Decomposition of Head Related Impulse Responses by Selection of Conjugate Pole Pairs

Kenneth John Faller II[1], Armando Barreto[1], Navarun Gupta[2] and Naphtali Rishe[3]

Electrical and Computer Engineering Department[1] and School of Computer and Information Sciences[3]
Florida International University
Miami, FL 33174 USA

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

Abstract - **Currently, to obtain maximum fidelity 3D audio, an intended listener is required to undergo time consuming measurements using highly specialized and expensive equipment. Customizable Head-Related Impulse Responses (HRIRs) would remove this limitation. This paper reports our progress in the first stage of the development of customizable HRIRs. Our approach is to develop compact functional models that could be equivalent to empirically measured HRIRs but require a much smaller number of parameters, which could eventually be derived from the anatomical characteristics of a prospective listener. For this first step, HRIRs must be decomposed into multiple delayed and scaled damped sinusoids which, in turn, reveal the parameters (delay and magnitude) necessary to create an instance of the structural model equivalent to the HRIR under analysis. 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.**

I. Introduction

The emergence of inexpensive and powerful computers has expanded virtual spatial audio to many areas. Virtual spatial audio is the use of digital signal processing (DSP) techniques to assign an artificial sense of directionality to digital sound signals.

Currently, there are two approaches to virtual spatial audio: multi-channel and two-channel approaches. The multi-channel approach uses multiple (more than two) speakers placed around the listener at strategically defined locations (e.g., Dolby 5.1 array) to physically reproduce the directionality of sounds generated around the listener. This approach produces emulated spatial sounds in a limited listening region which are then perceived by the listener, much like he/she would perceive naturally occurring sounds. However, this relies on the proper positioning of the speakers around the listener, which limits the use of the approach to stationary uses such as a home theater system.

The two-channel approach uses DSP techniques to create binaural sound pairs (Left ear signal, Right ear signal) for virtual spatial audio digitally so that they can be delivered to the listener through stereo headphones. It is known that sound signals are altered by the physical environment (e.g., floor, ceiling, walls, listener's torso, listener's head, and listener's outer ear) as they travel from the source to the eardrums of the listener. The two-channel approach strives to replicate this process synthetically, so that the listener can locate the virtual spatial audio source, at the location being emulated. The synthetic transformation is performed by application of special digital filters that are characterized by their impulse responses, called head-related impulse responses (HRIRs). Every position around the listener will have a specific HRIR, for each ear, associated with it. Convolving a sound signal with each HRIR for a desired location modifies the signal in a way that is similar to modifications the environment would have produced on the signal.

Logically, HRIRs depend on the anatomical features (outer ear, head, and torso) of the listener. As a result, HRIRs for each different location differ from person to person. Ideally, the HRIRs of each prospective listener would have to be measured empirically, at numerous source locations, in order to achieve highly convincing virtual spatial audio. However, this requires specialized personnel and expensive equipment that includes a small, wide bandwidth speaker and miniature microphones placed in the ear canal of the subject (Fig. 1)

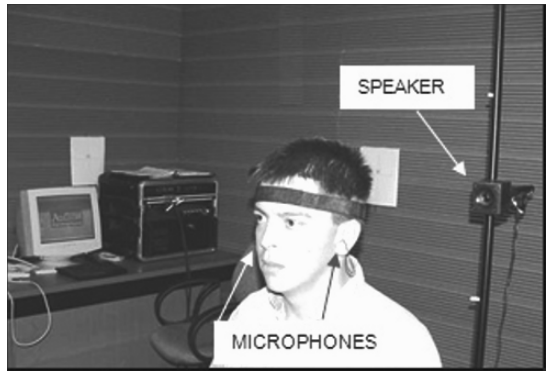

Fig. 1 Empirical HRIR measurement at FIU.

