

## CAPTURING AND RECREATING AUDITORY VIRTUAL REALITY

R. DURAISWAMI, D. N. ZOTKIN, N. A. GUMEROV and A. E. O'DONOVANj

*Perceptual Interfaces & Reality Lab., Computer Science & UMIACS*

*Univ. of Maryland, College Park*

*E-mail: {ramani,dz,gumerov}@umiacs.umd.edu, adamod@gmail.com*

Reproduction of auditory scenes is important for many applications. We describe several contributions to the capture and recreation of spatial audio that have been made over the past few years by the authors.

*Keywords:* Surrounding loudspeaker array; Auralization; Higher-order Ambisonics; Computational room acoustics

### 1. Introduction

The recreation of an auditory environment in virtual reality, augmented reality, or in remote telepresence is a problem of interest in many application domains. It involves first the capture of sound, and second its subsequent reproduction, in a manner that can fool the perceptual system to believe that the rendered sound is actually where the application requires it to be and consequently to make the user feel as if they are present in the real audio scene. Our natural sound processing abilities, such as acoustic source localization, selective attention to one stream out of many, and event detection, are often taken for granted; however, the percept of the spatial location of a source arises from many cues, including those that arise due to the scattering of the sound off the listeners. Accurate simulation of those scattering cues is necessary for the spatial audio reproduction task. We review the principles for virtual auditory space synthesis, describe problems that have to be solved in order for it to be convincing, and outline solutions available, including some developed in our lab. Numerous applications are possible, including audio user interfaces, remote collaboration, scene analysis, remote education and training, entertainment, surveillance, and others.

## 2. 3D Audio Reproduction

The goal of audio reproduction is to create realistic three-dimensional audio so that users perceive acoustic sources as external to them and located at the correct places.<sup>12</sup> One obvious solution is to place arrays of loudspeakers and just play the sounds from wherever they are supposed to be, panning as appropriate to permit sound placement in-between, e.g., as in.<sup>1</sup> Such a setup is obviously cumbersome, expensive, non-portable, and noisy to other people in the same environment, and is only used in special environments where these issues are not of concern. The following discussion assumes that the synthesis is done over headphones. Three building blocks of the virtual auditory reproduction system<sup>2</sup> are: head-related transfer function based filtering, reverberation simulation, and user motion tracking.

**Head-related transfer function:** Sound propagating through space interacts with the head and body of the listener, causing the wave reaching the ear to be modified from that was emitted at the source. Furthermore, due to the geometry of human bodies and in particular to the complex shape of the pinna, changes in sound spectrum depend greatly on the direction from which the sound arrive, and to a lesser extent on range. These spectral cues are linked to the perception of sound direction and are characterized as the head-related transfer function (HRTF) – the ratio between the sound spectrum at the ear and that at the source.<sup>7</sup> These are the filters to be applied to transform the original sound source to the perceived one.

Inter-personal differences in body geometry make the HRTF substantially different among individuals.<sup>9</sup> For accurate scene reproduction the HRTF can be measured for each participant, which may be time-consuming. Various methods of selecting the HRTF from a pre-existing database that “best-fits” individual in some sense were tried by researchers with matching either physical parameters (e.g. from a ear picture<sup>2</sup>) or perceptual experience (e.g., by asking which audio piece sounds closest to being overhead<sup>5</sup>). Successful attempts of computing HRTF using numerical methods on head/ear mesh have also been reported, though they usually require very significant computation time<sup>15,17</sup>

**Reverberation Simulation:** In addition to the direct sound from the source, we hear multiple reflections from boundaries of the environment and objects in it, termed reverberation. From these we are able to infer the room size, wall properties, and perhaps our position in the room. Each reflection is not heard separately, but rather they are perceptually joined in one auditory stream. The perception of reverberation is complicated; roughly speaking, the energy decay rate in the reverberation tail implies

room size, and relative magnitude at various frequencies is influenced by materials. The ratio of direct to reverberant sound energy is used to judge distance to the source. Hence, reverberation is a cue that is very important to perception of sound as being external to the listener. Also, the reflections that constitute the reverberation are directional (i.e., arrive to the listener from specific directions just like the direct sound) and should be presented as directional in an auditory display at least for the first few reflections.

**User Motion Tracking:** A key feature of any acoustic scene that is external to the listener is that it is stationary with respect to the environment. That is, when the listener rotates, auditory objects move in the listener-bound frame so as to “undo” the listener’s rotation. If this is not done, listeners subconsciously assume that the sound’s origin is in their head (as with ordinary headphone listening), and externalization becomes very hard. Active tracking of the user’s head and adjustment of the auditory stream on the fly to compensate for the head motion/rotation is necessary. A tracker providing at least the head orientation in space is needed for virtual audio rendering. The stream adjustment must be made with sufficiently low latency so as not to create dissonance between the expected and perceived streams.<sup>8</sup>

Note that all these three problems do not exist if the auditory scene is rendered using loudspeaker array. Indeed, HRTF filtering for the correct direction and with correct HRTF is done by the mere act of listening; room reverberation is added naturally, given that the array is placed in some environment; and motion tracking is also unnecessary because in this case the sources are in fact external to the user and obviously shift when the user rotates. In essence, the loudspeaker array setup attempts to represent the “original” scene – with sources physically surrounding the listener. However, such a setup is expensive/non-portable and can be used only in large facilities, and may still find it difficult to reproduce specific scenes different from the environment in which the reproduction is being done.

