Audio Engineering Society

Convention Paper

Presented at the 119th Convention
2005 October 7–10 New York, NY, USA

High Order Spatial Audio Capture and Binaural Head-Tracked Playback over Headphones with HRTF Cues

Ramani Duraiswami[1,2], Dmitry N. Zotkin[1], Zhiyun Li[1,2], Elena Grassi[1], Nail A. Gumerov[1], and Larry S. Davis[1,2]

[1]*Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD 20742 USA*

[2]*Department of Computer Science, University of Maryland, College Park, MD 20742 USA*

Correspondence should be addressed to Ramani Duraiswami (ramani@umiacs.umd.edu)

ABSTRACT

A theory and a system for capturing an audio scene and then rendering it remotely are developed and presented. The sound capture is performed with a spherical microphone array. The sound field at the location of the array is deduced from the captured sound and is represented using either spherical wave-functions or plane-wave expansions. The sound field representation is then transmitted to a remote location for immediate rendering or stored for later use. The sound renderer, coupled with the head tracker, reconstructs the acoustic field using individualized head-related transfer functions to preserve the perceptual spatial structure of the audio scene. Rigorous error bounds and Nyquist-like sampling criterion for the representation of the sound field are presented and verified.

1. INTRODUCTION

Large parts of our brains are devoted to the processing of sound cues extracted by the auditory system, and sound plays an important role in the way we interact with the world. People are able to locate and separate sound sources in diverse environments ranging from large open spaces to small crowded rooms. We are able to form judgments about the dynamic environment we occupy and its materials. When objects collide or interact, we are able to perform judgments about the types of objects, their speeds, and other qualities. These percepts are *in addition* to information we extract from speech or the enjoyment we derive from music.

The spatial location of the various elements of the sound scene plays an important role in our perceptual auditory ability. However, when we wish to render (that is, to recreate) auditory scenes using playback mechanisms, such as speakers or headphones, it is extraordinarily dif-

ficult to make listeners unambiguously perceive the location of auditory objects in the scene.

We present here a theory, synthetic validation, system development description, and experimental results for a system that captures a spatial sound scene using arrays of microphones (in particular spherical microphone arrays) and subsequently plays it back in a way that can evoke spatial presence. The theory follows from a consideration of the wave nature of sound. We represent the captured sound field in terms of spherical basis functions of the wave equation or in terms of the Herglotz wave function basis (also known as plane-wave expansions). Converting the captured sound into these base representations requires one to solve the fitting problem. Using the theory of band-limited representations, we are able to develop expressions for representation of the captured sound in closed form. Our theory provides error bounds that can be used to design arrays for capture of sounds of particular frequencies and to build approximations of the spatial sound field that are valid in a domain of a particular size.

These representations of the sound-field can then be used to recreate the spatial sound field for various applications, akin to holography. We focus our attention on the reproduction of the captured sound over headphones for a listener in an immersive reality system. We show how the use of a plane wave basis allows us to render spatial sound with head-related transfer functions, and we develop efficient algorithms for doing this.

We first test our theory by applying it to the reconstruction of synthetic acoustic scenes. Next, we build a practical microphone array system for sound capture and a head-tracked headphone-based reproduction system. We used these systems to capture and reproduce real scenes. Our algorithms are shown to be capable of operating in real time on a regular personal computer platform.

2. BACKGROUND AND PREVIOUS WORK

Using just two receivers (ears), humans are able to localize sound with amazing accuracy [1]. In addition to differences in the time of arrival or in the level between the signals reaching the two ears (interaural time delay, ITD, and interaural level difference, ILD), listeners determine the ”true” source position using additional localization cues arising from sound scattering. Just as a compact disk in white directional light appears to exhibit various colors because of the light scattering by ridges on the surface with characteristic ridge size that is similar to the wavelength of light, the “color” (i.e., the relative magnitudes of the various frequency components) of broadband sound is changed by its interaction with the environment, with the human body, and especially with the pinna. These objects have sizes comparable to the wavelength of sound. The scattering by the human body and by the external ears provides cues to source position. Scattering off the environment (room walls, etc.) provides further localization cues.

If our goal is to reproduce the sound received at the two ears (say in a headphone-based reproduction system) from a given source in a particular environment for some individual, we must reintroduce modifications that would have been made to the sound received at that individual’s ears if he or she were present in that environment. Also, it is extremely essential to track the person’s head movements and to render the scene stabilized with respect to these motions. While making these transformations to the sound may appear challenging, we are helped by the fact that the propagation of sound is a linear process, and the effects of the anatomical scattering, as well as ITDs and ILDs, can be described by a head-related impulse response (HRIR), or alternatively by its Fourier transform, the head-related transfer function (HRTF). Similarly, environmental scattering can be characterized by a room impulse response (RIR), or alternatively by its Fourier transform, the room transfer function (RTF).

The combined effect of the environment and the anatomy is given by a convolution of these two filters. Features related to the RIR (walls and surfaces) occur at a different length scale than the smaller features of the ear and head. Therefore, they are separated in the frequency domain and can be treated separately in the first approximation.

Knowing the HRIR and the RIR, one can, in principle, reconstruct the pressure waveforms that would reach a listener’s ears for any arbitrary source waveform arising from the particular location. Although the way in which the auditory system extracts information from the stimuli at the ears is only partially understood, the pressure at the eardrums is a sufficient stimulus: if the sound pressure signals generated in the rendering system are identical to the stimulus presented at the listener’s ears in the real scene and change the same way with her motion, she will get the same perception as she would have had in the real scene, including the perception of the presence of sound sources at their correct location in exocentric space, the environmental characteristics, and other aspects.

