Inverse Problems in Science and Engineering
Vol. 17, No. 1, January 2009, 51–63

Taylor & Francis
Taylor & Francis Group

A new inverse processing approach to the modelling of head-related transfer functions for audio spatialization

K.J. Faller II[a]*, A. Barreto[a] and N. Rishe[b]

[a]*Department of Electrical and Computer Engineering, Florida International University, Miami, Florida, USA;* [b]*School of Computing and Information Sciences, Florida International University, Miami, Florida, USA*

(*Received 16 April 2007; final version received 30 November 2007*)

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. Attempts to generalize 3D audio have 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 article 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 a 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.

Keywords: head-related transfer functions (HRTFs); head-related impulse responses (HRIRs); singal decomposition; customizable spatial audio; structural pinna model; 3D audio

1. 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 (Figure 1).

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

*Corresponding author. Email: kfall001@fiu.edu

ISSN 1741–5977 print/ISSN 1741–5985 online

DOI: 10.1080/17415970802082849
http://www.informaworld.com

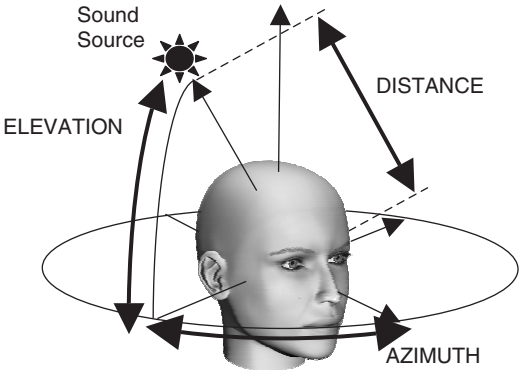

Figure 1. Diagram of a sound source in 3D space.

but requires expensive equipment and relies on 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.) [2,4]. The two-channel approach tries to model this modification. These modifications are modelled by means of a pair (left and right) of head-related transfer functions (HRTFs) for each position around the listener. This modelling 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 [2]). 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. The inverse problem being addressed in this paper is to decompose the HRTFs obtained from the measurement system into smaller subcomponents that are easily related to physical features of an intended listener.

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 a significant impact because it would broaden the availability of high fidelity HRTFs to the overall computer user population.

2. Methodology

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

2.1. *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 [1]. Later, in [3] a pinna model consisting of a summation of multiple-delayed and scaled-damped sinusoidals 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 modelled 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 horizontal line at the top represents the direct transmission of sound while the other horizontal 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.

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

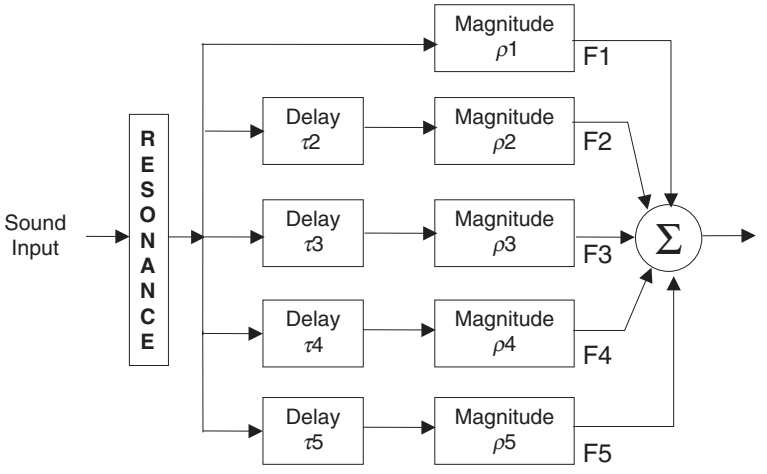

Figure 2. Block diagram of the pinna model.

values which can be used to instantiate the model in order to create customized HRTF for a desired source location.

2.2. *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. Modelling 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 [5].

