EFFICIENT CONVERSION OF X.Y SURROUND SOUND CONTENT TO BINAURAL HEAD-TRACKED FORM FOR HRTF-ENABLED PLAYBACK

Dmitry N. Zotkin, Ramani Duraiswami, and Nail A. Gumerov

Perceptual Interfaces and Reality Laboratory
UMIACS, University of Maryland, College Park
{dz,ramani,gumerov}@umiacs.umd.edu

ABSTRACT

Binaural presentation of X.Y sound is usually performed using virtual audio principles – that is, by attempting to virtually reproduce the setup of the X+Y loudspeakers in the reference room configuration. The computational cost of such playback is linear in the number of channels in the X.Y setup. We present a novel scheme that computes, offline, a spatio-temporal representation of the sound field in the listening area and store it as a multipole expansion. During head-tracked playback, the binaural signal is obtained by evaluating the multipole expansion at the ear position corresponding to the current user pose, resulting in a fixed playback cost. The representation is further extended to incorporate individualized HRTFs at no additional cost. Simulation results are presented.

Index Terms— Acoustic fields, audio systems, binaural presentation, head-related transfer function, surround sound.

1. INTRODUCTION

State-of-the-art consumer technology for reproduction of music and movie soundtracks consists of six or more loudspeakers spatially arranged in accordance with Dolby Digital 5.1 or DTS standards. The sense of presence created can be compelling; however, a 5.1 surround sound system is obviously stationary and cannot be "taken with the user". On the other hand, there is a significant increase currently in the number of people who listen to music on various portable devices using headphones. It would be of interest to fully and accurately reproduce the impression obtained by the user in the real 5.1 setup, or in other X.Y setups, by means of binaural (headphone) delivery.

In a real environment, sound that reaches the ear of a person differs from the original source sound due to scattering of the sound off the environment and off the person. Of particular interest is the participation of the person's outer ear (pinna) in the scattering. Pinna asymmetry leads to changes in the spectrum of the sound at the ear that are dependent on the direction of the sound source. These changes, along with inter-aural time and level delays due to spatial separation of two ears and head shadowing, are responsible for our ability to localize sounds. All localization cues are encoded in the head-related transfer function, HRTF [1], which is the ratio of the Fourier transform of the sound at the left (or right) ear $\Phi_{l,r}(k, r, \theta, \varphi)$ to that of the sound at the head center location $\Phi_c(k)$ as if the listener were absent

$$H_{l,r}(k, r, \theta, \varphi) = \frac{\Phi_l(k, r, \theta, \varphi)}{\Phi_c(k)}. \tag{1}$$

Here (r, θ, φ) are the spherical coordinates of the source location and k is the wavenumber (related to frequency f as $kc = 2\pi f$, where c is the sound speed). When the source range is farther than several head diameters, the dependence on r is weak and is usually neglected.

In the reference X.Y setup, the sound emanating from each loudspeaker is filtered with the appropriate HRTFs for the speaker direction, and a sum of HRTF-filtered sounds is heard by each ear. Furthermore, in a real environment a multitude of reflections from the walls and from objects leads to an effect commonly known as reverberation, which gets filtered similarly. Finally, a listener in the reference setup may move and turn, thereby changing the positions of the speakers in a user-bound frame and therefore changing the received sound (so user tracking is necessary for binaural delivery). All these effects must be reproduced faithfully for binaural playback to recreate the impression of the physical X.Y setup.

One approach is to simulate the physical system [2] by processing each virtual source via a pipeline that includes computing the positions of early reflections for the source, retrieving HRTFs for speaker/reflection directions according to the user pose, adding a reverberation tail, and convolving the source signal with the thus-obtained system's impulse response corresponding to the loudspeaker. This is repeated for each source, and the results mixed and played over headphones. Very compelling results are achieved with one or two simulated sound sources; however, the method's computational cost grows with the number of sources.

An alternative approach is taken here, which has a cost independent of the number of channels. If the physical arrangement of X.Y setup and the user position are given, a spatio-temporal representation of the impulse response for each loudspeaker in the neighborhood of the user position can be obtained. The Soundfield microphone and related Ambisonics recording technique [3] make an attempt to preserve the directionality of the sound field; however, the sound field representation used is far from complete and is insufficient for creating a fully immersive experience. The spatio-temporal representation proposed by us fully preserves the field directionality and can be used for playback with HRTFs. Further, when the signals (e.g., the tracks of the music piece) for each channel are available, it is possible to sum up contributions of all channels into *a single* representation that captures *the total acoustic field produced by all loudspeakers, HRTF effects, and reverberation*. A constant-cost computation is then used to generate the signal for binaural presentation based on current head-tracking data.