2.1. Anatomical scattering

The HRTF is a function of source direction and frequency, with a weaker dependence on the distance to the sound source [1]. If the sound source is located at azimuth φ, elevation θ, and range r in a spherical coordinate system, then the left and right HRTFs H^L and H^R are defined as the frequency-dependent ratio of the sound pressure level (SPL) at the corresponding eardrum ψ^L and ψ^R to the free-field SPL at the center of the head as if the listener were absent ψ^F:

$$H^L(k;\theta,\varphi,r) = \frac{\psi^L(k;\theta,\varphi,r)}{\psi^R(k;r)}, \qquad (1)$$

with a similar equation for H^R. In the following we will suppress the dependence on the wavenumber k. Usually, the dependence of the HRTF on the range r is also suppressed as it is expected to be small for relatively distant sources. In particular, for the plane wave sound impinging on a listener from a specific direction the range approaches infinity and the HRTF dependence on range can be ignored.

People have different sizes, and their external ears (pinnae) exhibit considerable variability in shape. As a consequence, the HRTF exhibits considerable inter-personal differences. Binaural sound rendering over headphones works extremely well when the listener's own HRTFs are used to synthesize the localization cues [2]. However, measuring the HRTFs is a complicated procedure. Because of that, 3D audio systems typically use a single set of HRTFs previously measured from a particular human or manikin subject. Localization performance generally suffers when a listener listens to directional cues synthesized from non-individual HRTFs [3], leading to two particular kinds of localization errors commonly seen with 3D audio systems: front/back confusions and elevation errors.

2.2. Environmental modeling

To complete the spatialization of sound, environmental scattering cues (also known as reverberation) must be incorporated in the simulation of auditory space [4]. When sound is produced in a reverberant space, the associated reverberation may often be perceived as a background ambience that is separated from the foreground sound. The loudness of the reverberation relative to the loudness of the foreground sound is an important distance cue.

The RIR characterizes environmental scattering and includes effects due to reflection at the boundaries, sound absorption, diffraction around edges and obstacles, and low frequency resonance effects. It is a function of both the source and the receiver locations. A geometrical approach to finding the impulse response is to trace all sound paths between the source and the receiver. While the geometric model may give somewhat reasonable results, it cannot account for propagation around objects, or scattering off edges, and the room impulse response obtained must be modified to account for such scattering. Also, the process of tracing the paths and modifying the RIR may be expensive and time consuming.

Various computational algorithms for RIR calculation have been proposed. A simple image model for box-like rooms was presented in [5]. The model was extended in [6] to the case of arbitrary piecewise-planar rooms and in [7] to the case of a directional and/or a shadowing source or receiver. Statistical approximation of the late reverberation tail was considered in [8]. More advanced reverberation computation methods that account for diffraction effects and are more computationally efficient has been recently developed (see, e.g., [9]).

A further difficulty in incorporating the room impulse response in the playback is the RIR length, which for large rooms can be as long as a few seconds. The source sound must be convolved with the RIR to account for reverberation. Convolution in the time domain can be performed with low latency (as low as one sample), but its complexity is quadratic (a product of the source signal length and the room RIR length). Convolution in the frequency domain (using the FFT) is much faster (the transforms just need be multiplied point by point), but has a delay of at least a single frame. Partitioned convolution using a geometrically increasing lengths of the frame (1, 2, 4, 8, ...) was proposed as a way of achieving efficient convolution without input output delay [10], though this algorithm is claimed to be proprietary. More sophisticated versions to distribute the convolution load according to some desired strategy have been proposed recently [11].

2.3. Recreating spatial audio

For successful reproduction of the localization cues, it is important to keep the left and right audio channels separated. This is easy to achieve when the listener is using headphones. When using loudspeakers, however, there is significant "crosstalk" between each speaker and the opposite ear of the listener. This crosstalk severely degrades localization performance and must be eliminated, which is possible in some situations with a relatively sim-

ple filtering approach [12]. The listener must be stationary and centered between the two loudspeakers (at the "sweet spot") in order for the crosstalk to be cancelled (or, alternatively, the listener's ear positions could be tracked to adapt canceling filters in real-time). Loudspeaker 3D audio systems are effective in desktop computing environments, as in such case there is usually only a single listener who is almost always centered between the speakers and facing the monitor [13], [14]. Extension of this approach to multiple sources and complex environments is very difficult.

A more general synthesis technique is to recreate the original sound field around the listener. Doing that automatically eliminates the crosstalk and sweet-spot problems in loudspeaker systems. One sound field recreation method is to implement the Kirchhoff Integral Equation approach, where specifying the correct acoustic pressure and its normal derivative (proportional to the velocity) on the boundary of the domain can, in principle, achieve an arbitrary acoustic field in the domain. An approximate implementation of this, called Wave Field Synthesis (WFS), partially specifies some of the boundary terms. It was proposed and implemented in [15] and [16]. However, no analysis of the error introduced by the various approximations was provided. In [17], a general framework for implementing the WFS method in real-time for remote rendering was proposed. A microphone array beamformer was used with a localization and tracking system to identify sound sources, and a loudspeaker array was used to approximately implement the principle. To work properly, however, a robust and accurate localization and tracking system and a highly directive beamformer are required. Further, in the absence of error bounds, the approximations made in this approach again cannot be characterized. Finally, extension of WFS to multiple sources and equalization for other rooms are topics of current research.

Another technique that was developed as a result of work in speaker-based higher order spatial audio reconstruction is "ambisonics" [18]. Here measurements at a point are made with a "sound-field" microphone, and the speakers attempt to recreate the sound field, which is expanded in low order (1st order, originally) spherical harmonics at the measurement point. Extensions to higher orders are being pursued [19]. The theory underlying these methods still needs to be fully established.

In contrast, head-tracked headphone-based systems can recreate spatial audio quite convincingly [20] for known source locations and known head-related transfer functions. However, the problem with these systems is that when we want to recreate the scene for multiple sources and complex environments, it is necessary to render the environmental reflections for each source and mix them with the head-related transfer functions cues. As the RIR changes with both source and listener positions (and is different for the two ears), the rendering process must be repeated for each ear for each of many sources. For complex scenes, the rendering cost can become overwhelming. In [21] an interesting, graphics-inspired culling strategy and hierarchical scene composition were proposed to make complex scene rendering efficient, and a tenfold increase in performance for several scenes was demonstrated.