### **2.1. *Signal Processing Pipeline***

We describe here the typical signal processing pipeline involved in virtual auditory scene rendering. The goal is to render several acoustic sources at certain positions in a simulated environment with given parameters (room dimensions, wall materials, etc.) Assume that the clean (reverberation-free) signal to be rendered for each source is given and that the HRTF of the listener is known. For each source, the following sequence of steps is done and the results are summed up together to produce auditory stream.

- Using source location and room dimensions, compute reflections (up to a certain order) using simple geometry.
- Obtain current user head position/rotation from the headtracker.
- Compute locations of the acoustic source and of its reflections in listener-bound coordinate system.
- Compose impulse responses (IR) for left and right ears by combining the HRIRs for the source and reflection directions.
- Convolve the source signal with left/right ear IR using frequency-domain convolution.
- Render the resulting audio frame to the listener and repeat.

In practice, due to limited computational resources available, reflections up to a certain order are computed in real time; the remaining reverberation tail is computed in advance and is appended to the IR unchanged. Also, artifacts such as clicks can arise at the frame boundaries due to the IR filter being changed significantly between frames. Usual techniques such as frame overlapping with smooth fade-in/fade-out windows can be used to eliminate artifacts.

### 3. Audio System Personalization

As mentioned before, the accuracy of the spatial audio rendering depends heavily on the knowledge of the HRTF for the particular user of the rendering system. The rendering quality and the localization accuracy degrade greatly when “generic” HRTF is used.<sup>9</sup> Therefore, it is necessary to acquire the HRTF for each user of the headphone spatial audio rendering system. Such acquisition can be done in several different ways, including direct HRTF measurement, selection of “best-fitting” HRTF from a database based on anthropometry, or HRTF computation using numerical methods on the pinna-head-torso mesh. Some relevant work done in the author’s lab is described below.

#### 3.1. *Fast HRTF Measurement*

We seek to measure the HRTF  $H_l(k; s)$  defined as

$$H_l(k; s) = \frac{\Psi_l(k; s)}{\Psi_c(k)}, \quad H_r(k; s) = \frac{\Psi_r(k; s)}{\Psi_c(k)},$$

where the signal spectrum at the ear is  $\Psi_l(k; s)$  and at the center of the head in the listener’s absence is  $\Psi_c(k)$ ; indices  $l$  and  $r$  signify left/right ear. Here,  $s$  is the direction to the source and we use wavenumber  $k$  in place

of the frequency  $f$  for notational convenience; note that  $k = 2\pi f/c$ . The traditional method of measuring the HRTF (e.g.,<sup>10</sup>) is to position a sound source (e.g., a loudspeaker) in the direction  $s$  and play the test signal  $x(t)$ . Microphones in the ears record the received versions of the test signal  $y(t)$ , which in the ideal case is  $x(t)$  filtered with the head-related impulse response (HRIR) – the inverse Fourier transform of  $H_{l,r}(k; s)$ . Measurements are done on a grid of directions to sample the continuous HRTF at a sufficient number of locations. The procedure is slow because the test sounds have to be produced sequentially from many positions, with sufficient interval. This necessitates use of equipment to actually move the loudspeaker(s) between positions. As a result, measurement at an acceptable sampling density (about a thousand locations) requires an hour or more.

An alternative fast method was developed recently by us.<sup>3</sup> The method is based on Helmholtz' reciprocity principle.<sup>11</sup> Assume that in a linear time-invariant scene the positions of the source and the receiver are swapped. By reciprocity, the recording made would be the same in both cases. The fast HRTF measurement method thus places the source in the listener's ear and measures the received signal simultaneously at many microphones at a regular grid of directions. The time necessary for the measurement is thus reduced from hours to seconds.

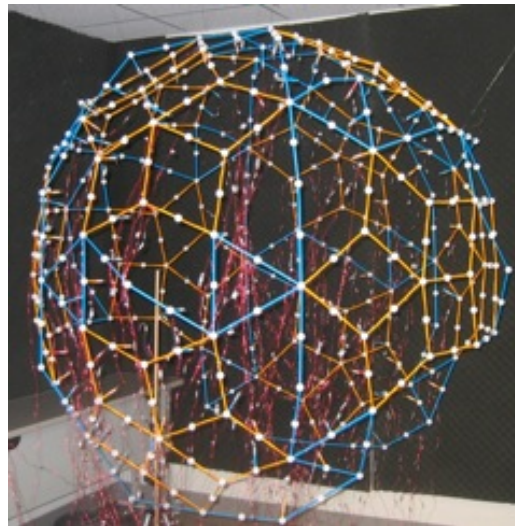


Fig. 1. A prototype fast HRTF measurement system based on the acoustic reciprocity principle. The diameter of the structure is about 1.4 meters.

The test signal  $x(t)$  is selected to be wideband (to provide enough energy in frequencies of interest) and short (to be able to window out equipment/wall reflections). In our case, the test signal is about 96 samples (2.45 ms long) up-sweep chirp. The sampling rate is 39062.5 Hz. The test signal is fed to the Knowles ED-9689 microspeaker, which is wrapped in a silicone plug so as to provide acoustic isolation and very safe sound levels, and is inserted into the subject's ear. 128 Knowles FG-3629 microphones are mounted at the nodes of an open sphere-like structure (Figure 1).