K. Elleithy (ed.), Advances and Innovations in Systems, Computing Sciences and Software Engineering, 259–264.

Since it is not possible to provide access to this measurement process for every potential user of virtual spatial audio, commercial developers have resorted to the use of "generic" HRIR pairs obtained experimentally from a mannequin of "average anatomical dimensions" (e.g., MIT's measurements of a KEMAR Dummy-Head Microphone [1]) or using a limited number of subjects to represent the general population (e.g., the CIPIC Database [2]). These databases include HRIR pairs for many different positions around the listener, defined in terms of their azimuth (θ), elevation (ϕ) and distance (r) in a spherical coordinate system (Fig. 2)

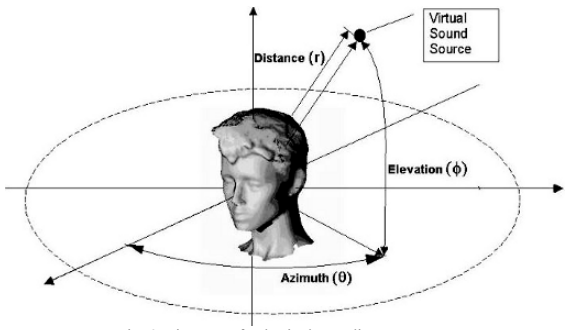

Fig. 2 Diagram of spherical coordinate system

Unfortunately, this type of "generic" HRIRs yield only an approximate sense of source location in many users, lacking the high spatialization fidelity of individual HRIRs [3].

The overall purpose of our research is to develop customizable HRIRs from a generic dynamic model. The generic model can be customized using physical measurements of the listener to provide similar spatialization fidelity as measured HRIRs. The current representation of HRIRs prohibits customization using geometric characteristics of the intended listener. Therefore, we believe that decomposition of HRIRs into partial components will allow their re-generation from a reduced number of parameters that are related to the geometry of each intended listener. Efficient HRIR customization could have significant practical impact because it would extend the benefits of high-fidelity audio spatialization to the overall computer user population.

II. Methodology

The following subsections describe the methodology used in this paper.

A. Structural Pinna Model

Brown and Duda in [4] proposed a "structural" model for binaural sound synthesis. In this approach, effects of the head, shoulders and pinna (outer ear) are "cascaded" to create a transfer function that contains all the spectral cues necessary to generate synthesized binaural sound. However, they did not provide a method to define the parameters of their pinna sub-model. A customizable functional model developed by Algazi models a listener's head with only 3 simple anatomical measurements [5].

In [6] we proposed a pinna model in which the sound entering the ear canal is the summation of signals with different delays. The delays are a result of waves bouncing off of the geometrical structures of the pinna, into the ear canal. The effect of the pinna cavities is modeled with a resonator. Therefore, the HRIRs were broken down into one direct wave and three delayed waves. Recent research by our group has achieved improvements in the decomposition of HRIRs augmenting the model with an additional delayed wave. A block diagram of this augmented model is shown in Figure 3.

In this model, the parallel paths represent the multiple signals entering the ear canal. Each indirect signal will arrive at the ear canal after a delay, τ_i, with respect to the direct wave. Additionally, the indirect signal will also have less energy, represented by a magnitude factor ρ_i , in comparison with the direct wave. The pinna model shown in Figure 3 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.

An efficient method must be found to obtain the parameters in the model of Figure 3 from HRIRs obtained empirically as long sequences of impulse response samples. This will enable the development of databases of these parameters (at numerous azimuths and elevations) for subjects whose relevant anatomical characteristics will also be measured. Our expectation is that once the data set is large enough, empirical relationships can be developed between the model parameters and the anatomical features. At that point the geometric characteristics of a new intended user could be measured and "converted" to parameter values to instantiate the model at a desired location.

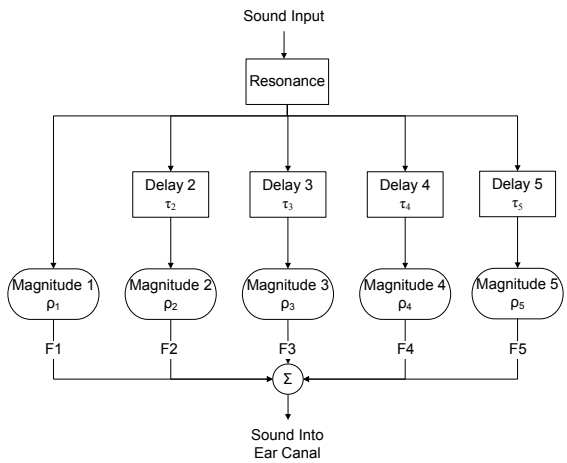

Fig.3: Block diagram of the pinna model

B. Iterative Decomposition Method

The impulse response of the model shown in Figure 3 will be the superposition of a damped sinusoidal (i.e., the impulse response of the resonator) with other damped sinusoids that appear delayed and scaled. Decomposition of a measured HRIR into this kind of sinusoidal components will reveal the

delays and scaling factors that should be used in the model to create an instance that will have a close approximation of the HRIR being decomposed as impulse response.

