

A NEW INVERSE PROCESSING APPROACH TO THE MODELING OF HEAD-RELATED TRANSFER FUNCTIONS FOR AUDIO SPATIALIZATION

Kenneth J. Faller II

*Department of Electrical and Computer Engineering
Florida International University
Miami, Florida, U.S.A.
kfall001@fiu.edu*

Armando Barreto

*Department of Electrical and Computer Engineering
Florida International University
Miami, Florida, U.S.A.
barretoa@fiu.edu*

Naphtali Rishé

*School of Computing and Information Sciences
Florida International University
Miami, Florida, U.S.A.
rishen@cis.fiu.edu*

ABSTRACT

Currently, achieving high-fidelity sound spatialization requires the prospective user to undergo lengthy measurements in an anechoic chamber using highly specialized equipment. This, in turn, has increased the cost and reduced the availability of high-fidelity spatialization to the general public. An attempt to generalize 3D audio has been made using the measurement of a KEMAR dummy head or creating a database containing a sample of the public. Unfortunately, this leads to increased front/back reversals and localization errors in the median plane. Customizable Head-Related Impulse Responses (HRIRs) would reduce the errors caused by general HRIRs and remove the limitation of the measured HRIRs. This paper reports an initial stage in the development of customizable HRIRs. The ultimate goal is to develop a compact functional model that is equivalent to empirically measured HRIRs but requires a smaller number of parameters that could be obtained from the anatomical characteristics of the intended listener. In order to arrive at such model, the HRIRs must be decomposed into multiple scaled and delayed damped sinusoids, which would reveal the parameters that the compact model needs to have an impulse response similar to the measured HRIR. Previously this type of HRIR decomposition has been accomplished through an exhaustive search of the model parameters. A new method that approaches the decomposition simultaneously in the frequency (Z) and time domains is reported here.

INTRODUCTION

Virtual spatial audio has been applied in many areas from computer video games to assistive technology for the visually impaired. Spatial audio is the ability that allows humans to locate a sound source in three-dimensional (3D) space (Fig. 1).

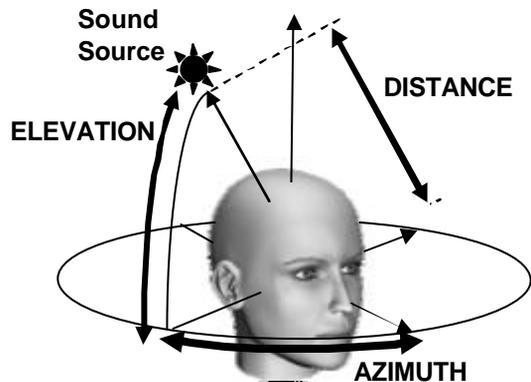


Fig. 1: Diagram of a sound source in 3D space

There are currently two methods for creating virtual spatial audio: the multi-channel and the two-channel approach. The multi-channel approach consists of physically positioning speakers around the listener (e.g., Dolby[®] 5.1 array). This method is effective but requires expensive equipment and relies in the relative positions of the listener and the speakers, which limits its portability.

The two-channel approach uses digital signal processing (DSP) techniques to create binaural (left and right channel) virtual spatial audio from a monaural source that can be delivered to a prospective listener through headphones or speakers. The two primary cues for sound spatialization are inter-aural time differences (ITDs) and inter-aural intensity differences (IIDs). However, ITDs and IIDs are not the only cues used by humans for sound localization.

It is known that sound is spectrally modified as it travels from the sound source to the listener's ear by the listener's anatomical features (torso, head, external ear, etc.) and environment (walls, floor, etc.) [1, 4]. The two-channel approach tries to model this modification. These modifications are modeled by means of a pair (Left and Right) of Head-Related Transfer Functions (HRTFs) for each position around the listener. This modeling framework also allows for the assignment of a virtual position of a digital sound, i.e., the "spatialization" of a digital sound. The implementation of this in a computer requires convolving a sound signal with the impulse responses (HRIRs) of the two HRTFs corresponding to the virtual position. When playing the left and right signals resulting from this convolution to a listener, he/she will perceive the sound as if it emanated from the desired location in 3D space.