To measure the HRTF, a calibration signal is first obtained to compensate for individual differences in channels of the system. The speaker is placed at the center of the recording mesh, and a test signal recorded at all microphones as  $y_i^c(t)$ ;  $i$  is the microphone index. The recorded signal was windowed so as to exclude reflections from the room boundaries. The test signal  $x(t)$  is played through the microspeaker and recorded at  $i$ th microphone at  $s_i$  as  $y_i(t)$ . This is repeated 48 times, and the resulting signal is time-averaged to improve SNR. The same thresholding and windowing is used on  $y_i(t)$ . The estimated HRTF  $H_i(k; s_i)$  is then determined as  $Y_i(k)/Y_i^c(k)$ , where capitals signify Fourier transform. The original signal  $x(t)$  is not used in calculations but is present implicitly in  $y_i^c(t)$  modified by channel responses.

The computed HRTF is windowed in the frequency domain with a trapezoidal window in order to dampen it to zero at the frequencies where measurements cannot be made reliably. In particular, the microspeaker is inefficient at low frequencies ( $\lesssim 700$  Hz), and there is no energy in the test signal above 14 kHz. The HRTF in the very low / very high frequency ranges is gradually tapered to zero. Inverse Fourier transform is used to generate HRIR, which is made minimum phase with appropriate time delays obtained from thresholding  $y_i(t)$ .

### 3.2. *HRTF Approximation Using Anthropometric Measurements*

The HRTF is a transfer function describing filtering of the sound source by the listener's anatomy, and HRTF features (such as ridges and valleys in the spectrum) are conceivably created by sound scattering in the outer ear and by the head and torso. It is therefore reasonable to assume that listeners having anatomical features "similar" to the certain extent would have similar HRTFs as well. Some authors (e.g.,<sup>18</sup>) have looked into this problem by building an HRTF model as a function of certain morphological measurements. However, the analysis was done on a limited number

of subjects. Later, several HRTF databases became available. One of them is the CIPIC database,<sup>10</sup> which includes certain anthropometric measurements in addition to the HRTF data for 45 subjects. We explored a simple way of spatial audio rendering system customization by finding the “best-matching” database subject in terms of ear parameters and then further amending the HRTF in accordance with head and torso parameters.<sup>19</sup>

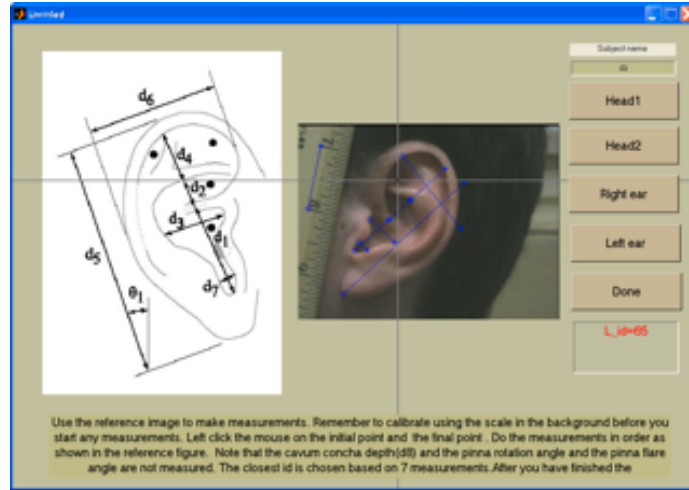


Fig. 2. A screenshot of our HRTF customization software.

Figure 2 shows the main interface of the system. On the left side, the reference image is shown with the seven ear dimensions  $d_1, \dots, d_7$  identified (they are, in order, cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, and intertragal incisure width). The reference image is used to guide the operator in process of measuring these dimensions in the picture of real ear. On the right side, the image of the user’s ear is acquired using a digital camera, along with a ruler to provide a scale. The operator identifies and marks feature points on the ear as shown. If  $\hat{d}_i$  is the value of the  $i^{th}$  parameter in the image, and  $d_i^k$  is the value of the same parameter for the  $k^{th}$  subject in the database, then the matching is performed by minimizing the error measure  $E^k$ :

$$E^k = \sum_{i=1}^7 \frac{(\hat{d}_i - d_i^k)^2}{\sigma_i^2}.$$

Here  $\sigma_i^2$  is the variance of the  $i^{th}$  parameter across all subjects in the database. Subject  $k$ ,  $k = \arg \min_k E^k$ , is chosen to be the best match to the user. In the case shown, it is the subject with ID 45. Note that matching is done separately for left and right ears, which sometimes results in different best-matching subjects for left and right ears because of asymmetries in individual anatomy.

The HRTF of the best-matching subject is further refined using the head-and-torso (HAT) HRTF model.<sup>20</sup> This is a three-parameter (head radius, torso radius, and neck height) analytical model for computation of HRTF at low frequencies (below approximately 3-4 kHz). The HAT algorithm models head and torso as two separated spheres and simulates the wave propagation path(s) for various source positions. Note that contributions to the HRTF caused by the torso, the head, and the pinnae are more or less separated on the frequency axis; in particular, pinna-induced features are generally located above 4-6 kHz. Therefore, it can be assumed that the database matching method would produce reasonably matching HRTF at higher frequencies only, as only the pinna features are used for matching. To compensate for that, we take another frontal photograph of the subject, measure three HAT parameters from it, compute HAT model HRTF, and blend it with the database HRTF, using HAT model HRTF exclusively below 0.5 kHz, progressively blending in database HRTF and blending out HAT model HRTF between 0.5 kHz and 3 kHz, and using database HRTF exclusively above 3 kHz.