Time-domain methods for this decomposition have been suggested before. In [6-8], two of the these methods, based on the Prony and Steiglitz-McBride (STMCB) signal modeling methods, were compared for decomposition of HRIRs. All these methods sought to apply second-order signal modeling to windowed sections of the HRIR that could be assumed to contain only a single damped sinusoid, which Prony and STMCB approximate with reasonable accuracy [9-11]. A full description of those methods can be found in [6-8].

A major drawback of this approach, however, is that the window sizes are not initially known. To discover the appropriate window sizes, a program was written to iterate through all possibilities. The windows were gradually widened starting from 2 to 10 sampling intervals for each window (for a total of five windows). In each tentative window the signal would be approximated using one of the modeling methods (second-order Prony or STMCB) and each possible sequence of second-order approximations (considered at the appropriate delays) would be summed together resulting in a candidate HRIR. All the possible resulting candidate HRIRs would be temporarily stored and eventually compared to the original measured HRIR using Equations 1 and 2 to assess their individual similarity or "fit" to the original HRIR. The candidate HRIR with the highest fit at the end of this process would be kept as the "reconstructed" HRIR that represents the most accurate decomposition achievable for that original (measured) HRIR. Analysis of the results from this process showed that, in general, it approximates the original HRIR with relatively high accuracy. Figure 4 shows the components extracted from a measured HRIR by this process.

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

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

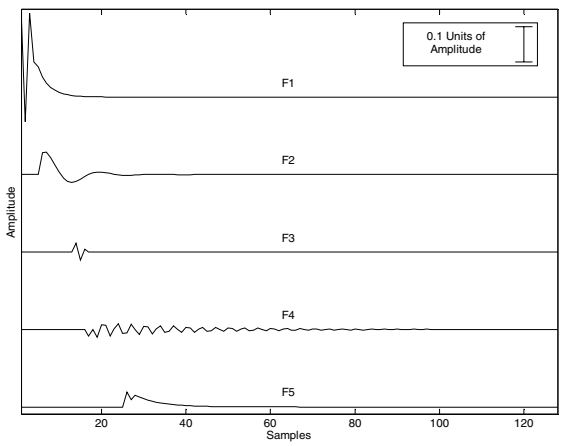

Fig.4: Five damped sinusoidal components obtained from a measured HRIR

Although this iterative process resulted in high fits for most of the HRIRs explored (about 96% average fit), the iterative 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 iterative search method.

Another drawback is that when the delay between sinusoids is small (less than 5 samples), the second-order STMCB or Prony sequential methods alone tend to inaccurately reconstruct the signal. To verify this, a single damped sinusoidal (x) was created and tested with the iterative method using Prony and STMCB. A short window containing only the first three samples from x was processed by STMCB and Prony in an attempt to approximate the original signal. The results of STMCB (xs) and Prony (xp) are shown in Figure 5. The approximations xs and xp appear to capture the details of the beginning part of x but fail to approximate the rest of it. This will lead to inaccurate approximation of the parameters for the pinna model. These drawbacks have prompted us to develop a new, faster and potentially more accurate method of HRIR decomposition into sequential damped sinusoids.