2.4. Goal: Auditory image-based rendering

In both loudspeaker and headphone-based approaches, we can mainly playback a simple audio scene containing only one or a few sources. To playback general scenes, in which the sound may be coming from many real or virtual sources, we need to model the interaction of each source with the environment as well as the interaction of all the reflected waves with the human listener. This process can be time consuming. Our goal here is to capture the existing sound field with sufficient detail in a region of space and then reproduce that sound field remotely for one or more listeners, evoking in them the perception that they would have had if they were present in the original scene. The main advantage of this approach is that instead of a synthesizing the scene from scratch and having to carefully model the sources and their interaction with the environment, we can simply reuse the actual existing wave field. To do this we must first capture the sound field in a way that retains the spatial information and then play it back.

2.4.1. Capturing spatial audio

Use of arrays of microphones can be an effective approach for capturing the spatial structure of the sound in the scene. In [22] and [23] various aspects of general sensor and microphone arrays are elaborated. To capture the 3D sound field, we prefer the symmetric spherical microphone array configuration. In [24] the sound field was captured using an open spherical microphone array in free space. Microphones can also be positioned on the surface of a rigid sphere to make use of the scattering [25] and to decompose the captured sound field using spherical harmonics. In [26] the use of a plane-wave basis to analyze sound fields was suggested, and in [27] the

design of spherical arrays was considered. However, a complete error analysis of the reconstructed field and the frequency dependence of the algorithms were not presented. We build on these works and further improve the analysis to provide explicit error bounds, frequency dependence, and efficient and optimal algorithms for performing the analysis of the captured sound and subsequent synthesis.

2.4.2. Rendering spatial audio directly from recordings

The goal of image-based rendering is to directly play the sound scene from the recorded sound, without having to explicitly model the presence of multiple acoustic sources in the scene or the multipath nature of the room interactions. The early technique of Ambisonics [18] used a low order spherical harmonic sound-field representation and attempted to recreate the sound field in the playback. The system that appears to come closest to ours is the Motion Tracked Binaural system [28]. Here, eight microphones are placed on the equator of a solid head-sized sphere, and the sound is recorded. During playback, the recorded sound that is closest to the ear position is played back. Head tracking allows the listener to rotate their head about an axis passing vertically through their body and the center of the head, and the reproduced scene is made stationary with respect to a listener moving in this way. Some methods for incorporating elevation effects using modeled HRTFs are also included. While this system appears to be a promising beginning, it does not allow for general motion or individualized HRTFs, and it does not seem that a solid theory exists to extend the method. The goal of this paper is to present and to verify such a theory.

3. MATHEMATICAL FORMULATION

3.1. Wave and Helmholtz Equations

Propagation of acoustic pressure perturbations $p'(\mathbf{r},t)$ in a homogeneous medium is described by the wave equation

$$\frac{\partial^2 p'(\mathbf{r},t)}{\partial t^2} = \frac{1}{c^2}\nabla^2 p'(\mathbf{r},t), \tag{2}$$

where c, t, and $\mathbf{r}$ are the speed of sound, time, and the radius vector of the field point, respectively, and ∇^2 is the Laplacian. The wave equation can be converted to the Helmholtz equation in the frequency domain

$$\nabla^2\psi(\mathbf{r}) + k^2\psi(\mathbf{r}) = 0, \quad k = \frac{\omega}{c}, \tag{3}$$

where k is the wavenumber and $\omega = 2\pi f$ is the circular frequency corresponding to the frequency f, by applying the Fourier transform

$$\psi(\mathbf{r},\omega) = \int_{-\infty}^{\infty} e^{i\omega t} p'(\mathbf{r},t)\, dt. \tag{4}$$

The problem under consideration then can be reduced to solving the Helmholtz equation for a number of frequencies. To obtain the time domain solution, we can take the inverse Fourier transform

$$p'(\mathbf{r},t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega t}\psi(\mathbf{r},\omega)\, d\omega. \tag{5}$$

The Helmholtz equation is an elliptic equation. Therefore, its solution can be specified by imposing conditions on the boundary of the domain. For scattering problems, where we seek solutions of the Helmholtz equation on the surface and outside the scatterer, the total field can be decomposed into the incident field and the scattered field:

$$\psi(\mathbf{r}) = \psi_{in}(\mathbf{r}) + \psi_{scat}(\mathbf{r}), \tag{6}$$

where the incident field is the field in the absence of the scatterer. For an infinite domain, the scattered field satisfies the Sommerfield radiation condition (Eq. (7) left), and on the surface S of the scatterer the total field satisfies the impedance boundary condition (Eq. (7) right)

$$\lim_{r\to\infty}\left[r\left(\frac{\partial\psi_{scat}}{\partial r} - ik\psi_{scat}\right)\right] = 0, \left(\frac{\partial\psi}{\partial n} + i\sigma\psi\right)\Big|_S = 0. \tag{7}$$

Here σ is the complex surface admittance (inverse of the impedance), and the $\partial/\partial n$ denotes the partial derivative in the direction of the normal outward to the region. For sound-hard surfaces $\sigma = 0$. More complex boundary conditions that take into account wave propagation inside the scatterer can be considered. However, when the material of the scatterer has much higher impedance than the outer medium (e.g., a air-to-solid interface), the sound-hard condition models the actual boundary conditions arising from continuity of pressure and normal velocities on the boundary very well.

3.2. Representations of the incident field

Our goal is to sample the sound field using a spherical microphone array and to build a representation of the incoming sound field (the "incident field"). We assume that the region in which we are modeling this sound field does

not contain any scattering objects or sound sources. This means that in this neighborhood the incident field is a regular function of $\mathbf{r}$. The sound field representation we are seeking is a decomposition of the wave field based on directions. As the original sound field satisfies the wave equation, we want to build its representation that also satisfies the wave equation and therefore is valid not only on the capture surface but also in a spatial neighborhood of it. We can then use this representation and further processing to account for scattering off the individual and to achieve an audio playback that is accurate and therefore provides a sense of spatial presence.