The effect of the anatomical features on the HRTFs implies that these transfer functions will be different for every intended listener. Originally, this required individually measuring the intended listener in an anechoic chamber using expensive and cumbersome equipment. A solution to this is the utilization of "generic" HRTFs (e.g., MIT's measurements of a KEMAR Dummy-Head Microphone [7] and the CIPIC Database [1]). Although this provides a general solution, it is known that "generic" HRTFs may lead to front/back reversals and elevation errors in the perception of the sound location for a given specific subject [8]. Hence, the "best" binaural virtual spatial audio is still achieved through measured HRTFs, which requires that all potential listeners be measured in an anechoic chamber or similar facility.

It would be highly desirable to "customize" HRTFs for each listener utilizing relevant geometrical measurements of his/her head and outer ears (pinnae). Unfortunately, state-of-the-art HRTF measurement systems report individual HRTFs in terms of long (128, 256, 512)

sequences of impulse response samples, which are not "tunable" according to the geometrical characteristics of a different listener (other than the one from which they were measured). As a result, the overall purpose of our research is to develop customizable HRIRs from a generic dynamic model that involves a smaller number of parameters.

Using the physical measurements of the intended listener, the generic model can be customized to provide similar spatialization fidelity as measured HRTFs. As mentioned earlier, the current representation of the HRTFs is not "tunable" using the geometric characteristics of a prospective user. Therefore, we believe that decomposition of the HRTFs into partial components will allow re-generation from a reduced number of parameters that are related to the geometry of each listener. This would have significant impact because it would broaden the availability of high fidelity HRTFs to the overall computer user population.

METHODOLOGY

The following subsections will describe the methodology used in our study.

A. Pinna Model

An alternative approach that has been suggested is based on the use of customizable HRIR "structural" models. Brown and Duda [4] have proposed that a "structural" model for binaural sound synthesis should "cascade" the effects of the listener's head (e.g., diffraction, inter-aural delay, etc.) with the local monaural effects of the geometry of the pinna or outer ear. Algazi has already proposed a customizable model of a listener's head, which only requires three anatomical measurements [2]. Later, in [3] a pinna model consisting of a summation of multiple delayed and scaled damped sinusoids was proposed.

In [3], extending upon the models proposed by Brown, Duda and Algazi, we proposed a pinna model in which the sound entering the ear canal is the summation of signals with different delays. In this model the cavities of the ear are modeled as resonators. The sound signal is also delayed and scaled, which represents the reflection of sounds bouncing off the geometrical structures of the ear.

Figure 2 shows a block diagram for the model proposed. The sound is first processed by a resonator and then scaled by a magnitude factor ρ_i and delayed by a delay factor τ_i in each of the

parallel paths. The vertical line farthest to the left represents the direct transmission of sound while the other vertical lines represent indirect paths to the ear canal. The pinna model shown in Figure 2 only requires 11 parameters (the resonator is represented by two parameters), and could be “cascaded” with Algazi’s functional head model to represent a complete HRIR.

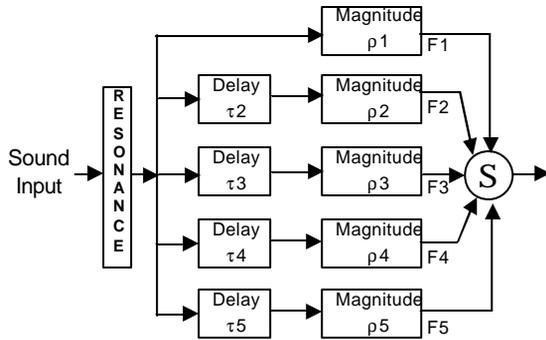


Fig. 2: Block diagram of the pinna model

The parameters in Figure 2 must be obtained from empirically measured HRTFs which will allow for the creation of a database of these parameters (at numerous azimuths and elevations). It is expected that, if the data set is large enough, a relationship can be established between the model parameters and the anatomical features. With this database, a new subject’s geometric characteristics could be “converted” to parameter values which can be used to instantiate the model in order to create customized HRTF for a desired source location.

B. Automated Decomposition Method

The pinna model proposed by [3], as mentioned in the previous section, required the decomposition of the measured HRTFs in order to create customized HRTFs. In [3], the decomposition process was achieved by manually adjusting a windowed portion of the measured HRIR. This window portion was used in an attempt to reconstruct the damped sinusoidal believed to reside in that window segment. Modeling methods like Prony or similar are used to try to obtain the entire damped sinusoidal.