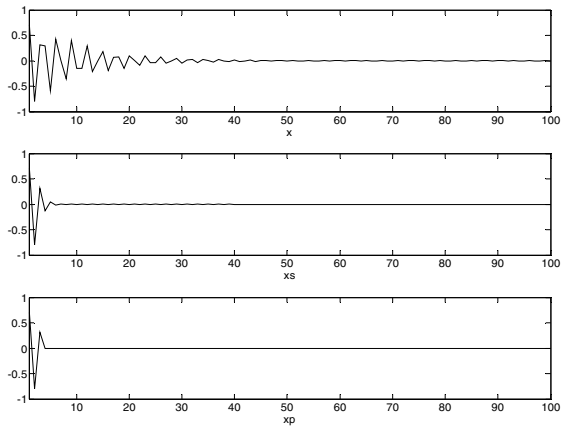

Fig. 5: x (top) vs. xs (middle) and xp (bottom)

C. Pole-Decomposition Method

In previous HRIR decomposition approaches the goal was always to isolate a segment of the HRIR that could be considered constituted by a single damped sinusoid. Under that assumption a second-order modeling approach (Prony or STMCB) was used to deal with every window along the HRIR. However, the correct demarcation of the boundaries for

these second-order windows was crucial to the accuracy of the process and, therefore, all probable window widths, for each of the sequential windows, had to be considered. This resulted in a search tree with a branching factor that remained high (e.g., 9) from the root node all the way to the leaf nodes.

In the new decomposition approach the end of the analysis windows does not need to be defined in advance. Instead a higher-order approximation is used on the complete remnant of the HRIR (at any point during the decomposition) to pre-define multiple damped sinusoids contained in the HRIR remnant, and then they are individually isolated according to their pole signature in the Z-domain, and pursued as candidates for the second-order representation of the particular HRIR segment in question.

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 6) [12]. Hence, a damped sinusoid in the Z-domain can be described with the following general equation:

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

where k is a scalar and p_1 and p_2 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 characterize the corresponding time domain sequence as a specific damped sinusoid.

In this new approach, instead of iterating through all possible window width combinations, an attempt is made to identify multiple delayed and damped sinusoids in the complete HRIR remnant available. Each of the viable damped sinusoids will be separated according to their conjugate pole signatures in the Z-domain. Then each damped sinusoid will be investigated as the approximation of that particular segment. The end of the segment is not pre-determined, but instead will be defined by the time index at which the reminder of the previous HRIR remnant minus the second-order approximation being investigated surpasses a predetermined threshold. That point will be considered the time at which a new damped sinusoidal contribution starts. The origin for analysis will be shifted to that point and the process will be repeated, except that using a modeling order which is two less than the previous modeling order used.

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. An experiment using simulated damped sinusoids was performed in order to verify this pole decomposition method. Three damped sinusoids with different magnitudes and delays where created and summed together, to be analyzed by the pole decomposition method.

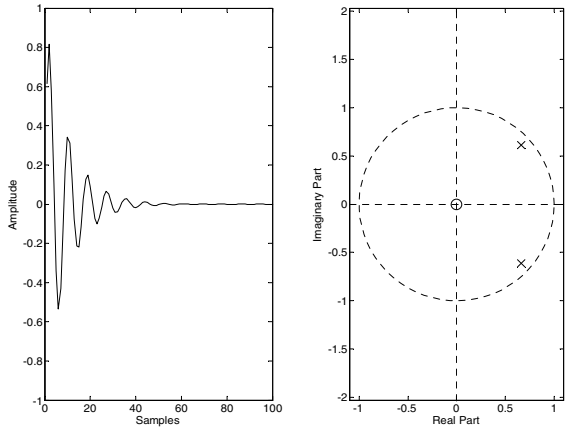

Fig. 6: Time domain and Zero-Pole plot of a single damped sinusoidal

The damped sinusoids used in this example were created using equation 4 where N is the length of the signal, n = 0,…,N-1, d_i is the negative damping factor and ω_d is the digital frequency. Once the three sinusoids (x1, x2 and x3) were created, the desired delays (τ_2 and τ_3) were applied to the last two sinusoids respectively, resulting in x2s and x3s. Finally, the sinusoids were summed point-to-point to produce the test signal (x). In this example N=100, τ_2=3, τ_3=6, ω_d=0.711, 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 7.