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 N is the length, n is the current sample, o is the original HRIR, c is the candidate HRIR, e is the error and f is the percent fit. 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

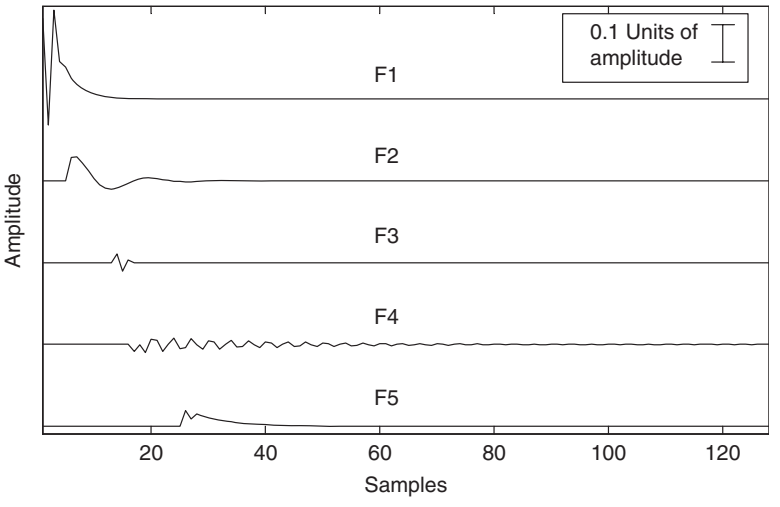

Figure 3. Five damped sinusoids obtained from a measured HRIR.

high accuracy. Figure 3 shows the components extracted from a measured HRIR by this process.

$$e(n) = o(n) - c(n) \tag{1}$$

$$f = 1 - \frac{(1/N)\sum_{n=1}^{N}(e(n)^2)}{(1/N)\sum_{n=1}^{N}(o(n)^2)} \tag{2}$$

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 window combinations, it will generate $9 \times 9 \times 9 \times 9 \times 9 = 59{,}049$ 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 may be needed to model late components that are not realized using only five windows. With five 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 five 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 three sample width was used in an attempt to approximate the entire signal. The approximation signals resulting from the STMCB (xs) and Prony (xp) methods are also shown in Figure 4. As can be seen in the results, xs and xp fail to properly reconstruct the original signal x. This would lead to inaccurate approximations of the parameters for the

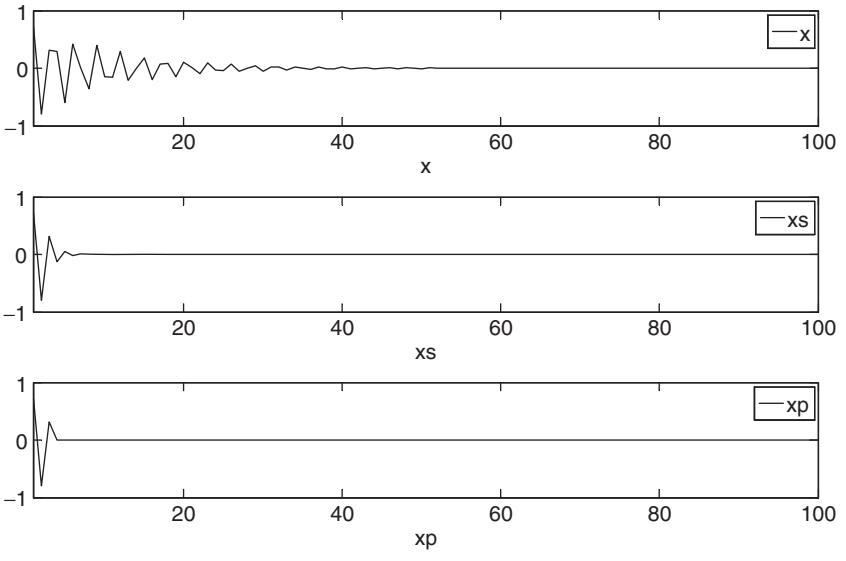

Figure 4. x (top) vs. xs (middle) and xp (bottom).

pinna model. These drawbacks have prompted us to develop a new, faster and potentially more robust method of HRIR decomposition into sequential damped sinusoids.