We have developed an iterative method to automate the decomposition process in the time domain. Windowing the HRIR, in a similar fashion to the manual method, the iterative method compares the reconstructed HRIR with the measured HRIR. The highest fitting

reconstructed HRIR is considered the best approximation to the measured segment. A complete description of this original automated method is available in [6].

A major drawback of this method is that the window sizes are initially unknown. The program has to iterate through all possibilities. These windows are gradually opened progressively from 2 to 10 samples for each of the five windows studied in each HRIR [6]. Each of the windows will approximate a damped sinusoid and each possible sequence of second-order approximations (considered at the appropriate delays) would be summed together resulting in a candidate HRIR. All the candidate HRIRs are temporarily stored and compared to the measured HRIR using Equations 1 and 2 to assess their individual similarity to the original measured HRIR or “fit” where “MS” is the mean square. The candidate HRIR with the highest “fit” is considered to be the reconstructed HRIR that most accurately portrays the measured HRIR. Analysis of the results from this process showed that, in general, it approximates the original HRIR with relatively high accuracy. Figure 3 shows the components extracted from a measured HRIR by this process.

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

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

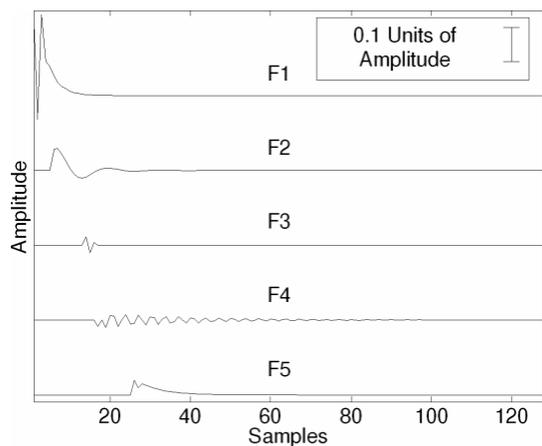


Fig. 3: Five damped sinusoids obtained from a measured HRIR

The results of the iterative process resulted in high fits (96% average fit for the cases studied in

[6]). Although this is promising, the iterative process is computationally expensive even with just five windows processed in this study. A tree diagram of this iterative process reveals that, by iterating through all possible widow combinations, it will generate $9 \times 9 \times 9 \times 9 = 59,049$ leaf nodes. If any other windows are added, they multiply the number of combinations by 9 for each additional window, which exponentially increases the computation time. Unfortunately, to obtain the best combination of windows all possibilities must be explored and the reconstructed HRIR for each combination must be compared to the measured HRIR. It became apparent that additional windows maybe needed to model late components that are not realized using only 5 windows. With 5 windows this process is already taxing, but the prospective addition of more windows further underscored the need for a less computationally complex approach.

Another drawback of this method is the potential inability to accurately reconstruct damped sinusoids when the delay between them is small (less than 5 samples). To explore this potential limitation, a damped sinusoid (x) was created and is shown in Figure 4 (top). Signal x was then tested with the iterative decomposition method using Prony and Steiglitz-McBride (STMCB) function approximations. A small window with a 3 sample width was used in an attempt to approximate the entire signal. The approximation signals resulting from the STMCB (x_s) and Prony (x_p) methods are also shown in Figure 4. As can be seen in the results, x_s and x_p fail to properly reconstruct the original signal x . This would lead to inaccurate approximations of the parameters for the pinna model. These drawbacks have prompted us to develop a new, faster and potentially more robust method of HRIR decomposition into sequential damped sinusoids.

C. Inverse Processing Approach

The purpose of this method is to decrease the computation time and to increase the accuracy when the delay between the damped sinusoids is small. In the iterative approach outlined in the previous section, the windowed portion of the HRIR was assumed to contain only one second-order damped sinusoid. Under this assumption, the windowed portion of the HRIR can be processed with the Prony or STMCB decomposition method in order to obtain the

entire damped sinusoid for that windowed portion. Unfortunately, the window size was unknown and, to accurately reconstruct the HRIR, all possible window combinations had to be explored.

