its poles in the z-plane, the instantiation of a model such as the one shown in Figure 2 would require the definition of only 9 parameters, all of which can reasonably be expected to relate to geometrical measurements of the outer ear of the subject. This approach, therefore, satisfies the requirements of a reduced-parameter pinna model, which could be "cascaded" with Algazi's functional head model to represent a given HRTF.

Since, however, the parameters of our model cannot be measured directly, our first goal is to convert HRIR sequences measured by specialized equipment (e.g., AuSIM's HeadZap HRTF measurement system). Once the values for our model are known for multiple source positions and for a large enough number of experimental subjects, from whom anatomical measurements are also available, empirical rules will be developed to assign parameter values from anatomical measurements. Such rules could then be used to assign a custom set of parameters to the model for spatial audio generation for any subject if his/her geometric measurements are known. This paper describes a new approach for the definition of the model parameters from the decomposition of a measured HRIR for one of the ears of a given subject, i.e., the conversion of a "traditional" measured HRIR sequence of values to a smaller set of model parameters.

2.2 Previous approaches for the determination of pinna model parameters from measured HRIRs

According to the model shown in Figure 2 a measured HRIR sequence will be conformed by the superposition of several damped sinusoidals (which are the impulse response of the resonator), appearing scaled by a magnitude factor ρ_i and delayed by a latency τ_i. ($\tau_1 = 0$). Therefore the magnitude factors and delays needed for the model will become apparent if the original measured HRIR sequence is decomposed into a number of scaled (ρ_i) and delayed (τ_i) damped sinusoids.

Previously [4-6] this process of HRIR decomposition into damped sinusoids has been attempted by sequential application of second-order Prony or Steiglitz-McBride (STMCB) signal modeling algorithms to consecutive windows defined on the measured HRIR sequence. The aim of that sequential process was to always restrict the analysis to a partial window of data where only one damped sinusoidal (second-order approximation) is expected to be present. So, an initial window is defined from the beginning of the HRIR sequence to a point where the second damped sinusoid is estimated to start, i.e., τ_2. The amplitude of this first estimated sinusoid is considered to be the value of the magnitude factor ρ_1. The sinusoid estimated for the initial interval (F1) is then extrapolated to the end of the HRIR sequence and it is subtracted from it, to remove the influence of this first damped sinusoid from the rest of the measured HRIR. At this point the residue obtained is analyzed in the same manner as the original HRIR, except that the origin of analysis is re-established at time τ_2. The next stage of decomposition will only use the window of data between τ_2 and the point in which the onset of the third damped sinusoid component is estimated to occur, τ_3. The damped sinusoid estimated in the second stage of decomposition (F2) will also be extrapolated and subtracted from the complete extent of the HRIR remnant still being analyzed. The amplitude of this second estimate will be considered to be ρ_2. To begin the third stage of decomposition the origin of analysis will be re-established at τ_3, and the process can be repeated through as many decomposition stages as damped sinusoids are sought. An example of results from the process is shown in Figure 3.

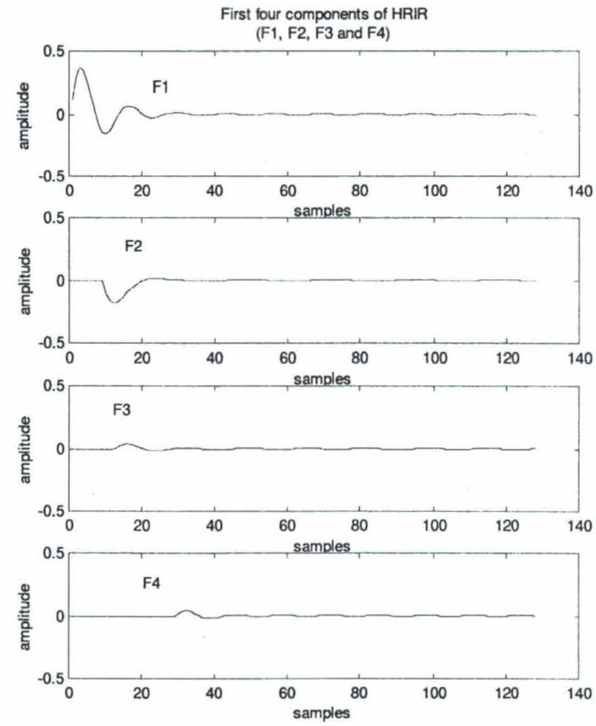

Fig. 3. Four damped sinusoidal components obtained from a measured HRIR.