The localization accuracy was tested on eight subjects in two different conditions (with continuously repeating 1-second white noise bursts and with just one noise burst and silence afterwards). In both tasks, subjects were asked to point to the perceived location of the sound source with their nose. A head tracking unit (Polhemus Fastrack) was used to measure the pointing direction, and error was computed as an angle between the selected direction and the true one. It was generally found that incorporation of the HAT model consistently improves the localization accuracy for all subjects (by about 25% on average). Subjectively the HAT-processed HRTF also improves the quality of the scene (subjects describe the rendered sound as having “depth”, being more “focused”, and being more “stable”). The results obtained with ear-matching personalization were mixed, with some subjects showing limited improvement and others showing no changes. Based on the experiments, the HAT model is routinely incorporated into HRTF-based experiments in our lab. More detailed description of the algorithms, protocols, experiments, and results is available in.<sup>19</sup>

### 3.3. HRTF Analysis and Feature Extraction

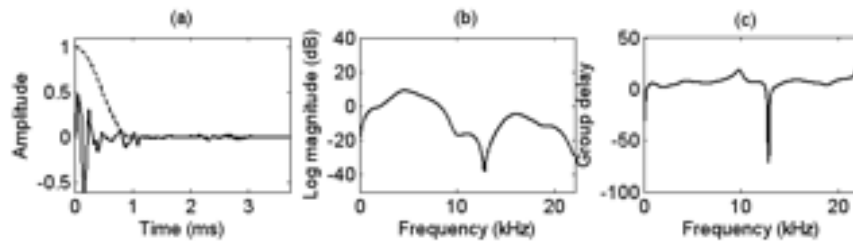


Fig. 3. An illustration to the HRTF feature extraction algorithm. a) HRIR (solid line) and 1.0 ms half-Hann window (dashed line). b) The spectrum of the windowed signal. c) The corresponding group-delay function.

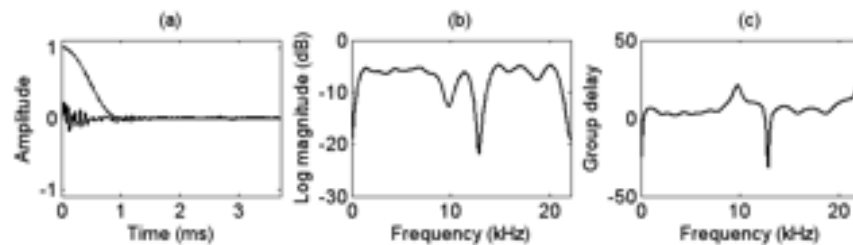


Fig. 4. Same as Figure 3 but for HRIR LP residual as an input instead of original HRIR.

To further advance towards reliable HRTF personalization method that does not require performing HRTF measurement, we have developed a method for extracting prominent features from HRTF and relating them to ear geometry.<sup>21</sup> A motivation for our work was that while many authors (e.g.,<sup>22</sup>) provide models for composing HRTF from known anthropometry, little work has been done in the opposite direction – i.e., to analyze the measured HRTF of a subject and decompose it into components.

The HRTF analysis method is a combination of several signal processing algorithms designed to extract HRTF notches caused by pinna and distinguish them from features caused by other body parts (e.g., shoulder reflection). In order to reject effects caused by body parts other than pinna, the HRTF is first converted to HRIR using an inverse Fourier transform

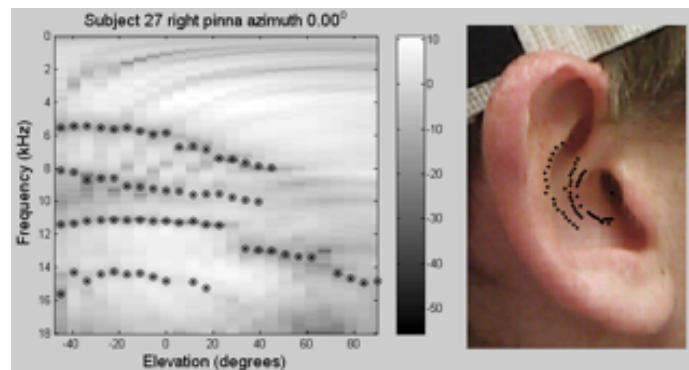


Fig. 5. The spectral notch frequencies extracted from HRTF measurements of CIPIC database subject 27 right pinna (left) are converted to ear dimensions and are marked on the pinna image of the same subject (right).

and then windowed using a half Hann window with the window length of 1.0 ms starting at the signal onset as shown in Figure 3(a). This particular window length was used to eliminate torso reflection (seen at around 1.6 ms) and knee reflection (at 3.2 ms). The spectrum of the windowed signal is shown in Figure 3(b). The notch is defined here more clearly than on the spectrum of the unwindowed signal (plot not shown). For further resolution enhancement, we extract the spectral notches from the group-delay function rather than from the magnitude spectrum, which was shown to be beneficial even for short data frames.<sup>23</sup> The group delay function is the negative of the derivate of the signal phase spectrum and can be computed directly via a combination of two Fourier transforms. It is shown in Figure 3(c) and clearly has a better resolution than a simple spectrum.

To reduce artifacts caused by windowing operation and remove broad HRTF features caused by resonances in the pinna cavities (and thus to better isolate narrow HRTF notches), we apply the 12th order linear prediction (LP) analysis to the original HRIR and then use the LP residual instead of the original HRIR as an input to the steps described above. Figure 4 shows the results of the processing of HRIR LP residual. It can be seen that the spectrum is flat now, allowing for higher accuracy in identifying the positions of the notches.