3.2.1. Spherical Wave Functions

Because the sound field satisfies the Helmholtz equation, it can be expanded into an infinite series over the basis of elementary spherical regular solutions $\left\{R_n^m(k;\mathbf{r})\right\}$ of Helmholtz equation:

$$\psi_{in}(k;\mathbf{r}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} A_n^m R_n^m(k;\mathbf{r}), \qquad (8)$$

where A_n^m are the expansion coefficients. The basis functions of the expansion are

$$R_n^m(k;\mathbf{r}) = j_n(kr)Y_n^m(\theta,\varphi),\ n=0,1,2,...;\ m=-n,...,n. \qquad (9)$$

Here (r,θ,φ) are the spherical coordinates of the location $\mathbf{r}$, $j_n(kr)$ are spherical Bessel functions of the first kind, and $Y_n^m(\theta,\varphi)$ are the normalized spherical harmonics

$$Y_n^m(\theta,\varphi) = (-1)^m\sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\cos\theta)e^{im\varphi}, \qquad (10)$$

where $n=0,1,2,...$ and $m=-n,...,n$. They are related to the associated Legendre functions $P_n^{|m|}(\mu)$, which can be defined by the Rodrigues formulae:

$$\begin{aligned} P_n^m(\mu) &= (-1)^m\left(1-\mu^2\right)^{m/2}\frac{d^m}{d\mu^m}P_n(\mu), \qquad (11)\\ P_n(\mu) &= \frac{1}{2^n n!}\frac{d^n}{d\mu^n}\left(\mu^2-1\right)^n. \end{aligned}$$

Here $P_n(\mu)=P_n^0(\mu)$ are the Legendre polynomials. For future reference, we note the addition theorem satisfied by the spherical harmonics: if $\mathbf{s}_l$ and $\mathbf{s}_j$ are two points on the unit sphere with coordinates $\mathbf{s}_l=(\theta_l,\varphi_l)$ and $\mathbf{s}_j=(\theta_j,\varphi_j)$, then the spherical harmonics and the Legendre polynomial of order n of the angle between $\mathbf{s}_l$ and $\mathbf{s}_j$ satisfy the relation

$$P_n(\mathbf{s}_l\cdot\mathbf{s}_j) = \frac{4\pi}{2n+1}\sum_{m=-n}^{n} Y_n^{-m}(\mathbf{s}_l)\,Y_n^m(\mathbf{s}_j). \qquad (12)$$

3.2.2. Band-limited plane-wave expansions

While plane-waves are often used to represent sound sources in the far-field, they also constitute a remarkable basis for the wave equation that can represent the sound field in both the near and far-fields. The properties of the plane-wave representation of the wave field are being exploited for developing faster versions of the fast multipole method for the Helmholtz equation [29]. Here we use the plane-wave basis to decompose the sound field. This basis, also called the Herglotz wave function, represents the sound field in integral form and can be interpreted as expressing the incident field as a superposition of plane waves $e^{ik\mathbf{s}\cdot\mathbf{r}}$ propagating[1] in all possible directions $\mathbf{s}$, where the magnitude and the phase of the plane wave in the direction $\mathbf{s}$ is characterized by the complex amplitude $\mu_{in}(\mathbf{s})$. Thus we write

$$\psi_{in}(\mathbf{r}) = \frac{1}{4\pi}\int_{S_u} e^{ik\mathbf{s}\cdot\mathbf{r}}\mu_{in}(\mathbf{s})\,dS(\mathbf{s}), \qquad (13)$$

where the integration is taken over all directions (notionally over the surface of a unit sphere), $\mathbf{s}$ is the unit vector on the unit sphere with Cartesian coordinates $\mathbf{s}=(\sin\theta\cos\varphi,\sin\theta\sin\varphi,\cos\theta)$, and $\mu_{in}(\mathbf{s})$ is a surface function, also known as the far-field signature function. We again reiterate that this is *not* a far-field representation of the sound field – plane waves provide an exact basis over which we can approximate *any* regular soundfield.

In fact, incident field representations (8) and (13) are closely related due to the Gegenbauer expansion of the plane wave [30]

$$\begin{aligned} e^{ik\mathbf{s}\cdot\mathbf{r}} &= 4\pi\sum_{n=0}^{\infty}\sum_{m=-n}^{n} i^n Y_n^{-m}(\mathbf{s})\,R_n^m(\mathbf{r}), \qquad (14)\\ R_n^m(\mathbf{r}) &= \frac{i^{-n}}{4\pi}\int_{S_u} e^{ik\mathbf{s}\cdot\mathbf{r}}Y_n^m(\mathbf{s})dS(\mathbf{s}), \end{aligned}$$

[1] We take the convention that the plane wave is propagating in the direction $\mathbf{s}$. Some authors adopt the convention that $\mathbf{s}$ is the direction the wave is coming from. Care must be taken when comparing expressions.

where we use the notation $Y_n^m(\mathbf{s})$ instead of $Y_n^m(\theta,\varphi)$ (i.e., $\mathbf{s}$ is a point on the unit sphere). This shows that the expansion coefficients in Eq. (8) and the signature function in Eq. (13) are related as

$$\begin{aligned}\mu_{in}(\mathbf{s}) &= \sum_{n=0}^{\infty}\sum_{m=-n}^{n} i^{-n} A_n^m Y_n^m(\mathbf{s}), \qquad (15)\\ A_n^m &= i^n \int_{S_u} \mu_{in}(\mathbf{s})\, Y_n^{-m}(\mathbf{s})\, dS.\end{aligned}$$

In practice we work with samples of a function; therefore, the integral over the sphere needs to be performed via some sort of quadrature, which replaces the integration over the sphere by a summation of function values at selected quadrature points multiplied by appropriate weights (Eq. (16), left). We will return to the particular quadrature we choose later – here it suffices to mention that the locations and weights of the quadrature points are chosen to compute the integral over the sphere surface exactly for the functions that can be represented by a "band-limited" spherical harmonic expansion of a particular degree up to p (Eq. (16), right)

$$\int_{S_u} F(\mathbf{s})\, dS = \sum_{j=0}^{L_Q-1} F(\mathbf{s}_j)\, w_j,\ F(\mathbf{s}) = \sum_{n=0}^{p-1}\sum_{m=-n}^{n} C_n^m Y_n^m(\mathbf{s}), \qquad (16)$$

where C_n^m is the set of arbitrary coefficients and L_Q is the number of quadrature points. Thus the concept of band-limitedness is applicable to both the plane-wave and the truncated spherical wave-function expansions. In the subsequent work we assume we work in a band-limited representation in either case. The goal is to choose a band-limit that is sufficiently accurate for our purposes. To make this choice, we need to establish the error bounds.