In the new decomposition method, the windows do not have to be predefined. Instead, an attempt is made to isolate the damped sinusoids according to their pole signatures in the Z-domain. A high-order approximation is used on the complete HRIR remnant (at any stage during the decomposition) and the candidates for the damped sinusoids are isolated by identifying the conjugate pole pairs in the resulting high-order approximation. This is possible because, in general, a single damped sinusoidal is represented by conjugate poles within the unit circle and a zero at the origin in the Z-domain (Fig. 5). Hence, Equation 3 where k is a scalar and p_1 and p_2 are complex poles can describe a damped sinusoid in the Z-domain. According to this equation, if the scalar k and the poles are known then, using the inverse Z-transform, it is possible to characterize the corresponding time domain sequence as a specific damped sinusoid.

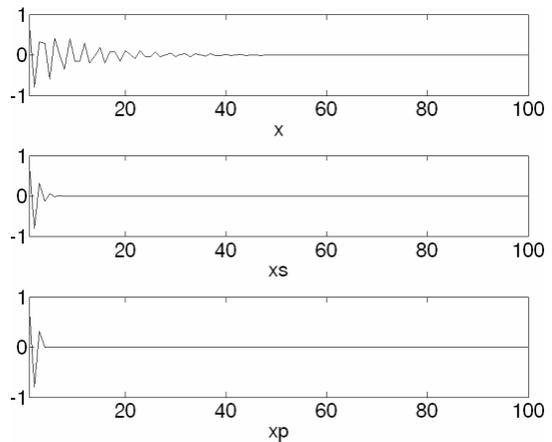


Fig. 4: x (top) vs. x_s (middle) and x_p (bottom)

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

This new approach is an iterative approach as well, but it differs from the previous approach in that it does not try every possible window width combination. Instead, each of the viable damped sinusoids that can be isolated by their conjugate pole pairs in the Z-domain is used as a possible candidate. Then each damped sinusoid will be

investigated as the approximation of the current segment. The end of the segment is not predefined. Instead, each candidate sinusoid will be subtracted from the previous HRIR and the time index at which the remainder surpasses a predetermined threshold will be considered the beginning of the next segment. The remainder of the HRIR will be shifted to this point and the process will be repeated with one less modeling order than in the previous iteration.

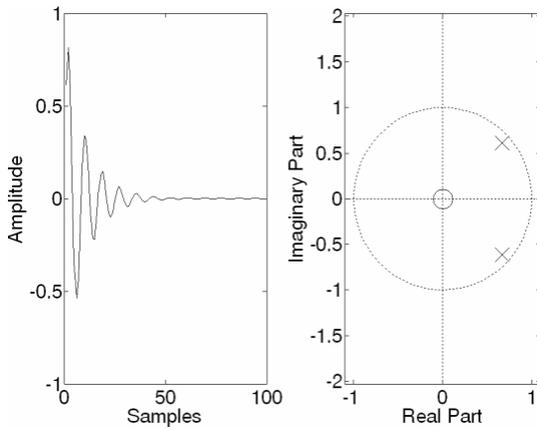


Fig. 5: Time domain and zero-pole plot of a single 2nd order damped sinusoidal

In contrast with the window-based method, an analysis of the tree diagram of the new iterative process reveals a much simpler and compact structure. This is due to the decrease in the number of possible combinations that need to be explored, since at each subsequent node the branching factor decreases by one. For example, if 5 damped sinusoids are sought in total, then only $5 \times 4 \times 3 \times 2 \times 1 = 5! = 120$ leaf nodes will exist. In order to verify this, an experiment was performed. To simplify the explanation, only three damped sinusoids were summed together, each having different delays and magnitudes. The details of the experiment are outlined in the following subsection.

D. Simulation of Inverse Processing Approach

Equation 4 was used to create the three damped sinusoids for this simulation. The sinusoids were each N points in length, $n = 0, \dots, N-1$, d_i is the negative damping factor and ω_d is the digital frequency. Desired delays (t_2 and t_3) were applied to x_2 and x_3 respectively to obtain x_{2s} and x_{3s} . Finally x_1 , x_{2s} and x_{3s} were summed point-to-point resulting in the test signal,

x . In this example $N=100$, $t_2=3$, $t_3=6$, $\omega_d=0.711$, $d_1=-0.1$, $d_2=-0.125$ and $d_3=-0.15$. The three signals (x_1 , x_{2s} and x_{3s}) and the resulting signal (x) are shown in Figure 6.