To verify the method, the notch frequencies were extracted for HRTFs of several subjects of the CIPIC HRTF database. A representative example is shown in Figure 5. On the left, the positions of prominent notches in the spectrum are marked. Note that each notch position traces a certain

contour as sound source elevation changes. On the right, the extracted notch frequencies are projected onto the pinna photograph assuming that the notch frequency corresponds to the first anti-resonance between incoming wave and reflected wave. The contours obtained on the pinna photograph clearly trace the outlines of anatomical features of the ear, showing that the results are meaningful and that the shape of spectral contours is indeed related to fine details of the pinna shape. The method could potentially be used for HRTF personalization, either by HRTF synthesis “from scratch” based on pinna measurements or by modifying HRTF of one subject to match another subject in accordance with differences in ear geometry. A much more detailed description of the algorithm and a large number of additional examples are available in.<sup>21</sup>

### 3.4. Numerical HRTF Computation

Yet another alternative to measuring HRTF directly is to compute it using numerical methods. This requires one to have reasonably accurate representation of the surface of the subject’s body, including head, torso, and fine pinna structure. The representation considered is usually a triangular surface mesh. The necessary mesh resolution is determined by the highest frequencies for which the HRTF is to be computed; the rule of thumb for most numerical methods is to have at least 5-6 mesh elements per wavelength. A fine discretization with very large number of elements is therefore necessary for upper frequency hearing limit of 20 kHz (wavelength of 1.7 cm). For faster convergence, it is also desirable that the mesh consists of mostly close-to-equilateral triangle patches (as opposed to narrow, elongated triangles).

The acoustic potential  $\Phi(k; r)$  at a wavenumber  $k$  in any volume that does not enclose the acoustic sources must conform to the Helmholtz (wave) equation with the Sommerfeld radiation condition

$$\nabla^2 \Phi(k; r) + k^2 \Phi(k; r) = 0, \quad \lim_{r \rightarrow \infty} \left[ r \left( \frac{\partial \Phi(k; r)}{\partial r} - ik \Phi(k; r) \right) \right] = 0,$$

where  $r$  is a radius-vector of an arbitrary point within the volume. Assume that an object (a head mesh, or head-and-torso mesh) is located within the volume and is irradiated by a plane wave  $e^{ikr \cdot s}$  propagating in a direction  $s$ . Denote by  $S$  the boundary of the object. For simplicity, the sound-hard boundary condition

$$\left. \frac{\partial \Phi(k; r)}{\partial n} \right|_S = 0$$

is usually assumed, although impedance boundary condition can be stipulated as well. Any boundary element method can be used to iteratively solve the wave equation in presence of the irradiating field and to obtain the set of potentials  $\Phi_m(k; r)$  at all surface elements  $S_m$  (e.g.,<sup>24</sup>). Note that by definition the HRTF is nothing but a value of the potential at the surface element corresponding to the ear canal entrance and is immediately available once the wave equation is solved.

However, such computation method is extremely wasteful. Indeed, the result of computation is the set of potentials for all  $S_m$  while only one value is ultimately needed (however, all the values are necessary for iterative solution procedure). Moreover, all computations have to be re-done from scratch for another direction  $s$ .

Similarly to how the physical HRTF measurement method can be sped up, the numerical computations can also be made several orders of magnitude faster using reciprocity principle. The same object mesh is used; however, the monopole source is placed directly on the boundary of the object at the place corresponding to the ear canal entrance, just like the loudspeaker is placed in the reciprocal HRTF measurement method. The BEM computations are then done to compute the acoustic field around the object, and the computed field is sampled at points corresponding to the locations of microphones in the reciprocal HRTF measurement method. An example paper using this technique is.<sup>25</sup>

However, the processing power required for direct implementation of BEM is extremely high. For reference, in<sup>25</sup> a head mesh with about 22000 elements is used, and the HRTF computations are done only up to the frequency of 5.4 kHz; still, it took about 28 hours to perform computations for one frequency only. Admittedly,<sup>25</sup> is eight years old; however, even with the present level of technology it is impossible to get up to 20 kHz because much finer mesh should be used at higher frequencies, and the BEM complexity grows as  $O(N^6)$ .

Based on our work in,<sup>26</sup> we have developed a method for the numerical HRTF computation to be done several orders of magnitude faster than other existing work. The key point in our work is to use fast multipole methods (FMM) to speed up the iterative process of computing the acoustic potential on the mesh.<sup>17</sup> Usually, the field computations are done pairwise for all mesh patches at each iteration, resulting in  $O(N^2)$  cost per iteration. In FMM, the patches are grouped into sets according to certain rules, and for each set the field produced by all patches is computed at  $O(N)$  cost. The computed fields are then applied to all patches also at  $O(N)$  cost.

The total running time for the FMM BEM solver scales approximately as  $O(N^{1+\alpha})$ , for some  $\alpha < 0.5$ .

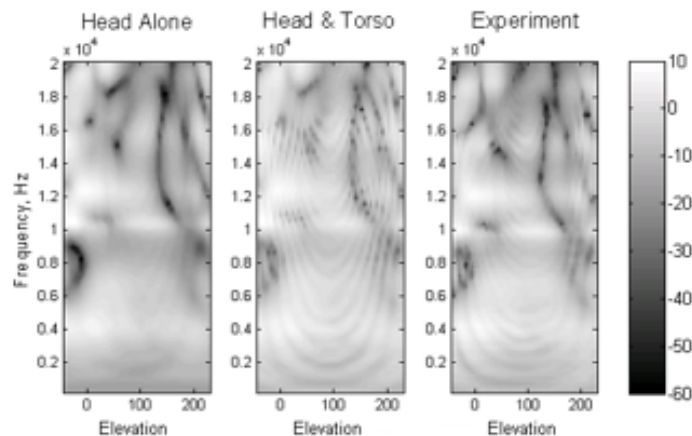


Fig. 6. BEM-obtained HRTF (in dB) for the KEMAR head mesh alone. Middle: BEM-obtained HRTF for the KEMAR head and torso mesh. Right: Experimentally measured KEMAR HRTF data. In all plots, azimuth is zero and elevation is varying from -45 to 225 degrees. “Large” KEMAR pinna is used.