2.3. *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 (Figure 5). Hence, Equation 3 where k is a scalar and $p1$ and $p2$ 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.

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

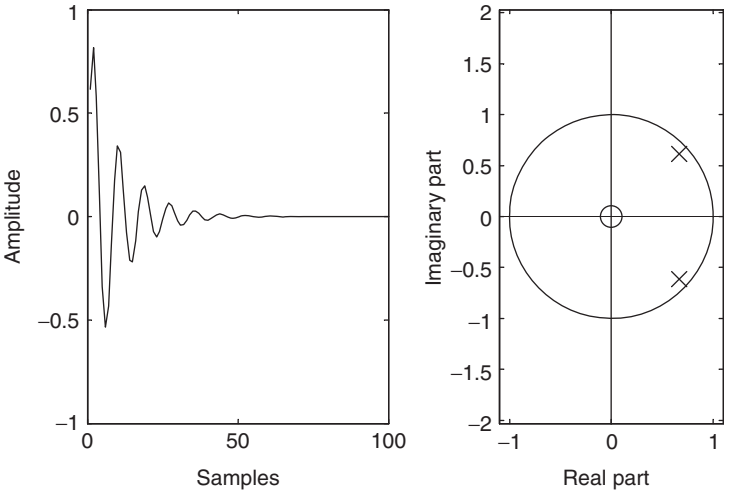

Figure 5. Time domain and zero-pole plot of a single second-order damped sinusoidal.

This new approach is also an iterative approach, 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 a modelling order that is two less than in the previous iteration.

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 five 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.

3. 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, $n = 0, \ldots, N-1$, d_i is the negative damping factor and ω_d is the digital frequency. Desired delays (τ_2 and τ_3) were applied to x2 and x3, respectively to obtain x2s and x3s. Finally x1, x2s and x3s were summed point-to-point resulting in the test signal, x. In this example $N = 100$, $\tau_2 = 3$, $\tau_3 = 6$, $\omega_d = 0.711$, $d_1 = -0.1$, $d_2 = -0.125$ and $d_3 = -0.15$. The three signals (x1, x2s and x3s) and the resulting signal (x) are shown in Figure 6.

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

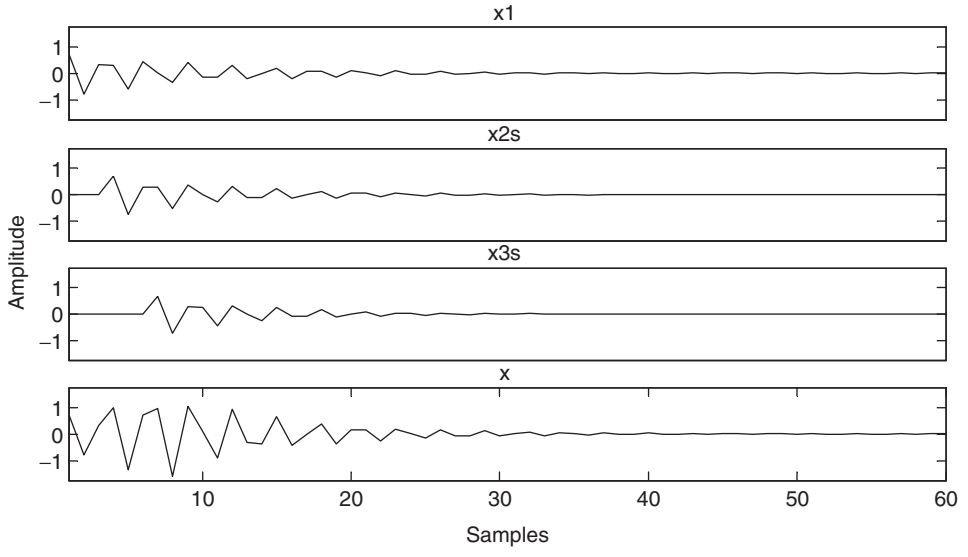

Figure 6. Plot of the three damped sinusoids (x1, x2s with delay τ_2 and x3s with delay τ_3) and the sum of them (x).