3.3. Choice of band-limit via the error bounds

We intend to measure the sound-field using a spherical microphone array. Our goal is to build a sampled version of the plane wave representation and to establish an error bound for it. We also aim at constructing this sampled version from the measurements taken with the spherical microphone array. Finally, using the reconstructed representation, we aim to recreate sound that a listener would have heard in the vicinity of the location of the microphone array. All these tasks use a band-limited representation of the sound, and error bounds that explicitly show that the sound field is in fact reconstructed with prescribed accuracy for a given spatial band-limit parameter p are crucial to this scheme.

3.3.1. Truncated Spherical Wavefunction Expansions of a plane wave

As the series (8) converges absolutely and uniformly in the domain of interest, the function $\psi_{in}(\mathbf{r})$ can be approximated by a band-limited functions $\psi_{in}^{(p)}(\mathbf{r})$. Here, by band-limited functions we mean the functions that are represented by the first p^2 terms of expansions (8), that is, when (8) is truncated at the maximum degree $p-1$. The value of p chosen depends on the frequency and the size of the domain as shown below.

Let us consider the truncation error in representing a single plane wave input field by using the regular spherical basis functions. It can be evaluated based on the Gegenbauer expansion (14), where

$$\begin{aligned}\epsilon_p(\mathbf{s},\mathbf{r}) &= e^{ik\mathbf{s}\cdot\mathbf{r}} - 4\pi \sum_{n=0}^{p-1}\sum_{m=-n}^{n} i^n Y_n^{-m}(\mathbf{s})\, R_n^m(\mathbf{r}) \qquad (17)\\ &= \sum_{n=p}^{\infty} (2n+1)\, i^n\, j_n(kr)\, P_n\left(\frac{\mathbf{r}\cdot\mathbf{s}}{r}\right)\end{aligned}$$

Assume that the domain of interest can be enclosed inside a sphere of radius R. In this case the following general error bound can be found using the inequality [30]

$$|j_n(kr)| \leqslant (kR)^n / (2n+1)!!$$

(where !! indicates double factorial) and the error bound for the Taylor expansion of the exponent:

$$\begin{aligned}|\epsilon_p(\mathbf{s},\mathbf{r})| &= \left|\sum_{n=p}^{\infty} (2n+1)\, i^n\, j_n(kr)\, P_n\left(\frac{\mathbf{r}\cdot\mathbf{s}}{r}\right)\right| \qquad (18)\\ &\leqslant \sum_{n=p}^{\infty} (2n+1)\,|j_n(kr)| \leqslant \sum_{n=p}^{\infty} \frac{(kR)^n}{(2n-1)!!}\\ &< \sum_{n=p}^{\infty} \frac{(kR)^n}{2^{n-1}(n-1)!}\\ &\leqslant \frac{2}{p!}\left(\frac{kR}{2}\right)^{p+1} \exp\left(\frac{kR}{2}\right) = \delta_p, \quad p \geqslant 1.\end{aligned}$$

For relatively low ($kR < 1$) or moderate ($kR \approx 1$) frequencies, equation (18) provides relatively low p (e.g. for $kR = 2$ we have $|\epsilon_p| < 2e/p!$). For higher frequencies ($kR \gg 1$) asymptotic analysis (e.g., see [31]) shows

that p should be always larger than kR and

$$|\epsilon_p(\mathbf{s},\mathbf{r})| \lesssim \exp\left\{-\frac{1}{3}\left[2\frac{p-kR}{(kR)^{1/3}}\right]^{3/2}\right\} = \delta_p, \quad kR \gg 1. \tag{19}$$

Since in all cases the error $\epsilon_p(\mathbf{s},\mathbf{r})$ can be uniformly bounded and the incident field can be represented as a superposition of plane waves, we obtain from Eq. (13) the overall error of approximation of the incident field by the band-limited function $\psi_{in}^{(p)}(\mathbf{r})$ inside a sphere of radius R:

$$\left|\psi_{in}(\mathbf{r}) - \psi_{in}^{(p)}(\mathbf{r})\right| \leqslant \frac{1}{4\pi}\int_{S_u} |\epsilon_p(\mathbf{s},\mathbf{r})|\,|\mu_{in}(\mathbf{s})|\,dS(\mathbf{s}) \tag{20}$$

$$\leqslant \max|\epsilon_p(\mathbf{s},\mathbf{r})| \max|\mu_{in}(\mathbf{s})| \lesssim \delta_p \max|\mu_{in}(\mathbf{s})| = \epsilon_s,$$

where δ_p can be selected according to Eq. (18) or Eq. (19). Note that in the latter case the formula can be inverted to determine the truncation number based on the specified accuracy ϵ_s:

$$p \approx kR + \frac{1}{2}\left(3\ln\frac{\max|\mu_{in}(\mathbf{s})|}{\epsilon_s}\right)^{2/3}(kR)^{1/3}, \quad kR \gg 1. \tag{21}$$

Two terms in the sum in the Eq. (21) can actually be viewed as the initial value of p chosen to assure convergence plus a small correction, which is based on the desired accuracy and logarithmically grows with it.

For the practical implementation, the evaluation can go both ways – to compute the accuracy for a given truncation number using Eq. (19) or to evaluate the necessary truncation number for a given demanded accuracy using Eq. (21). The desired accuracy ϵ_s in Eq. (21) can be defined as a percentage of the maximum sound field magnitude (e.g., if we want the error ϵ_s not to exceed 2% of the maximum sound field decomposition coefficient $\mu_{in}(\mathbf{s})$, then the value of the fraction under the logarithm in (21) is $1/0.02 = 50$ and for $kR = 20$ the second term in Eq. (21) is approximately equal to 7). When performing a multifrequency analysis, it is best in practice to increase p along with the frequency as guided by (21) to avoid errors due to numerical instabilities in the special function routines.