2. BACKGROUND

In a volume with no acoustic sources enclosed, the propagation of the acoustic wave is governed by Helmholtz equation

$$\nabla^2\psi(k, \mathbf{r}) + k^2\psi(k, \mathbf{r}) = 0, \tag{2}$$

where $\psi(k,\mathbf{r})$ is the Fourier transform of the pressure. Solutions of the Helmholtz equation can be expanded as a series of spherical basis functions – the regular $R_n^m(k,\mathbf{r})$ in finite regions and the singular $S_n^m(k,\mathbf{r})$ in infinite regions:

$$R_n^m(k,\mathbf{r}) = j_n(kr)Y_n^m(\theta,\varphi);\ S_n^m(k,\mathbf{r}) = h_n(kr)Y_n^m(\theta,\varphi), \quad (3)$$

$$Y_n^m(\theta,\varphi) = (-1)^m\sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\cos\theta)e^{im\varphi},$$

where (r,θ,φ) are spherical coordinates of the radius vector $\mathbf{r}$, $j_n(kr)$ and $h_n(kr)$ are the spherical Bessel and spherical Hankel functions, respectively, $Y_n^m(\theta,\varphi)$ are the orthonormal spherical harmonics, and $P_n^m(\mu)$ are the associated Legendre functions.

Any regular acoustic field $\Phi(k,\mathbf{r})$ near a point $\mathbf{r}^*$ in a region that does not contain sources can be represented as a sum of regular functions with some complex coefficients $C_n^m(k)$ as

$$\Phi(k,\mathbf{r}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} C_n^m(k)R_n^m(k,\mathbf{r}-\mathbf{r}^*). \quad (4)$$

In practice the outer summation is truncated:

$$\Phi(k,\mathbf{r}) = \sum_{n=0}^{p}\sum_{m=-n}^{n} C_n^m(k)R_n^m(k,\mathbf{r}-\mathbf{r}^*). \quad (5)$$

The parameter p is called the truncation number. Its choice depends directly on k and on the radius D of the region in which the approximation (5) is used to represent the field [4].

3. FREE-FIELD CASE

Impulse response (IR): Let loudspeaker c be located at point $\mathbf{r}^{(c)}$ in the reference configuration, the listener's head center be at the origin, and the region of interest be an area near the origin (in contrast with [5], where area of interest is the whole room). We first computer a representation of the field in the region without the listener.

The reverberation in a simple "shoebox" room can be computed using the Allen-Berkley model [6]. The reverberation in this model is represented by a collection of image sources located outside of the room. The position and the strength of each image source are computed from geometric constraints and wall absorption coefficients [6].

Assume that computed image sources are located at points $\mathbf{r}_l^{(c)}$ and have strengths $Q_l^{(c)}$. The potential $\Phi^{(c)}(k,\mathbf{r})$ formed by all image sources is equal to the sum of potentials generated by each source,

$$\Phi^{(c)}(k,\mathbf{r}) = \sum_l Q_l^{(c)}\frac{e^{ik|\mathbf{r}-\mathbf{r}_l^{(c)}|}}{4\pi|\mathbf{r}-\mathbf{r}_l^{(c)}|}. \quad (6)$$

Using multipole translation identity [4] and assuming that there are no sources within the region of interest, we can represent the potential as

$$\Phi^{(c)}(k,\mathbf{r}) = \sum_l Q_l^{(c)}\left[ik\sum_{n=0}^{p}\sum_{m=-n}^{n} S_n^{-m}(k,\mathbf{r}_l^{(c)})R_n^m(k,\mathbf{r}))\right]. \quad (7)$$

The summation can be regrouped as follows:

$$C_n^{(c)m}(k) = \sum_l Q_l^{(c)}S_n^{-m}(k,\mathbf{r}_l^{(c)}), \quad (8)$$
$$\Phi^{(c)}(k,\mathbf{r}) = ik\sum_{n=0}^{p}\sum_{m=-n}^{n} C_n^{(c)m}(k)R_n^m(k,\mathbf{r}).$$