$$x_i(n) = e^{d_i * n} \cdot \sin(\omega_d \cdot p \cdot n) \quad (4)$$

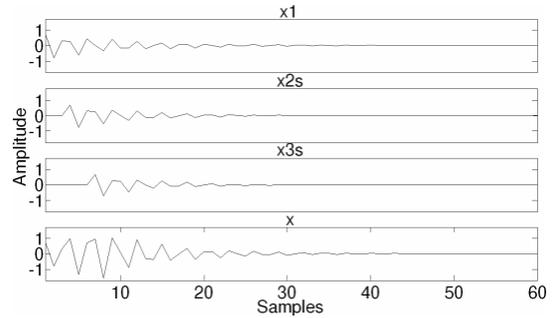


Fig. 6: Plot of the three damped sinusoids (x_1 , x_{2s} with delay t_2 and x_{3s} with delay t_3) and the sum of them (x)

The following paragraphs outline the steps for the inverse processing algorithm. First, a sixth order STMCB approximation process was applied to the complete test signal x . In some cases, the number of damped sinusoids contained in a signal may be unknown but in this case it is known that the test signal contains three second order damped sinusoids. Therefore, a sixth order STMCB is initially used. The result of the sixth order STMCB will have the pole structure shown in Figure 7.

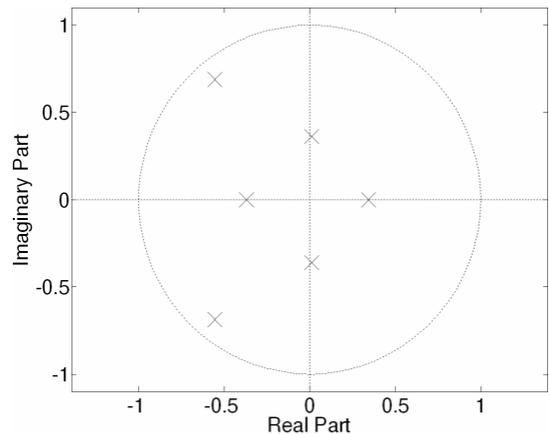


Fig. 7: Poles obtained from the 6th order STMCB approximation of the complete sequence x

As seen in the figure, there are two conjugate pole pairs. Those two pairs will be investigated as possible candidates for the first damped sinusoidal. In this case, since the STMCB order was six, the resulting approximation could have up to three possible conjugate pole pairs. (To clarify further, if the STMCB order was q then there would be up to $2q$ conjugate pole pairs in the resulting approximation).

Once the conjugate pole pairs are isolated, their corresponding time domain representations are each subtracted from x resulting in a residue sequence like the one shown in Figure 8. A threshold is then applied to the residue to determine the end of the segment and the location of the onset of the remaining sinusoids. The threshold for this particular example was set to 25% of the maximum peak of the residue sequence. An empirically defined threshold is presented later for use with HRIRs.

In Figure 8, an estimate of t_2 is the time at which the residue sequence surpasses the threshold. This will be considered the onset of the next damped sinusoid and is as the beginning point for the second stage of the decomposition. As mentioned in the previous section, a damped sinusoid is removed at each subsequent stage. Therefore, there should be one less damped sinusoid in each new remnant. This results in the application of a fourth-order STMCB in the second stage which yields 4 poles. These four poles are used to synthesize up to two candidates for the second damped sinusoid which will be subtracted from x . This process is repeated until the last damped sinusoid is extracted from x .

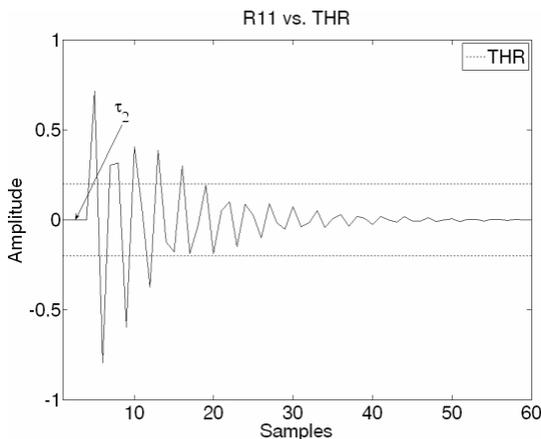


Fig. 8: Plot of the first remnant with threshold lines (THR)

After M stages of decomposition, all the possible combinations must be explored to determine which candidate result best approximates x . This results in a search tree with $M!$ leaf nodes, with each node representing M delayed and scaled damped sinusoids that must be added together to obtain an approximation to the full original signal x . All the combinations must be tested against x using Equations 1 and 2, which will yield their individual fits. The combination with the highest fit is considered the best approximation to the original overall signal. In the example described in this subsection, the winning combination of candidate damped sinusoids achieved a 99.99% fit when compared to the original signal x . Furthermore, the individual damped sinusoids matched x_1 , x_2 s and x_3 s very closely, as well.