3.4. Scattering off a spherical microphone array

Our goal is to build the appropriate representation of the incident sound field from the sound measured after scattering off the sphere. Thus we must undo the effects of scattering off the sphere from the measurements to arrive at the incident field. In particular, we will assume that the incoming sound field is of finite bandwidth and is to be represented in the band-limited plane-wave basis. The Rayleigh solution of the problem of scattering of a plane wave off a sound-hard sphere of radius a is classical and can be found, for example, in a recent paper [25]. For a general band-limited incident field with coefficients of expansion A_n^m and sound-hard boundary conditions, the solution at a point with angular coordinates $\mathbf{s}$ on the sphere is

$$\begin{aligned}\psi_S(\mathbf{s}) &= \frac{i}{(ka)^2}\sum_{n=0}^{p-1}\frac{1}{B_n(ka)}\sum_{m=-n}^{n}A_n^m Y_n^m(\mathbf{s}), \\ B_n(ka) &= h_n'(ka) + (\sigma/k)\,h_n(ka),\end{aligned} \tag{22}$$

where subscript S denotes that ψ_S is the measured field on the sphere surface. Particularly, when the incident field is a band-limited plane wave propagating in the direction $\mathbf{s}'$, it follows from Eq. (14) that $A_n^m = 4\pi i^n Y_n^{-m}(\mathbf{s}')$. Using the addition theorem for spherical harmonics (12), we can write the measured sound field at $\mathbf{s}$ due to a plane wave propagating towards $\mathbf{s}'$ as

$$\psi_S(\mathbf{s};\mathbf{s}') = K(\mathbf{s};\mathbf{s}') = \frac{i}{(ka)^2}\sum_{n=0}^{p-1}\frac{i^n(2n+1)\,P_n(\mathbf{s}\cdot\mathbf{s}')}{B_n(ka)}. \tag{23}$$

Let the sound received at the sphere be denoted by a superposition of plane waves (13). Then the measured sound-field on the sphere is a superposition of the scattered plane waves, and we obtain it as

$$\begin{aligned}\psi_S(\mathbf{s}) &= \frac{1}{4\pi}\int_{S_u} K(\mathbf{s};\mathbf{s}')\,\mu_{in}(\mathbf{s}')\,dS' \\ &= \sum_{l=0}^{L_Q-1} w_l K(\mathbf{s};\mathbf{s}_l')\,\mu_{in}(\mathbf{s}_l'),\end{aligned} \tag{24}$$

where there are L_Q quadrature points at which we must compute the input coefficients $\mu_{in}(\mathbf{s}_l')$ to determine the input field that would have existed in the absence of the spherical array. One way to proceed is to choose particular measurement (microphone) locations $\mathbf{s}_j$ and particular quadrature points $\mathbf{s}_l$ and to solve the following (possibly over-determined) linear system of equations in $\mu_{in}(\mathbf{s}_l)$:

$$\psi_S(\mathbf{s}_j) = \sum_{l=0}^{L_Q-1} K(\mathbf{s}_j;\mathbf{s}_l)\,w_l\,\mu_{in}(\mathbf{s}_l), \quad j = 1,\cdots,L_M, \tag{25}$$

where L_M is the total number of microphones.

Using the fact that we work in a band-limited representation, we can also derive explicit expressions for the incoming plane-wave coefficients $\mu_{in}(\mathbf{s}_l)$ in terms of the measured $\psi_S(\mathbf{s}_j)$ on the sphere. These expressions will allow us to perform exact fitting in the space of band-limited functions of bandwidth p. By bandwidth p, we mean that the functions are represented by the first p^2 terms of expansions (8) (i.e., by the series (8) truncated at the maximum degree $n = p-1$). Following is the derivation of the explicit expressions for $\mu_{in}(\mathbf{s}_l)$.

The coefficients A_n^m representing the input field can be found from Eq. (22) due to the orthonormality of the system:

$$A_n^m = -i\,(ka)^2\, B_n(ka) \int_{S_u} \psi_S(\mathbf{s})\, Y_n^{-m}(\mathbf{s})\, \mathrm{d}S(\mathbf{s}). \quad (26)$$

Substituting this into the first Eq. (15) and using the addition theorem for spherical harmonics, we obtain for the function truncated to a bandwidth p:

$$\begin{aligned}
\mu_{in}(\mathbf{s}') &= \sum_{n=0}^{p-1} \sum_{m=-n}^{n} i^{-n} A_n^m Y_n^m(\mathbf{s}') \quad (27)\\
&= -i\,(ka)^2 \sum_{n=0}^{p-1} i^{-n} B_n(ka) \times \\
&\quad \times \int_{S_u} \psi_S(\mathbf{s}) \sum_{m=-n}^{n} Y_n^m(\mathbf{s}')\, Y_n^{-m}(\mathbf{s})\, dS(\mathbf{s}) \\
&= -\frac{i\,(ka)^2}{4\pi} \sum_{n=0}^{p-1} (2n+1)\, i^{-n} B_n(ka) \times \\
&\quad \times \int_{S_u} \psi_S(\mathbf{s})\, P_n(\mathbf{s}\cdot\mathbf{s}')\, dS(\mathbf{s}).
\end{aligned}$$

Using quadrature (16) for the sphere, we obtain

$$\begin{aligned}
\mu_{in}(\mathbf{s}_l) &= -\frac{i\,(ka)^2}{4\pi} \sum_{n=0}^{p-1} (2n+1)\, i^{-n} B_n(ka) \times \\
&\quad \times \sum_{j=0}^{L_M-1} w_j\, P_n(\mathbf{s}_l\cdot\mathbf{s}_j)\, \psi_S(\mathbf{s}_j) \\
&= \sum_{j=0}^{L_M-1} w_j\, M(\mathbf{s}_l;\mathbf{s}_j)\, \psi_S(\mathbf{s}_j), \quad (28)
\end{aligned}$$

where the symmetric kernel M is

$$\begin{aligned}
M(\mathbf{s}_l;\mathbf{s}_j) &= -\frac{i\,(ka)^2}{4\pi} \sum_{n=0}^{p-1} (2n+1)\, i^{-n} \times \quad (29)\\
&\quad \times B_n(ka)\, P_n(\mathbf{s}_l\cdot\mathbf{s}_j).
\end{aligned}$$

To determine the L_Q coefficients μ_{in}, a series of p matrix-vector products of size $L_Q \times L_M$ must be performed.