The potential at any point $\mathbf{r}$ in region of interest can now be computed using these equations, subject of course to the proper choice of p based on k and D. Note that the set of coefficients $C_n^{(c)m}(k)$ for channel c does not depend on point $\mathbf{r}$ at all. Such factorization of the total potential into receiver-position-dependent and receiver-position-independent part constitutes the basis for powerful set of algorithms known as fast multipole methods [7]. As such, the coefficients $C_n^{(c)m}(k)$ can be precomputed in advance. They fully describe the 3-D field in the region of interest. As the Fourier transform of a unit-strength impulse is equal to one at all k, one can also think that the set of these coefficients define a spatio-temporal impulse response (STIR) of a system because for a given evaluation point $\mathbf{r}$ one can directly compute a transfer function $\Phi^{(c)}(k,\mathbf{r})$ and get a corresponding impulse response by doing its inverse Fourier transform.

To reiterate, the difference between "traditional" method of rendering virtual audio (as in, e.g., [2]) and the proposed method is as follows. In the "traditional" method, the current position and orientation of the user is used to compute appropriate system impulse response from scratch (i.e., by using HRTFs corresponding to current positions of loudspeakers and image sources in used-bound coordinate system). In the proposed method, instead, the set of coefficients $C_n^{(c)m}(k)$ determining the STIR in the region where the user might potentially be is fully computed first (this step may be performed in advance), and then the traditional time-domain impulse response is computed for positions of user's ears by computing $\Phi^{(c)}(k,\mathbf{r}_e)$ (where $\mathbf{r}_e$ is the radius vector of the appropriate ear) and taking IFT.

Total field: Assume that for each channel c the signal $x^{(c)}(t)$ is to be played via this channel. Recall that the convolution in time domain is equivalent to the multiplication in frequency domain. If the set of $C_n^{(c)m}(k)$ is known for each channel, then the output signal for each ear located at point $\mathbf{r}_e$ is computed as

$$y(t) = \sum_c IFT(X^{(c)}(k)\Phi^{*(c)}(k,\mathbf{r}_e))(t), \quad (9)$$

where $\Phi^{*(c)}(k,\mathbf{r}_e)$ is a complex conjugation of $\Phi^{(c)}(k,\mathbf{r}_e)$, $X^{(c)}(k)$ is the Fourier transform of $x^{(c)}(t)$, and IFT is the inverse Fourier transform. The following trick can be applied. Observe that if $y(t) = IFT(Z_1(k)Z_2^*(k))(t)$ and $\hat{y}(t) = IFT(Z_1^*(k)Z_2(k))(t)$, then $\hat{y}(t)$ is a time-inverse of $y(t)$. As such, we can put conjugation onto the signal itself as long as we remember to time-inverse the obtained result:

$$\hat{y}(t) = \sum_c IFT(X^{*(c)}(k)\Phi^{(c)}(k,\mathbf{r}_e))(t). \quad (10)$$

Due to linearity of the Fourier transform, one can push the summation inside the IFT, use equation (8) to arrive at

$$D_n^m(k) = \sum_c X^{*(c)}(k)C_n^{(c)m}(k), \quad (11)$$
$$\hat{y}(t) = IFT(ik\sum_{n=0}^{p}\sum_{m=-n}^{n} D_n^m(k)R_n^m(k,\mathbf{r}_e))(t),$$

and obtain $y(t)$ via time inversion. Now coefficients $D_n^m(k)$ form a spatio-temporal field representation (STFR) and fully describe the acoustic field in the area of interest created during playback of $x^{(c)}(t)$ via c loudspeakers; contributions of all c channels are summed up to form $D_n^m(k)$. As a consequence, the computational load during playback is no longer dependent on the number of channels.

In practice, it is necessary to break $x^{(c)}(t)$ into frames as it is usually done for frequency-domain convolution, padding frames as appropriate to avoid wrap-around effects [8]. Then, separate $D_n^m(k)$ are computed for each frame, and the X.Y sound is stored in ready-for-playback format as sets of $D_n^m(k)$ for each frame. The original X.Y tracks are no longer necessary at that point; furthermore, as mentioned before, there exist only one set of $D_n^m(k)$ per input frame, containing all original channels. The storage requirements of such scheme is increased compared to storing only per-channel STIR $C_n^{(c)m}(k)$ (in the latter case only c sets of coefficients are stored, compared to the number of sets $D_n^m(k)$ equal to the number of frames in $x^{(c)}(t)$ in the former case). Higher storage requirements are however compensated by lower computational load during playback.