INVERSE PROCESSING OF HRIRs

The new method described in the previous subsection was used to decompose 14 actual measured HRIRs recorded, at a sampling rate of 98 KHz, with the AuSIM HeadZap system at FIU. The process was identical to the description in the previous section except that the total estimated number of damped sinusoids was $M = 5$ and the threshold was empirically defined (18% of the signal peak value). Since $M = 5$, the initial STMCB order used was 10.

The determination of the threshold level was achieved by finding the maximum of the average fit for the reconstructed HRIRs as the threshold was incremented in steps of 0.5% from 0% to 40% (of the peak amplitude of the remnant) for HRIRs measured from 14 subjects. The sound source was at $\pm 90^\circ$ azimuth (i.e., directly lateral from the ear measured) and elevations from -36° to 54° at increments of 18° were considered. Figure 9 shows the average fit found at different thresholds for an elevation of -36° . As can be seen, there is a maximum for a threshold of 18% of the peak amplitude (0.18) and very similar patterns were observed for other elevations. Therefore, 0.18 was selected as the threshold for the HRIRs decomposed in this project.

To assess the potential of the new decomposition method in a realistic context, HRIRs from 14 subjects for an elevation of 0° and azimuths from -150° to 180° at increments of 30° (along the horizontal plane) were decomposed using the old (window-based) and the new (pole pair-based) algorithms. The results for each ear are displayed in Table 1.

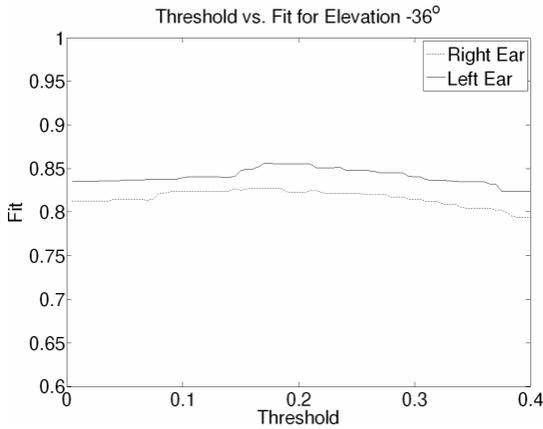


Fig. 9: Threshold versus fit for elevation -36°

Table 1: HRIR Decomposition Results

Method	Average Fit Left Ear	Average Fit Right Ear
Exhaustive, variable window width (old)	97.57%	97.57%
Inverse Processing w/ Threshold (new)	89.40%	88.15%

While the goodness of fit achieved by both methods is similar, the pole decomposition method has been found to be much faster than the old method, as detailed in the next section. Figures 10 to 12 show examples of high, average and low fit cases for HRIRs using the “new” method, respectively.

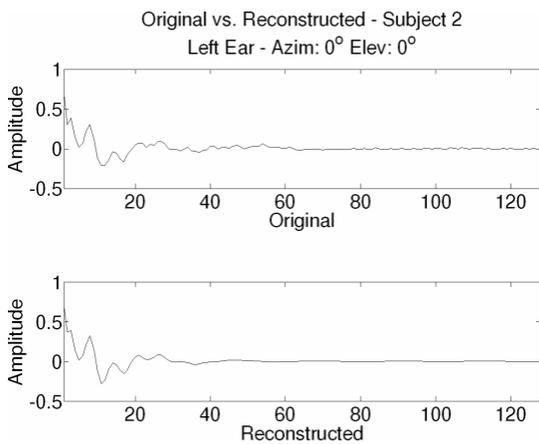


Fig. 10: Original (top) vs. reconstructed HRIRs for the left ear of subject 2 for azimuth 0° and elevation 0° - Highest Fit example