The goal of the process is to obtain a set of damped sinusoids which, when added, form a "reconstructed" HRIR which is a good approximation of the original, measured HRIR sequence. The goodness of fit of the reconstructed HRIR is assessed by means of Equations (1) and (2), where MS means mean square value:

$$\text{Error} = \text{Original HRIR} - \text{Reconstructed HRIR}, \qquad (1)$$

$$\text{Fit} = [1 - \{\text{MS(Error)}/\text{MS(Original HRIR)}\}]. \qquad (2)$$

It has been observed that proper recognition of the boundaries for the windows to be analyzed (τ_2, τ_3, etc.) was critical to the achievement of high fit values. Originally, these break points were determined in each stage of decomposition by tentatively widening the window of analysis, finding a tentative damped sinusoid and calculating the mean square error (MSE) between the tentative approximation and the values of the HRIR under analysis, within the tentative window. Typically the MSE value found would decrease as the window was widened, until it would reach the onset of the next damped sinusoidal present, where the MSE would spike. It was later found [5, 6] that an exhaustive search which tried all probable widths (range of 2 to 10 sampling intervals of 10.4 μs each) for all the sequential windows and selected the parameters from the combination of widths that resulted in the larger overall fit (between the complete reconstructed HRIR and the original measured HRIR) provided better average fits for a database of 14 subjects including a total of 2016 HRIRs (2 ears, 72 source positions per subject). However, the exhaustive search approach is extremely computationally intensive, even with just the 5 windows processed in those studies. In fact, the tree-diagram needed to track all possible width combinations of 5 sequential windows has 9x9x9x9x9=59,049 leaf nodes and the addition of any subsequent windows with this approach will multiply the number of leaf nodes by 9, per additional window. To truly select the best of all possible alternatives, all the branches of the tree need to be explored and the reconstructed HRIR defined at each leaf node compared with the measured HRIR to assess its fit. It became clear that increasing the number of windows of analysis (which may be necessary to model late components in the HRIRs) would be impractical using the exhaustive search method. This has prompted us to develop a new, faster method of HRIR decomposition into sequential damped sinusoids.

2.3 Pole Approximation of Damped Sinusoids

The goal of this new method is to avoid having to pre-set the width of each sequential window of data analyzed in each subsequent stage of the decomposition process. The need to isolate small windows of data was connected to the assumption that such windows could be defined so that they would only contain a single damped sinusoid and not a superposition of several of them. Under that premise the Prony or STMCB second order algorithms were applied in each window, seeking to approximate a single damped sinusoid. In general, a single damped sinusoidal component sequence will be represented by a conjugate pair of poles within the unit circle and a zero at the origin of the Z-plane (Figure 4) [7]. Hence, a damped sinusoid in the Z-domain can be described with the following general equation:

$$X(z) = \frac{k \cdot z}{(z - p_1)(z - p_2)} \qquad (3)$$

where k is a scalar and p1 and p2 are complex poles. According to Equation 3, if the scalar k and the poles are known then, using the inverse z-transform, it is possible to find the time domain representation of a damped sinusoid.

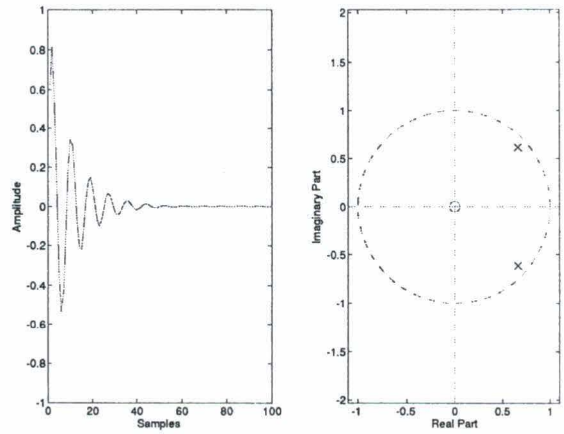

Fig. 4: Time domain and Zero-Pole plot of a single damped sinusoidal.

Based on these considerations, the new approach to the decomposition of HRIRs into scaled and delayed damped sinusoids will use the complete remnant of the HRIR available for processing at each decomposition stage (instead of a bounded window), searching for multiple damped sinusoids by application of a higher-order STMCB approximation for the whole remnant. The method will then isolate individual damped sinusoids by their complex conjugate pole signatures in the Z-domain, pursuing as many alternative outcomes as complex pairs can be identified for the specific decomposition stage in question. This also results in a tree-search approach. However, the branching factor of this search tree starts at the amount of damped sinusoids being extracted from the whole HRIR but decreases by one in every subsequent stage of the decomposition, which makes the number of leaf nodes much smaller than for the previous algorithm.