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

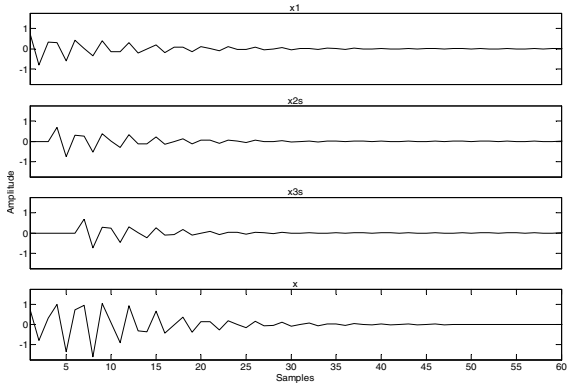

Fig. 7: Plot of the three damped sinusoids (x1, x2s with delay τ_2 and x3s with delay τ_3) and the sum of them (x)

The process starts by applying a sixth order STMCB approximation process to the complete x. Sixth order is used initially because the decomposition of x into three second – order signals (damped sinusoids) is sought. The results from the sixth-order STMCB approximation will have the pole structure shown in Figure 8. As seen in the figure, there are two complex conjugate pairs of poles. Each of these will be

investigated as a candidate to represent the first sinusoidal present in x (there could be up to three branches at the initial node of this search tree, if all the poles were complex).

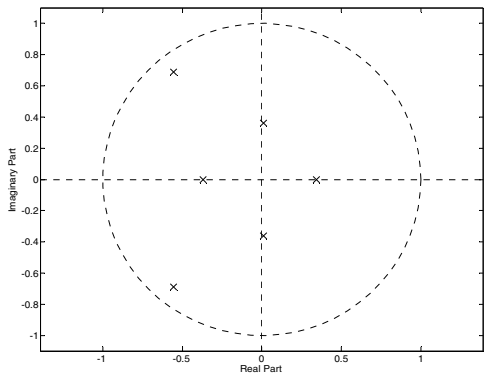

Fig. 8: Poles obtained from the sixth-order STMCB approximation of the complete sequence x

The investigation of each of these alternatives involves its subtraction from x to define a residue sequence, as shown in Figure 9, which will then be thresholded. The threshold level used for this segmentation was set at 25% of the signal peak, in this synthetic example. A slightly different threshold to process real HRIRs was found as described in the following section. This is the only adjustable parameter in our method and the rationale for the value recommended is also presented in the following section of the paper.

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 in it.

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.

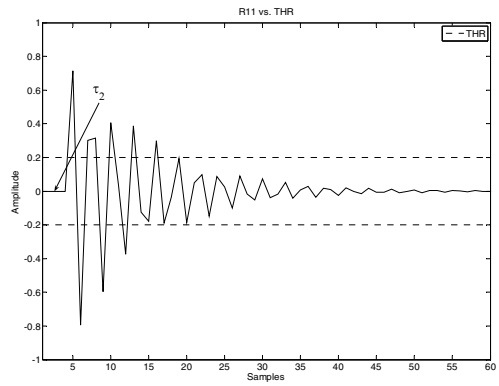

Fig. 9: Plot of the first remnant with threshold lines (THR)

III. POLE DECOMPOSITION OF MEASURED HRIRS

The method described in Section II-C was applied to the decomposition of 14 actual HRIRs, recorded from 14 subjects using the AuSIM HeadZap system at Florida International University (Fig. 1). The goal in each case was to obtain M = 5 damped sinusoidal components. Therefore, the order of the first STMCB approximation process was 10. The procedure was identical as the one explained for the decomposition of the synthetic sequence x, in Section II-C, with the exception that an empirically defined threshold level (18% of the signal peak value) was applied to each reduced remnant of the HRIR.

The empirical determination of the best threshold level to use in decomposing actual HRIR signals was performed by plotting the average fit for the reconstructed HRIR as the threshold used changed in increments of 0.005 for HRIRs measured from 14 subjects and corresponding to sound sources at +/-90º azimuth (i.e., directly lateral from the ear measured) and elevations from -36º to 54º at increments of 18º. For example, Figure 10 shows this plot for ϕ = -36º. As can be seen in this plot, there is a curvature that has a maximum at a threshold value of about 0.18. Similar observations were made for other elevations. Thus 18% was selected as the recommended threshold.

HRIRs from 14 subjects for an elevation of 0º and azimuths from -150º to 180º at increments of 30º (along the horizontal plane) were decomposed using the old and the new algorithms. The results for each ear are displayed in Table I.

TABLE I: HRIR DECOMPOSITION RESULTS

METHOD:	Average Fit Left Ear	Average Fit Right Ear
Exhaustive, variable window width (old)	97.57%	97.57%
Pole decomposition 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 11 to 13 show the highest, average and lowest fits for HRIRs using the "new" method, respectively.