We have computed the HRTF of the KEMAR manikin for which a lot of experimental measurements are available from different researchers. Our KEMAR mesh includes the torso and has approximately 445000 elements in total, and on a high-end desktop workstation (QX6700 2.66 GHz Intel CPU, 8 Gb RAM) we are able to compute the HRTF for one frequency in 35 minutes on average (total computation time for 117 frequencies spanning from 172 Hz to 20.155 kHz was 70 hours). We have performed comparison of our results with experimentally-measured KEMAR data and found very good agreement. One example is shown in Figure 6, where the contours of the HRTF features observed in the experimental data match those obtained in numerical simulations very well. The pinna mesh for this particular computation was obtained using a CT scan. We had also attempted to numerically compute HRTF using head/pinna meshes obtained from a 3-D laser scanner device and generally found that the agreement between computed and experimental HRTF is significantly worse, which is due to the fact that the laser scan is not able to reproduce the surface with required fidelity due to loss of features in concave areas. Obviously CT scan is not a good choice for pinna shape acquisition with human subjects be-

cause of associated time, cost, and radiation exposure concerns; perhaps a laser-scanned mesh can be made acceptable for BEM HRTF computation by stitching several meshes scanned at different angles. The research on obtaining quality meshes is ongoing.

#### 4. Auditory Scene Capture and Reproduction

We wish to perform a multi-channel audio recording in a way so that the spatially/temporally removed listener can perceive the acoustic scene to be fully the same as if present at the recording location, so that sound sources are perceived to be in the correct directions and distances, environmental information is kept, and the acoustic scene stays external with respect to the listener when the listener moves/rotates. In order to do that, such recording should capture the spatial structure of the acoustic field. Indeed, a single-microphone recording could faithfully reproduce the acoustic signal at a given point; but from such a recording it is obviously impossible to tell which parts of audio scene arrive from which directions. Also, the information about spatial structure of the acoustic field is necessary to allow for listener motion. A microphone array that is able to sample such spatial structure is therefore necessary.

An important distinction from the VAS rendering system described above here is that in VAS rendering the synthesis is done “from scratch” by placing sound sources for which clean signals are available at known places in known environment and then artificially adding environmental effects. In scene capture and rendering application, the scene structure information is unknown. A tempting approach is to localize all sound sources, isolate them somehow, and then render them using VAS; however, localization attempts with multiple active sources (such as at a concert) are futile; moreover, such approaches would remove all but the most prominent components from the scene, which contradicts the goal of keeping the environmental information intact. Therefore, scene capture and reproduction system attempt to blindly re-create the scene without trying to analyze it. More specifically, it operates by decomposing the acoustic scene into pieces that come from various directions and then renders those pieces to the listener to appear to come from these directions.<sup>4</sup> The grid of these directions is fixed in advance and is data-independent. The placement of sources would be reproduced reasonably well with such a model.

Spherical microphone arrays constitute an ideal platform for such system due to their inherent omnidirectionality. A 64-microphone array is shown in Figure 7. Assume that the array has  $L_q$  microphones located



Fig. 7. A prototype 64-microphone array with embedded video camera. The array connects to the PC via a USB 2.0 cable streaming digital data in real-time.

at positions  $s'_q$  on the surface of a hard sphere of radius  $a$ . Assume that the time-domain signal measured at each microphone is  $x(t; s'_q)$ . Perform Fourier transform of those signals to obtain the potentials  $\Psi(k; s'_q)$  at all microphones. The goal is to decompose an acoustic scene into waves  $y(t; s_j)$  arriving from various directions  $s_j$  (the total number of directions is  $L_j$ ) and then render them for the listener from directions  $\tilde{s}_j$  (which are the directions  $s_j$  converted to the listener-bound coordinate system) using his/her HRTF  $H_l(k; s)$  and  $H_r(k; s)$ . It is easier to work in the frequency domain. Let us denote the Fourier transform of  $y(t; s_j)$  by  $\lambda(k; s_j)$ . Further, denote by  $\Psi$  the  $L_q \times 1$  vector of  $\Psi(k; s'_q)$ , and denote by  $\Lambda$  the  $L_j \times 1$  vector of  $\lambda(k; s_j)$ . We want to build a simple linear mechanism for decomposition of the scene into the directional components so that it can be expressed by equation  $\Lambda = W\Psi$ , where  $W$  is the  $L_j \times L_q$  matrix of decomposition weights constructed in some way.

One way to proceed is to note that the operation that allows for selection of a part of audio scene that arrives from a specified direction is nothing but a beamforming operation (e.g.,<sup>13</sup>). A beamforming operation for direction  $s_j$  in frequency domain can be written as

$$\lambda(k; s_j) = -i \frac{(ka)^2}{4\pi} \sum_{n=0}^{p-1} (2n+1) i^{-n} h'_n(ka) \sum_{q=1}^{L_q} \Psi(k; s'_q) P_n(s_j \cdot s'_q),$$

where  $h'_n(ka)$  is the derivative of the spherical Hankel function of order  $n$

and  $P_n(\dots)$  is an associate Legendre polynomial of order  $n$  as well.  $p$  here is the truncation number and should be kept about  $ka$ ; increasing it improves the selectivity but also greatly decreases robustness, so a balance must be kept. This decomposition way can be expressed in the desired matrix-vector multiplication framework as

$$w(k; s_j, s'_q) = -i \frac{(ka)}{4\pi} \sum_{n=0}^{p-1} (2n+1) i^{-n} h'_n(ka) P_n(s_j \cdot s'_q).$$