For example, if 5 damped sinusoids will be extracted, only 5x4x3x2x1 = 5! = 120 leaf nodes will exist.

The details of the process are explained with a simulated example in the paragraphs below.

The example addresses the decomposition of a synthetic signal created by summing three delayed damped sinusoids using this new "pole approximation" method. The sinusoids where created using the following equation:

$$x_i(n) = e^{d_i * n} \cdot \sin\left(\omega \cdot \frac{180}{\pi} \cdot n\right) \qquad (4)$$

where N is the length of the signal, n = 1,…,N, d_i is the negative damping factor and ω is the frequency. Once the three sinusoids (x1, x2 and x3) are created, the desired delays (τ_2 and τ_3) are applied to the last two sinusoids respectively, resulting in x2s and x3s. Finally, the sinusoids are then summed point-to-point to produce the final signal (x). In this example N=100, τ_2=3, τ_3=6, ω=3, d_1=-0.1, d_2=-0.125 and d_3=-0.15. The three signals (x1, x2s and x3s) and the resulting signal (x) are shown in Figure 5.

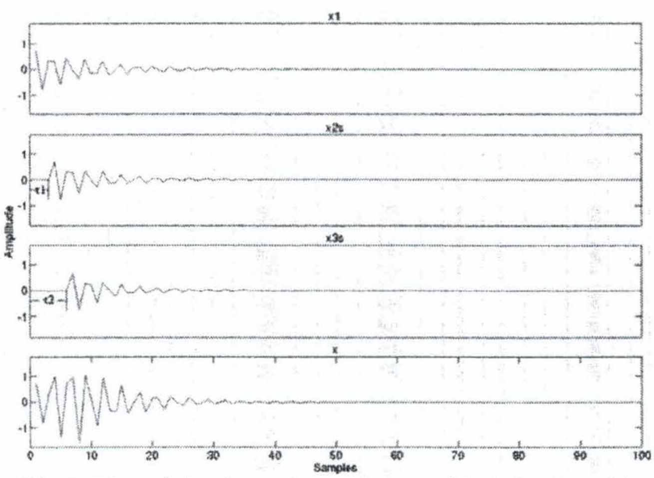

Fig.5: Plot of the three damped sinusoids (x1, x2s with delay τ_2 and x3s with delay τ_3) and the sum of them (x).

In this example the goal is to decompose x into three damped sinusoids. Therefore, the process starts by applying a sixth-order STMCB approximation process to the complete x sequence. The results from the sixth-order STMCB approximation will have the pole structure shown in Figure 6. As seen in Figure 6, x will result in 2 conjugate pairs. These pairs will be used to compute two separate impulse responses. One of which should approximate x1 accurately.

The damped sinusoidal impulse responses associated with the conjugate pole pairs shown in Figure 6 will be investigated as candidates to represent the first sinusoidal present in x (i.e., there will be up to three branches at the initial node of this search tree, if all the poles were complex). The investigation of each of these alternatives involves its subtraction from x to define a residue sequence, as shown in Figure 7, which will then be thresholded, according to Equation 5.

$$THR = \max(|signal|) * 0.25 \qquad (5)$$

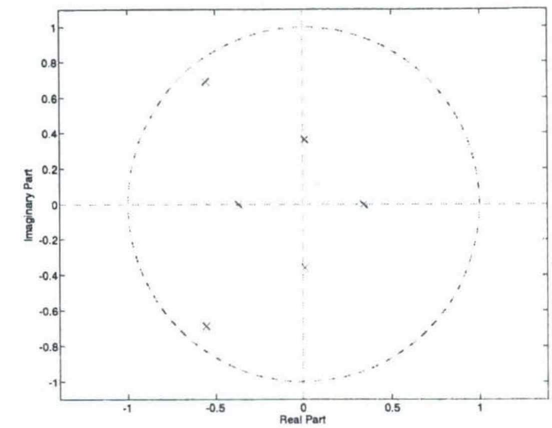

Fig.6: Poles obtained from the sixth-order STMCB approximation of the complete sequence x.

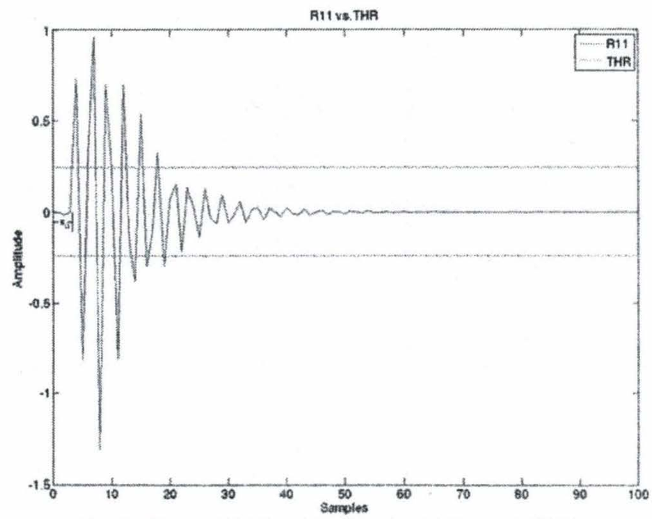

Fig.7: Plot of R11 with threshold lines (THR).