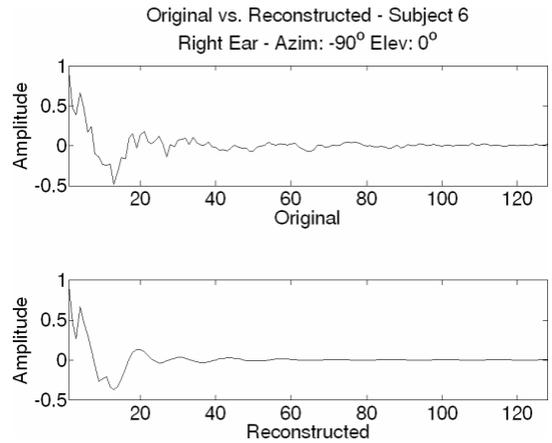


Fig. 11: Original (top) vs. reconstructed HRIRs for the right ear of subject 6 for azimuth -90° and elevation 0° - Average Fit example

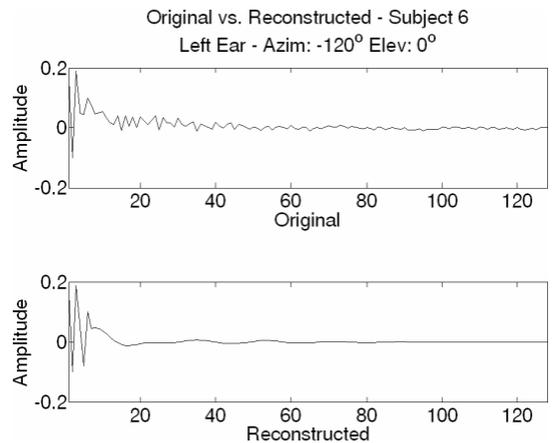


Fig. 12: Original (top) vs. reconstructed HRIRs for the left ear of subject 6 for azimuth -120° and elevation 0° - Lowest Fit example

CONCLUSION

Although the “old” method achieved higher fit than the “new” method (Table 1), there are some severe drawbacks to the “old” method. First, when the delay is small (less than five samples), the second order STMCB modeling method may inaccurately approximate the signal. Secondly, the average calculation time for the “old” method was about 100 times longer than for the “new” method when tested with the 14 measured HRIRs, using $M = 5$ (429 seconds compared to 4.2 seconds).

Therefore, according to these observations, it would be reasonable to recommend the “new” inverse processing method for the creation of a large database, based on the separation of damped sinusoids according to their pole pair signature in the Z-domain, especially if five or more components are sought. This kind of large-scale study will be necessary in order to establish a relationship between model parameters and the anatomical characteristic of the intended listener that we ultimately seek.

ACKNOWLEDGMENT

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

REFERENCES

1. Algazi, V., Duda, R., Thompson, D. and Avendano, C. The Cipc HRTF database 2001 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* New Paltz, NY, 2001.
2. Algazi, V.R., Avendano, C. and Duda, R.O. Estimation of a spherical-head model from anthropometry. *Journal of the Audio Engineering Society*, 49 (6). 472-479.
3. Barreto, A. and N. Gupta Dynamic Modeling of the Pinna for Audio Spatialization. *WSEAS Transactions on Acoustics and Music*, 1 (1). 77-82.
4. Brown, C.P. and Duda, R.O. A structural model for binaural sound synthesis. *Ieee Transactions on Speech and Audio Processing*, 6 (5). 476-488.
5. Charles, L.P. and Nagle, H.T. *Digital control system analysis and design (3rd ed.)*. Prentice-Hall, Inc., 1995.
6. Faller II, K.J., Barreto, A., Gupta, N. and Rishe, N. Decomposition and Modeling of Head-Related Impulse Responses for Customized Spatial Audio. *WSEAS Transactions on Signal Processing*, 1 (3). 354-361.
7. Gardner, B., Martin, K. and Massachusetts Institute of Technology. Media Laboratory. Vision and Modeling Group. *HRFT measurements of a KEMAR dummy-head microphone*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1994.
8. Wenzel, E.M., Arruda, M., Kistler, D.J. and Wightman, F.L. Localization Using Nonindividualized Head-Related Transfer-

Functions. *Journal of the Acoustical Society of America*, 94 (1). 111-123.