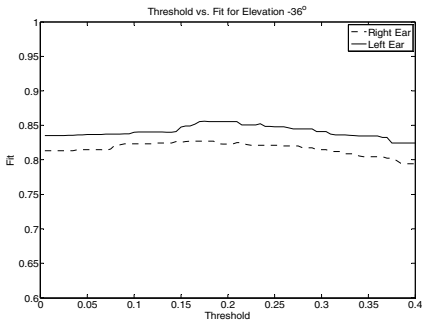

Fig.10: Threshold versus fit for elevation -36º

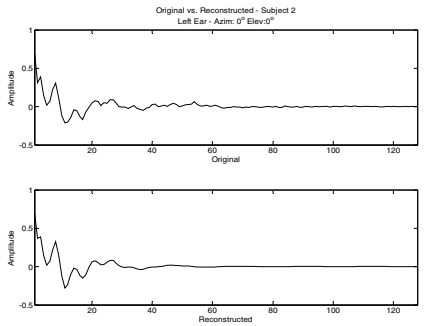

Fig.11: Original (top) vs. reconstructed HRIRs for the left ear of subject 2 for azimuth 0º and elevation 0º - Highest Fit example

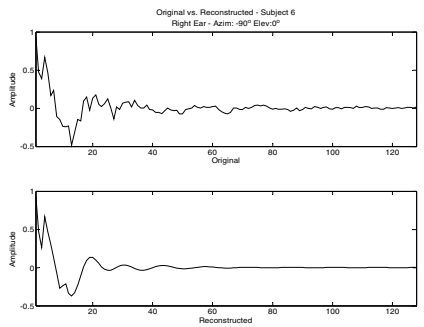

Fig.12: Original (top) vs. reconstructed HRIRs for the right ear of subject 6 for azimuth -90º and elevation 0º - Average Fit example

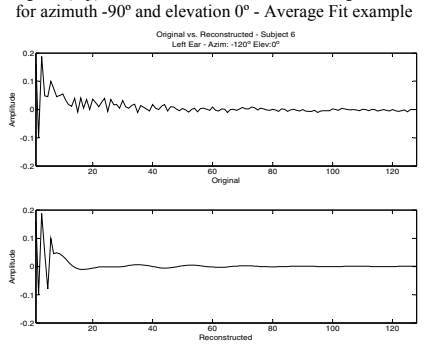

Fig.13: Original (top) vs. reconstructed HRIRs for the left ear of subject 6 for azimuth -120º and elevation 0º - Lowest Fit example

IV. Conclusion

The results shown in Table I indicate that the "old" method achieved a slightly higher average fit, but exhibited several drawbacks. First, the average calculation time was found to be about 100 times longer for the "old" method when a test set of 14 HRIRs were decomposed by both approaches (429 s to 4.2 s). Secondly, when the delay is small (less than 5 samples wide), the second-order STMCB sequential method alone tends to inaccurately reconstruct the signal.

Therefore, according to the observations indicated above, it seems that the new HRIR decomposition method, based on the separation of damped sinusoids according to their pole pair signature in the Z-domain, may be a more practical approach to the creation of a large database of decomposed HRIRs (particularly if more than 5 components will be sought), which is a pre-requisite to the establishment of relationships between model parameters and measurable anatomic characteristics of the subjects.

Acknowledgment

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

References

[1] B. Gardner, K. Martin, and Massachusetts Institute of Technology. Media Laboratory. Vision and Modeling Group., *HRFT measurements of a KEMAR dummy-head microphone*. Cambridge, Mass.: Vision and Modeling Group, Media Laboratory Technical R4port 280, Massachusetts Institute of Technology, 1994.

[2] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The Cipic HRTF database," in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* New Paltz, NY, 2001, pp. 99-102.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] L. Ljung, *System Identification - Theory For the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[10] T. W. Parks and C. S. Burrus, "Digital filter design," Wiley-Interscience, 1987, pp. 226-228.

[11] K. Steiglitz and L. McBride, "A technique for the identification of linear systems," *IEEE Transactions on Automatic Control*, vol. 10, pp. 461-464, 1965.

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