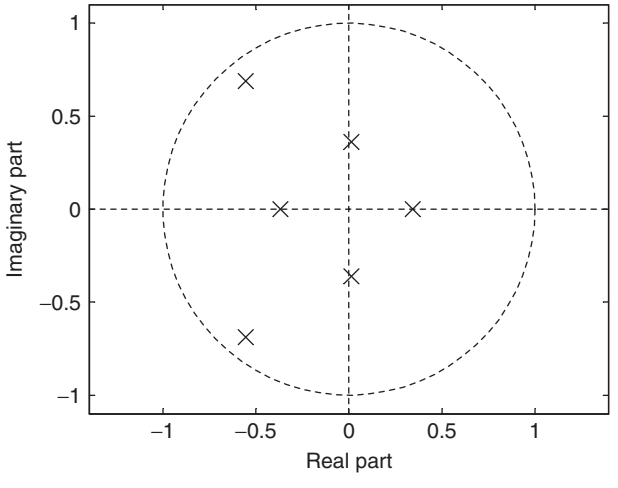

Figure 7. Poles obtained from the sixth-order STMCB approximation of the complete sequence x.

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.

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

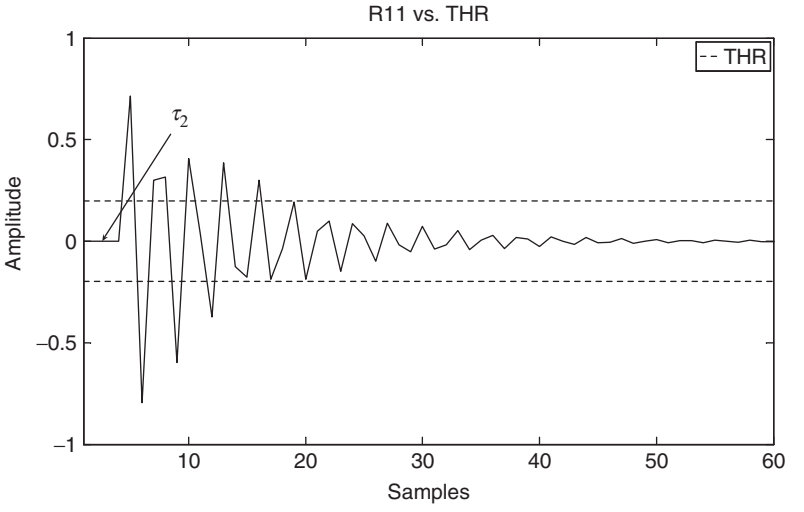

Figure 8. Plot of the first remnant with threshold lines (THR).

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 $q/2$ 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 τ_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 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 four 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.

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 x1, x2s and x3s very closely, as well.

4. 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 96 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 $+/-90^\circ$ 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.