Another method of decomposition<sup>16</sup> is to note that a relationship inverse to the desired one can be established as  $\Psi = F\Lambda$ , where  $F$  is a  $L_q \times L_j$  matrix with elements  $f(k; s'_q, s_j)$  equal to the potential caused at microphone at  $s'_q$  by the wave arriving from the direction  $s_j$ . The equation for  $f(k; s'_q, s_j)$  is

$$f(k; s'_q, s_j) = \frac{i}{(ka)^2} \sum_{n=0}^{p-1} \frac{i^n (2n+1) P_n(s_j \cdot s'_q)}{h'_n(ka)}.$$

Then, a linear system  $\Psi = F\Lambda$  is solved for  $\Lambda$ , exactly if  $L_q = L_j$  or in least squares sense otherwise. For this method, the weight matrix  $W_L = F^{-1}$  (generalized inverse for non-square  $F$ ).

The decomposed acoustic scene is stored for later rendering and/or transmitted over network to the remote listening location. When rendering the scene for the specific user, his/her HRTF are assumed to be known. During the rendering, the directions  $s_j$  with respect to the original microphone array position are converted to directions  $\tilde{s}_j$  with respect to the listener-bound coordinate frame, and the audio streams  $y_l(t)$ ,  $y_r(t)$  for left/right ears are obtained as<sup>14</sup>

$$Y_{l,r}(k) = \sum_{j=1}^{L_j} H_{l,r}(k; \tilde{s}_j) \lambda(k; s_j), \quad y_{l,r}(t) = IFT[Y_{l,r}(k)](t),$$

where IFT is the inverse Fourier transform. Essentially the scene components arriving from different directions are filtered with their corresponding HRTF taking into account listener head position/orientation and are summed up. The above description is a brief outline of auditory capture and reproduction principles, and many practical details are omitted for brevity and are available in e.g.<sup>27</sup> or.<sup>16</sup>

## 5. Sample Application: Audio Camera

One particularly promising application of spherical microphone array signal processing is the real-time visualization of the acoustic field energy. The

microphone array immersed in the acoustic field has the ability to analyze the directional properties of the field. The same spherical microphone array is used in this application as well, but the goal and the processing means are different. The goal is to generate an image that can be used by a human operator to visualize the spatial distribution of the acoustic energy in the space. Such image can further be spatially registered with and overlaid on the photograph of the environment so that the acoustically active areas can be instantly identified.<sup>28</sup>

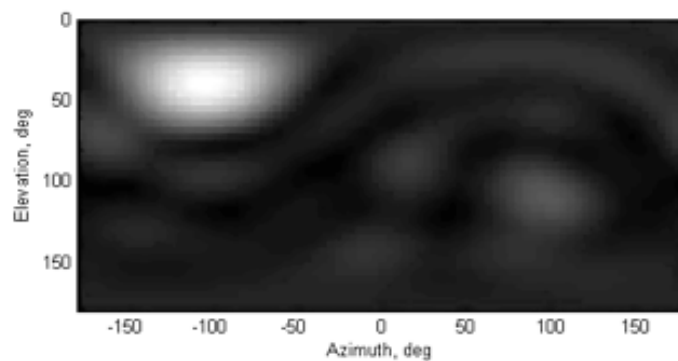


Fig. 8. Sample acoustic energy map. Map dimensions are 128 by 64 pixels. Pixel intensity is proportional to the energy in the signal beamformed to the direction indicated on axes.

To create the acoustic energy map, the beamforming operation is performed on a relatively dense grid of directions. The energy map is just a regular, two-dimensional image with the azimuth corresponding to X axis and the elevation corresponding to the Y axis. In this way, the array's full omnidirectional field of view is unwrapped and is mapped to the planar image. The energy in the signal obtained when beamforming to the certain direction is plotted as the pixel intensity for that direction. A sample map obtained with the array of Figure 7 is shown in Figure 8. It can be seen in the image that an active acoustic source exists in the scene and is located approximately in the direction  $(-100, 40)$  (azimuth, elevation). In the same way, multiple acoustic sources can be visualized, or a reflection coming from the particular direction in space can be seen. Furthermore, if the original acoustic signal is short (like a pulse), an evolution of the reverberant sound field can be followed by recording it and then playing it back in the slow motion. In such a playback, the arrivals of the original sound and then of its

reflections from various surfaces and objects in the room are literally visible painted over the places where the sound/reflections are coming from.<sup>30</sup> Such information is usually obtained indirectly in architectural acoustics for the purposes of adjusting the environment for better attendee experience, and a real-time visualization tool could prove very useful in this area. Furthermore, it has been shown that the audio energy visualization map is created according to the central projection equations – the same equations that regular (video) imaging devices conform to,<sup>29</sup> so the device has been tagged “audio camera”, and the vast body of work derived for video imaging (e.g., multi-camera calibration or epipolar constraints) can be applied to audio or multimodal (auditory plus traditional visual) imaging. The reader is encouraged to refer to the references mentioned herein for further “audio camera” algorithm details and examples.