4. HRTF-ENABLED PLAYBACK

Playback with sphere HRTFs: In the presence of the listener, the field at the ears is scattered off themselves and must be processed with HRTFs. In the absence of actual HRTFs and for simplicity, sometimes spherical HRTFs are used, as if the scatterer were a pinnaless spherical head of radius a. Let

$$\tilde{R}_n^m(k, a, \mathbf{r}_e) = (j_n(kr_e) - \frac{j_n'(ka)}{h_n'(ka)} h_n(kr_e)) Y_n^m(\theta_e, \varphi_e). \quad (12)$$

The modification of the field by the sphere can be obtained by substituting $R_n^m(k, \mathbf{r}_e)$ in (8) and (11) with $\tilde{R}_n^m(k, a, \mathbf{r}_e)$ [9].

Individual HRTFs: For greater fidelity, individualized HRTFs must be incorporated and (8) be modified. By definition, the HRTF $H(k, \mathbf{s}_j)$ is the potential created at the ear by the plane wave $e^{ik\mathbf{r}\cdot\mathbf{s}_j}$. Assume first that the user is fixed (later, we will allow the user to move and rotate) and that the incident $\Phi^{(c)}(k, \mathbf{r})$ is given by (8). The idea is to compute the potentials $\Psi_{(h)n}^m(k)$ at the ear created by each spherical mode $R_n^m(k, \mathbf{r})$ and then sum them up weighted by the input coefficients $C_n^{(c)m}(k)$. $\Psi_{(h)n}^m(k)$ is called *mode response*, and subscript h indicates that it incorporates the HRTFs of the particular user and is therefore individualized.

To incorporate HRTFs into the rendering, we decompose each spherical mode over the plane wave basis [10]. Any regular acoustic field $\Phi(k, \mathbf{r})$ in a region can be represented as a superposition of plane waves $e^{ik\mathbf{s}\cdot\mathbf{r}}$, with each plane wave weighted by $\mu(k, \mathbf{s})$:

$$\Phi(k, \mathbf{r}) = \frac{1}{4\pi} \int_{S_u} \mu(k, \mathbf{s}) e^{ik\mathbf{s}\cdot\mathbf{r}} dS(\mathbf{s}), \quad (13)$$

where integration is taken over all possible directions $\mathbf{s}$; $\mu(k, \mathbf{s})$ is known as a signature function and fully specifies the field in the region. The multipole and the plane-wave representations can be converted to each other via Gegenbauer expansion:

$$\begin{aligned} \mu(k, \mathbf{s}) &= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} i^{-n} C_n^m(k) Y_n^m(\mathbf{s}), \quad (14) \\ C_n^m(k) &= i^n \int_{S_u} \mu(k, \mathbf{s}) Y_n^{-m}(\mathbf{s}) dS(\mathbf{s}). \end{aligned}$$

In practice, integration over the surface is replaced with summation over L quadrature points on the sphere with quadrature weights w_j. Details of choosing L, relationship between p and L, and procedures for computing quadrature points and weights can be found in [10]. Thus,

$$\begin{aligned} R_n^m(k, \mathbf{r}) &= \frac{i^{-n}}{4\pi} \sum_{j=1}^{L} w_j Y_n^m(\mathbf{s}_j) e^{ik\mathbf{r}\cdot\mathbf{s}_j}, \quad (15) \\ e^{ik\mathbf{r}\cdot\mathbf{s}_j} &= 4\pi \sum_{n=0}^{p} \sum_{m=-n}^{n} i^n Y_n^{-m}(\mathbf{s}_j) R_n^m(k, \mathbf{r}). \quad (16) \end{aligned}$$

For a given mode $R_n^m(k, \mathbf{r})$, equation (15) decomposes it into a set of plane waves in the directions in the HRTF measurement grid. As the response to each of those plane waves is known (it is just the HRTF), the mode response $\Psi_{(h)n}^m(k)$ to $R_n^m(k, \mathbf{r})$ is

$$\Psi_{(h)n}^m(k) = \frac{i^{-n}}{4\pi} \sum_{j=1}^{L} w_j Y_n^m(\mathbf{s}_j) H(k, \mathbf{s}_j). \quad (17)$$

Recall that the incident field is described as

$$\Phi^{(c)}(k, \mathbf{r}) = ik \sum_{n=0}^{p} \sum_{m=-n}^{n} C_n^{(c)m}(k) R_n^m(k, \mathbf{r}). \quad (18)$$

As $\Psi_{(h)n}^m(k)$ is the response to $R_n^m(k, \mathbf{r})$, the response $\Psi_{(h)}^{(c)}(k)$ to the incident field $\Phi^{(c)}(k, \mathbf{r})$ for channel c is therefore

$$\Psi_{(h)}^{(c)}(k) = ik \sum_{n=0}^{p} \sum_{m=-n}^{n} C_n^{(c)m}(k) \Psi_{(h)n}^m(k). \quad (19)$$

This equation defines how the free-field STIR defined by $C_n^{(c)m}(k)$ is modified by the user's HRTFs.