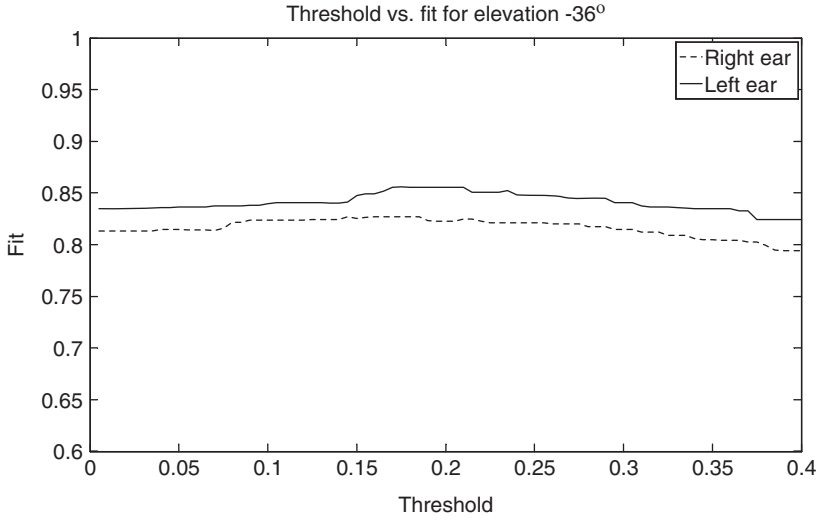

Figure 9. Threshold vs. fit for elevation -36°.

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 10–12 show examples of high, average and low fit cases for HRIRs using the 'new' method, respectively.

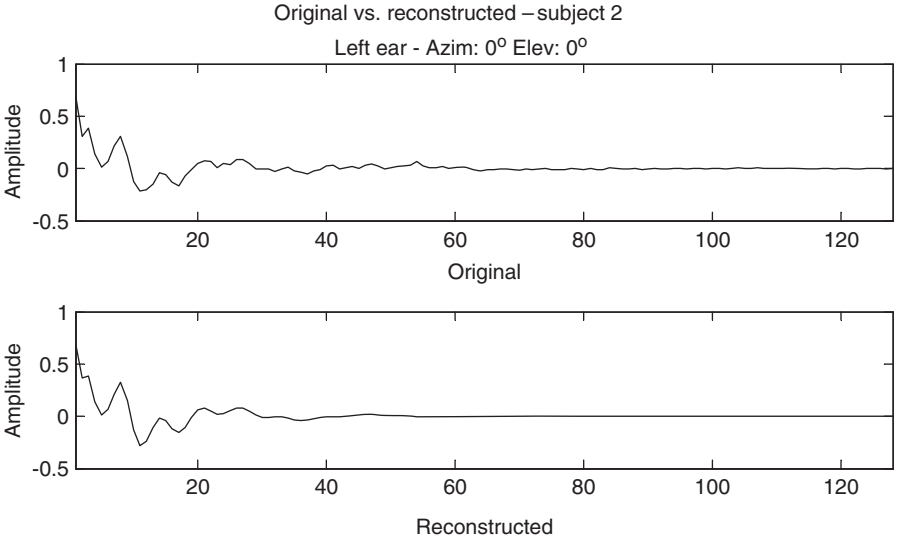

Figure 10. Original (top) vs. reconstructed HRIRs for the left ear of subject 2 for azimuth 0° and elevation 0° – highest fit example.