3.5. Quadrature for discrete integration

The band-limited plane-wave basis used here relies on quadrature over the sphere. As we work with band-limited functions, the quadrature must be able to accurately reproduce functions of the specified bandwidth p. Further, as these nodes are to be used as sampling points, intuitively we want them distributed "uniformly" on the spherical surface. A surprising result on quadrature over the sphere [32] is that any quadrature formula of order p should have more than p^2 quadrature points. It is known for the exact quadrature that if the bandwidth of all functions $\psi_S(\mathbf{s})$, $K(\mathbf{s};\mathbf{s}')$, and $\mu_{in}(\mathbf{s}')$ is p, it is sufficient to have $L = 4p^2$ nodes in a more or less arbitrary node distribution, but this number is too large. For special node distributions, L can be made smaller. For a spherical grid which is a Cartesian product of the grid equispaced in φ and the grid in θ in which the nodes are distributed as zeros of the Legendre polynomial of degree p with respect to $\cos\theta$ we have $L = 2p^2$; however, these points are quite inconveniently distributed, and it would be hard to manufacture such a microphone array.

If we are willing to use approximate quadrature formula, then it may be possible both to reduce the number of points and to have uniform point distribution. In [33] quadrature points and weights for various number of nodes are computed using an optimization procedure. These weights and positions are not related to the spherical harmonics, and no analytical guarantees on their integration are available. On the other hand, the point distributions in [33] are relatively uniform, which makes it possible to manufacture such microphone array.

To test the utility of these points for use with spherical harmonics, we performed empirical integration of spherical harmonic pairs of various orders over the sphere using the Fliege quadrature nodes and weights with 64 points and 324 points and tested their suitability for integrating spherical harmonics pairs of up to order 7 and 17,

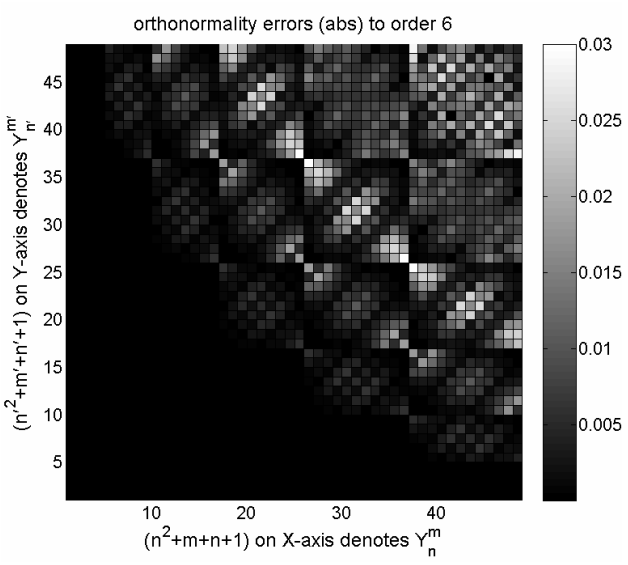

Fig. 1: The validity of the orthogonality relation $\int_{S_u} Y_n^m(\mathbf{s}) Y_{n'}^{-m'}(\mathbf{s}) dS$ is verified for $n, n' = 0, \ldots, 6$ and $m = -n, \ldots n$; $m' = -n', \ldots, n'$ using the 64 quadrature points calculated in [33]

respectively. These samples were seen to satisfy the orthogonality condition with low error. The corresponding error plot is shown in Fig. 1, where the 64 node quadrature of [33] is used and the difference between the theoretical value of the integral (1 when order and degree are equal and 0 in all other cases) and its computed value is shown as a pixel gray level. We note that the product of two Y_p^r and Y_q^s is a spherical harmonic of order $p+q$, and it may not be justified to expect reasonable performance beyond $p = 3$. Despite that, these nodes perform remarkably well. For higher order products, a maximum error of 0.03 is introduced, and most errors are much smaller. In practice we expect these errors to be of lesser order than other errors introduced in experiments (finite-bit representations, noise, etc.)

3.6. Headphone-based sound rendering

For a given source, individual (with known HRTFs), and environment, sound rendering over headphones has been done by a number of authors (see, e.g., [20]). These systems can be extended to a few sources, although the cost of rendering the environmental reflections (and their modification by anatomical scattering) makes their complexity prohibitive for large numbers of sources. Here our goal is to use the plane-wave representation obtained from the sound measured by the spherical array to do the rendering. If the head-related transfer functions are assumed to be obtained from far-field sources, as is customary in the literature where their dependence on range is neglected, then the modification of the incoming plane-wave in the frequency domain is particularly simple: all we need to do is to multiply the signal arriving at the head center by the appropriate HRTF.