The time at which the residual surpasses this threshold will be considered the onset of the next damped sinusoidal, i.e., the estimate of τ_2. As in the previous method, the decomposition process will continue on to a second stage after re-establishing the origin of analysis at the estimated τ_2. The assumption made in every subsequent decomposition stage is that there should be one less damped sinusoidal present in the new remnant (since one has just been removed in the previous stage). As such, a fourth-order STMCB approximation will be applied in the second decomposition stage, yielding 4 poles, which will then be used to synthesize up to two candidates for the second damped sinusoid extracted from x. The same pattern of steps will be applied through all subsequent stages of the decomposition, until the stage in which a second-order STMCB approximation will be applied to the last remnant to identify the last damped sinusoid.

After M stages of decomposition there will be M! leaf nodes in the search tree, each representing a set of M delayed and scaled damped sinusoids that, when added together, form candidate approximations to the original signal x. The fit of each of those M! candidate approximations with respect to x will be evaluated (Equations 1 and 2) and the candidate with the highest fit will be selected as the final decomposition of x. In our example, the winning candidate approximation had a 99.99% fit with the original x, and the individual damped sinusoids obtained through each stage of decomposition also matched x1, x2s and x3s very closely.

3 Results from HRIR Decomposition

The method described in section 2.3 was applied to the decomposition of 14 actual HRIRs, recorded from 14 subjects using the AuSIM HeadZap system at Florida International University. The goal in each case was to obtain M = 5 damped sinusoidal components. Therefore, the order of the first STMCB approximation process was 10. Otherwise, the procedure was identical as the one explained for the decomposition of the synthetic sequence x, in section 2.3. The final fit achieved by the new "pole approximation" method, and the execution time required to complete the decomposition, for each HRIR, were recorded. The same group of HRIRs was also decomposed using the previous, exhaustive search method. The final fit achieved and time required was also recorded for this method (Table I).

TABLE I: HRIR DECOMPOSITION RESULTS

METHOD:	Average Time	Average Fit
Exhaustive, variable window width (old)	429.61 sec	98.24%
Pole approximation (new)	4.23 sec	94.62%

4 Conclusions

The results shown in Table I clearly indicate that the new, pole approximation approach, is about 100 times faster than the exhaustive, variable window width (old) method, reaching a decomposition result in less than 5 seconds, while the previous approach would take more than 7 minutes, in the same computing platform. The goodness of fit of the new approximation was slightly lower than the one achieved through the exhaustive method for the set of HRIRs used in the comparison (94.62% instead of 98.24%). Therefore the new proposed method seems to provide a better alternative for this type of decomposition.

5 Acknowledgements

This work was sponsored by NSF grants IIS-0308155, CNS-0520811, HRD-0317692 and CNS-0426125.

References:

[1] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer-Functions," *Journal of the Acoustical Society of America*, vol. 94, pp. 111-123, 1993.

[2] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *Ieee Transactions on Speech and Audio Processing*, vol. 6, pp. 476-488, 1998.

[3] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *Journal of the Audio Engineering Society*, vol. 49, pp. 472-479, 2001.

[4] A. Barreto and N. Gupta, "Dynamic Modeling of the Pinna for Audio Spatialization," *WSEAS Transactions on Acoustics and Music*, vol. 1, pp. 77-82, January 2004.

[5] K. J. Faller II, A. Barreto, N. Gupta, and N. Rishe, "Decomposition and Modeling of Head-Related Impulse Responses for Customized Spatial Audio," *WSEAS Transactions on Signal Processing*, vol. 1, pp. 354-361, 2005.

[6] K. J. Faller II, A. Barreto, N. Gupta, and N. Rishe, "Performance Comparison of Two Identification Methods for Analysis of Head Related Impulse Responses," in *Advances in Systems, Computing Sciences and Software Engineering*, T. Sobh and K. Elleithy, Eds. Netherlands: Springer, 2006, pp. 131-136.

[7] L. P. Charles and H. T. Nagle, *Digital control system analysis and design (3rd ed.)*: Prentice-Hall, Inc., 1995.