## 6. Conclusion

Possible applications of systems that capture and recreate spatial audio are numerous; one can imagine, for example, such a system being placed in the performance space and transmitting the audio stream in real time to the interested listeners, enabling them to be there without actually being there. Similarly, a person might want to capture some auditory experience – e.g., a celebration – to store for later or to share with family and friends. Another application from a different area could involve a robot being sent to environments that are dangerous to or unreachable by humans and relaying back the spatial auditory experience to the operator as if he/she is being there. The auditory sensor can also be augmented with video stream to open further possibilities, and the research in spatial audio capture and synthesis is ongoing.

## References

1. V. Pulkki (2002). “Compensating displacement of amplitude-panned virtual sources”, Proc. 22th AES Conf., Espoo, Finland, 2002, pp. 186-195.
2. D. N. Zotkin, R. Duraiswami, and L. S. Davis (2004). “Rendering localized spatial audio in a virtual auditory space”, IEEE Transactions on Multimedia, vol. 6, pp. 553-564.
3. D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov (2006). “Fast head-related transfer function measurement via reciprocity”, J. Acoust. Soc. Am., vol. 120, pp. 2202-2215.
4. A. E. O’Donovan, D. N. Zotkin, and R. Duraiswami (2008). “Spherical microphone array based immersive audio scene rendering”, Proc. ICAD 2008, Paris, France.

5. P. Runkle, A. Yendiki, and G. Wakefield (2000). "Active sensory tuning for immersive spatialized audio", Proc. ICAD 2000, Atlanta, GA.
6. J. B. Allen and D. A. Berkeley (1979). "Image method for efficiently simulating small-room acoustics", J. Acoust. Soc. Am., vol. 65, pp. 943-950.
7. W. M. Hartmann (1999). "How we localize sound", Physics Today, November 1999, pp. 24-29.
8. N. F. Dixon and L. Spitz (1980). "The detection of auditory visual desynchrony", Perception, vol. 9, pp. 719-721.
9. E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman (1993). "Localization using non-individualized head-related transfer functions", J. Acoust. Soc. Am., vol. 94, pp. 111-123.
10. V. R. Algazi, R. O. Duda, D. P. Thompson, and C. Avendano (2001). "The CIPIC HRTF database", Proc. IEEE WASPAA 2001, New Paltz, NY, pp. 99-102.
11. P. M. Morse and K. U. Ingard (1968). "Theoretical Acoustics", Princeton Univ. Press, New Jersey.
12. C. Kyriakakis, P. Tsakalides, and T. Holman (1999). "Surrounded by sound: Immersive audio acquisition and rendering methods", IEEE Signal Processing Magazine, vol. 16, pp. 55-66.
13. J. Meyer and G. Elko (2002). "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield", Proc. IEEE ICASSP 2002, Orlando, FL, vol. 2, pp. 1781-1784.
14. R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis (2005). "High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues", Proc. AES 119th convention, New York, NY, preprint #6540.
15. M. Otani and S. Ise (2006). "Fast calculation system specialized for head-related transfer function based on boundary element method", J. Acoust. Soc. Am., vol. 119, pp. 2589-2598.
16. D. N. Zotkin, R. Duraiswami, and N. A. Gumerov (2009). "Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays", IEEE Transactions on Audio, Speech, and Language Processing, in press.
17. N. A. Gumerov, A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin (2009). "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation", J. Acoust. Soc. Am., in press.
18. C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile (2000). "Enabling individualized virtual auditory space using morphological measurements", Proc. First IEEE Pacific-Rim Conference on Multimedia, Sydney, Australia, pp. 235-238.
19. D. N. Zotkin, J. Hwang, R. Duraiswami, and L. S. Davis (2003). "HRTF personalization using anthropometric measurements", Proc. IEEE WASPAA 2003, New Paltz, NY, pp. 157-160.
20. V. R. Algazi, R. O. Duda, and D. M. Thompson (2002). "The use of head-and-torso models for improved spatial sound synthesis", Proc. AES 113th

- convention, Los Angeles, CA, preprint #5712.
21. V. C. Raykar, R. Duraiswami, and B. Yegnanarayana (2005). "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses", *J. Acoust. Soc. Am.*, vol. 118, pp. 364-374.
  22. V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson (2001). "Structural composition and decomposition of HRTFs", *Proc. IEEE WAS-PAA 2001*, New Paltz, NY, pp. 103-106.
  23. B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan (1984). "Significance of group delay functions in signal reconstruction from spectral magnitude or phase", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 610-623.
  24. T. Xiao and Q.-H. Liu (2003). "Finite difference computation of head-related transfer function for human hearing", *J. Acoust. Soc. Am.*, vol. 113, pp. 2434-2441.
  25. B. F. G. Katz (2001). "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation", *J. Acoust. Soc. Am.*, vol. 110, pp. 2440-2448.
  26. N. A. Gumerov and R. Duraiswami (2004). "Fast multipole methods for the Helmholtz equation in three dimensions", Elsevier Science, The Netherlands.
  27. B. Rafaely (2005). "Analysis and design of spherical microphone arrays", *IEEE Trans. Speech Audio Proc.*, vol. 13(1), pp. 135-143.
  28. A. E. O'Donovan, R. Duraiswami, and N. A. Gumerov (2007). "Real time capture of audio images and their use with video", *Proc. IEEE WASPAA 2007*, New Paltz, NY, pp. 10-13.
  29. A. E. O'Donovan, R. Duraiswami, and J. Neumann. "Microphone arrays as generalized cameras for integrated audio-visual processing", *Proc. IEEE CVPR 2007*, Minneapolis, MN.
  30. A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin (2008). "Imaging concert hall acoustics using visual and audio cameras", *Proc. IEEE ICASSP 2008*, Las Vegas, NV, April 2008, pp. 5284-5287.