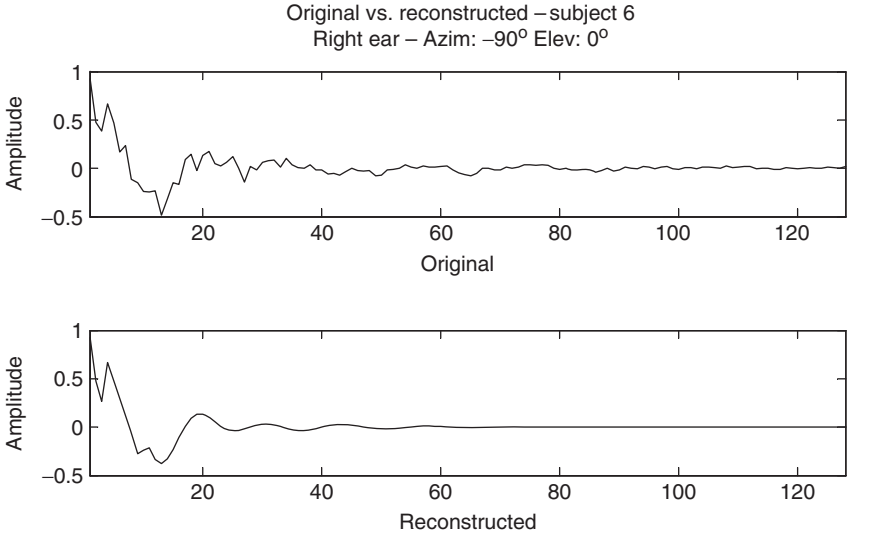

Figure 11. Original (top) vs. reconstructed HRIRs for the right ear of subject 6 for azimuth −90° and elevation 0° – average fit example.

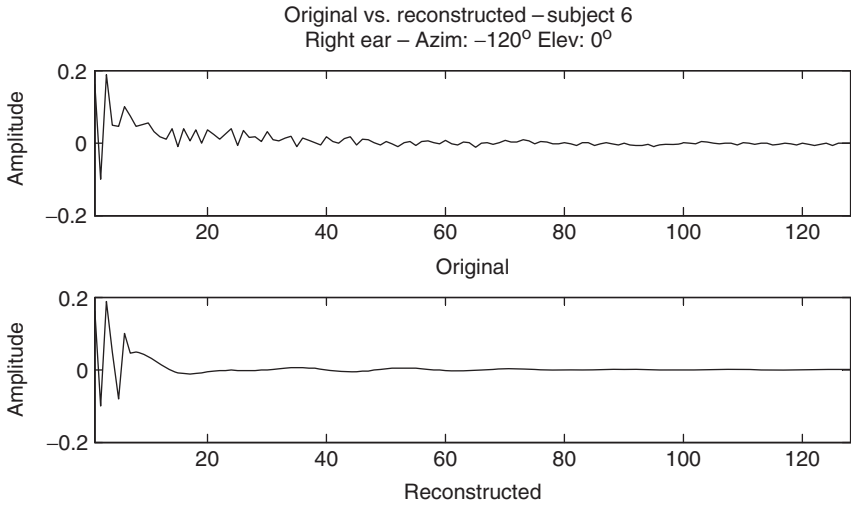

Figure 12. Original (top) vs. reconstructed HRIRs for the left ear of subject 6 for azimuth -120° and elevation 0° – lowest fit example.

5. Conclusion

Although the 'old' method achieved a 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 modelling 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 s compared to 4.2 s).

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 the relationship between model parameters and the anatomical characteristic of the intended listener that we ultimately seek.

Acknowledgement

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

References

[1] V. Algazi, C. Avendano, and R. Duda, *Estimation of a spherical-head model from anthropometry*, J. Audio Eng. Soc. 49 (2001), pp. 472–479.
[2] V. Algazi, R. Duda, D. Thompson, and C. Avendano, *The Cipic HRTF database 2001* IEEE Workshop on Applications of Signal Processing to Audio and Acoustics New Paltz, NY, 2001.
[3] A. Barreto and N. Gupta, *Dynamic modeling of the pinna for audio spatialization*, WSEAS Trans. Acoust. Music 1 (2004), pp. 77–82.

[4] C.P. Brown and R.O. Duda, *A structural model for binaural sound synthesis*, IEEE T. Speech Audi. P. 6 (1998), pp. 476–488.
[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 Trans. Signal Processing. 1 (2005), pp. 354–361.
[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., Springer, Netherlands, 2006, pp. 131–136.
[7] B. Gardner, K. Martin and Massachusetts Institute of Technology. Media Laboratory. *Vision and Modeling Group. HRTF measurements of a KEMAR dummy-head microphone. Vision and Modeling Group*, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1994.
[8] E.M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman, *Localization using nonindividualized head-related transfer-functions*, J. Acoust. Soc. Am. 94 (1993), pp. 111–123.