Now, consider the case when the user is allowed to move freely. Note that the motions of the user are relative to the field; e.g., if a user rotates, we can instead rotate the incident field in opposite direction. The incident field is described by $C_n^{(c)m}(k)$; the field obtained by translation and rotation of the incident field is described by some other coefficients $\hat{C}_n^{(c)m}(k)$. Given field's translation and rotation, it is possible to write a transform matrix that for a $O\left(p^3\right)$ cost computes $\hat{C}_n^{(c)m}(k)$ from $C_n^{(c)m}(k)$. The expressions are not given here for brevity; a full analysis and fast recurrences for transform matrix are given in [4]. As such, to allow for user motion, the equation (19) is modified by substituting $C_n^{(c)m}(k)$ with $\hat{C}_n^{(c)m}(k)$ transformed from $C_n^{(c)m}(k)$ using current user's pose.

Similarly, free-field STFR $D_n^m(k)$ can be combined with the user's HRTFs for fixed-cost playback:

$$\Psi_{(h)}(k) = ik \sum_{n=0}^{p} \sum_{m=-n}^{n} \hat{D}_n^m(k) \Psi_{(h)n}^m(k)$$

where $\hat{D}_n^m(k)$ is $D_n^m(k)$ subject to the translation/rotation operator according to the current user pose. Note that the computation all the way from the original signal to the $D_n^m(k)$ representation and computation of mode responses from individual user's HRTFs can be done offline, and the remaining (playback) computational load is relatively low.

An alternative approach is to embed HRTF directly in content as follows. Compute, via traditional methods such as [2], the potential at the ear for many possible head rotations for pulse signal being played via loudspeaker c.Then, fit it with multipoles, resulting in a set of coefficients $C_n^{(c)m}(k)$ forming a personalized, HRTF-embedded STIR, which can be used to prepare HRTF-embedded

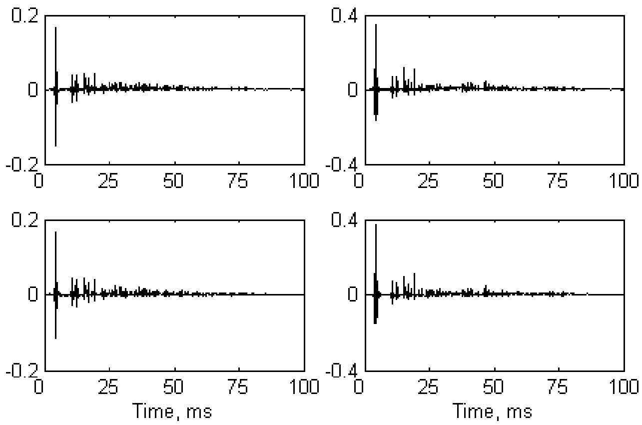

Fig. 1. Left column: Time-domain left-ear IR computed for left loudspeaker and the head rotated 30 degrees to the left (i.e., the person is facing the loudspeaker) using traditional algorithm (top) and proposed algorithm (bottom). Right column: Same comparison for right-ear IRs for the head rotated 80 degrees to the left.

STFR directly playable using equation (11) with current ear position as $\mathbf{r}_e$. In such content, music and HRTFs will be interwoven in a complex manner. As a possible commercial application, a "personalized music" service can be set up to acquire user's HRTFs and issue music in form of HRTF-embedded STFR.

5. RESULTS

We describe several practical issues and compare the proposed algorithm with traditional one in terms of execution speed and in the amount of data involved in the decomposition.

Truncation number: The truncation number depends on the radius of the region of interest D and on the wavenumber k. It is shown in [5] that theoretically the good choice of p is $p^* = (ekD - 1)/2$ and in practice $p = (3/4)p^*$ halves the computational load (due to number of expansion coefficients roughly proportional to p^2) with negligible increase in approximation error. If a listener is not allowed to move and can only rotate, then D is equal to the radius of the head a. In the experimental evaluation of the algorithm accuracy and speed (below), we use $a = 0.087m$, in which case the truncation number $(3/4)p^*$ is about 6 for $k = 73.2$ (i.e., at the frequency of 4 kHz). (Note that the truncation number should be varied as a function of k for further reduction of computation load and to prevent overfitting).

Playback with sphere HRTFs: Computed free-field STFR can be used for playback with sphere HRTFs by substituting $\tilde{R}_n^m(k, a, \mathbf{r}_e)$ for $R_n^m(k, \mathbf{r}_e)$ in equation (11). Listening experiments show that the obtained playback signal indeed exhibits the sphere-scattered characteristics (shadowing, diffraction, and bright spot). Of course, playback with sphere HRTFs can only convey azimuthal (left-to-right) information.

Computation of mode responses: The computational time involved in computation of spherical mode responses $\Psi_{(h)n}^m(k)$ was evaluated on 1.7 GHz Pentium Pro processor based PC. The HRTF measurement grid had 393 positions and was similar to the grid used in the original KEMAR HRTF measurements at MIT [12]. It was found that the computation of mode responses takes about thirty seconds for both ears.

Total field representation computation and playback: We used a sampling frequency of 8 kHz and simulated reverberation length of 200 ms. For each 200 ms fragment of the input sound, its STFR was computed; STFR of one fragment occupied approximately 62.4 kB. As a comparison, 200 ms worth of data in 5.1 format at 8 kHz with 16-bit precision occupies (uncompressed) 19.2 kB. We can see that the amount of data is increased compared to the original sound. However, the proposed algorithm's significant advantage is that the sound frame in STFR *always* occupies 62.4 Kb, no matter how many channels there are in the input X.Y signal, and playback cost is also independent of the number of channels.

The head-tracked playback was performed in real-time directly from STFR and mode responses without using the original X.Y sound, consuming about 70% of the available CPU computational power. We have compared the output signal produced by our algorithm for several random head orientations with the output signal produced by traditional virtual audio synthesis algorithm [2] for the same head orientations and found that the average difference in all cases over all samples is below 1.5% (Figure 1), showing good accuracy of multipole representation and reasonable choice of p. Increase in p decreases approximation error at the expense of increased computational load.

Use of the technique with higher sampling rate would require higher computational load and larger storage requirements due to increased truncation number. We have attempted to implement the technique with the CD-quality music (44.1 kHz sampling rate) and found that real-time playback is not feasible on our hardware; however, the Moore's law advances in computing power and new architectures (e.g., general-purpose GPU computing) should allow CD-quality implementations in the near future.

6. REFERENCES

[1] R. H. Gilkey and T. R. Anderson (eds.) "Binaural and Spatial Hearing in Real and Virtual Environments", L. Earlbaum Assoc., Mahwah, NJ, 1997.

[2] D. N. Zotkin, R. Duraiswami, and L. S. Davis. "Rendering localized spatial audio in a virtual auditory space", IEEE Trans. Multimedia, 6:553-564, 2004.

[3] R. K. Furness. "Ambisonics – An overview", Proc. 8th AES Int. Conf., Washington, D.C., pp. 181-189, 1990.

[4] N. A. Gumerov and R. Duraiswami. "Fast multipole methods for the Helmholtz equation in three dimensions", Elsevier, 2004.

[5] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov. "Fast evaluation of the room transfer function using the multipole method", IEEE Trans. Speech Audio Proc., in press.

[6] J. B. Allen and D. A. Berkley. "Image method for efficiently simulating small-room acoustics", J. Acoustic. Soc. Am., 65:943-950, 1979.

[7] V. Rokhlin. "Rapid solution of integral equations of classical potential theory", Computational Physics, 60:187-207, 1983.

[8] A. V. Oppenheimer and R. W. Schafer. "Digital signal processing", Prentice Hall, 1975.

[9] J. Meyer and G. Elko. "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield", Proc. IEEE ICASSP, 2:1781-1784, 2002.

[10] R. Duraiswami, Z. Li, D. N. Zotkin, E. Grassi, and N. A. Gumerov. "Plane-wave decomposition analysis for spherical microphone arrays", Proc. IEEE WASPAA, 150-153, 2005.

[11] S. Zhang and J. M. Jin. "Computation of special functions", Wiley, 1996. http://jin.ece.uiuc.edu/routines/routines.html

[12] W. G. Gardner and K. D. Martin. "HRTF measurements of a KEMAR", J. Acoustic. Soc. Am., 97:3907-3908, 1995.