In the present decomposition, the sound arriving at the head center consists of a superposition of weighted plane waves. To get the total sound field, we need to just take the sound field arriving from each of the HRTF measurement directions, weigh them with the HRTF, and sum up over all arrival directions. However, there is a potential issue here in that the spherical grid on which head-related transfer functions are measured can be quite different from the microphone array grid in both the number of points and the directions they correspond to. Fortunately, the band-limited representation provides a method for performing the interpolation in an easy manner. We just use Eq. (28) with l chosen to run over the HRTF grid and j over the array grid. In this case we can weigh the plane-wave densities corresponding to sample points l on the HRTF grid. Using Eq. (28) we can write the sound fields $\psi^L(k)$ and $\psi^R(k)$ received at the locations of the left ear and the right ear, respectively, in the plane wave basis as

$$\psi^L = \sum_{l=1}^{L_Q} w_l H_l^L \mu_{in}(\mathbf{s}_l), \quad \psi^R = \sum_{l=1}^{L_Q} w_l H_l^R \mu_{in}(\mathbf{s}_l), \tag{30}$$

where the L_Q nodes $\mathbf{s}_l$ (the HRTF grid locations) have quadrature weights w_l that ensure that functions with band-limit p are integrated exactly and $H_l^L(k)$ and $H_l^R(k)$ are the left and right HRTF values for the direction $\mathbf{s}_l$ and the wavenumber k. The dependence on k has been suppressed in the equation (30) for clarity, but it should be remembered that ψ^L, ψ^R, w_l, H_l^L, H_l^R, and computed $\mu_{in}(\mathbf{s}_l)$ do depend on k as the plane-wave sound field decomposition is done at each frequency of interest (i.e., at each k) independently.

Knowing the potentials at the left and right ears for all k of interest, we can then render the sound by going back to the time domain. The full algorithm implemented in our system, from sound capture to sound rendering, is presented below in pseudocode form.

HRTF-based spatial scene capture and rendering algorithm
Initial Input: array radius, number of microphones, number of HRTF samples and locations, desired quadrature order p
Preliminary (offline) processing:
determine a quadrature weights of order p for the microphone grid and for the HRTF grid
determine appropriate analysis band and correspondence between wavenumber and $p(k)$ using Eq. (21).
Online processing of data frames:
For frame i
Input data from array at each of the L_M microphones of length T
Prepare data and convert to frequency domain
for k ($k_{\min}$ to $k_{\max}$)
select $p(k)$
do fitting at the HRTF grid nodes
build $\psi(k)$ at the sphere center
Evaluate $\psi_{Left}(k)$ and $\psi_{Right}(k)$ at the HRTF grid nodes
next k
Perform an Inverse FFT to obtain sound in the time domain
Perform any filtering modifications
Playback
next frame

4. EXPERIMENTAL SYSTEM

We have built an experimental system to validate and test in practice the theoretical results described above. The experimental setup consists of a spherical microphone array (depicted in Fig. 2), two custom-made amplifier boxes, and a typical high-end office PC outfitted with two data acquisition boards for signal acquisition.

The microphone array is constructed from a spherical plastic lampshade of 10.1 cm radius. The microphones used are only 2.5 mm in diameter (Knowles FG-3329). The microphone grid is a Fliege's 64 quadrature points, which results in near-unit quadrature weights. Out of 64 positions, only 60 are used as the lower part of the sphere containing nodes 12, 24, 29, and 37 is cut off to accommodate microphone cables. Thus, there are 60 microphones in the grid.

During the recording, the data is collected simultaneously from all 60 microphones in the spherical array. The signal recorded by the microphones is decoupled, amplified, and low-pass filtered with custom preamplifiers (4th order Bessel filter, cut-off frequency of 18 kHz, gain 100) to avoid aliasing. Signals are sampled at a rate of 39.0625 kHz and are acquired through the A/D ports of two National Instruments PCI-6071E cards. The playback system consists of the Sennheiser HD-470 headphones and a 3SPACE FASTRAK Polhemus head-tracking system connected to a PC computer. The head-tracking system transmits position coordinates of the head of the subject relative to a transmitter to the PC serial port to stabilize the audio scene in the coordinate frame that is stationary with respect to a moving listener, which is crucial to achieve proper perception [20].

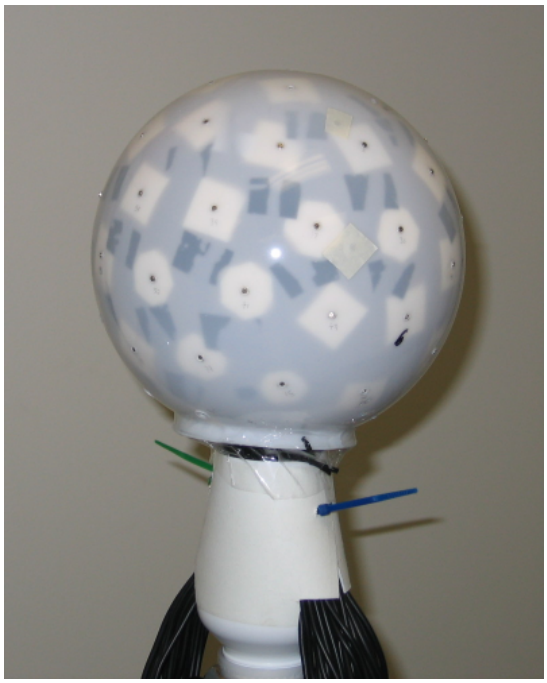

Fig. 2: Our spherical microphone array is constructed using 60 Knowles FG-3329 microphones on a 10 cm radius lampshade.

The useful frequency range of the microphone array is limited from below by the sphere radius a and from above by the number of microphones. When a is very small compared to the wavelength, wave propagation is not significantly affected by the sphere and the pressure variations at different microphones are very small and are likely to be masked by the measurement noise in the

real setup. In practice, the $ka \approx 0.3$ is the low frequency limit of the array, which translated approximately to the frequency of 160 Hz. The high frequency limit is dictated by the number of microphones being equal to p^2; given that p must be set approximately equal to ka, the high limit is $ka \approx 7$ (the frequency of approximately 4 kHz). At higher frequencies, spatial aliasing occurs. Therefore, all processing is limited to this frequency range. However, it is necessary to include both lower and higher frequencies in the rendering for the purposes of realism.

At low frequencies, head scattering effects are negligible; therefore, low-frequency content can be taken from any microphone in the array and mixed into rendering, which we do. The advantage of an array of the size similar to the size of the human head becomes apparent here. As far as high frequencies are concerned, the best solution is to build a microphone array with sufficient number of microphones. Alternately, an approach similar in spirit to [28] can be considered. The high-frequency part is composed from frequencies with wavelength of the order of the inter-microphone distance or smaller. Such waves are thus much smaller than the radius of the sphere, and they are well-attenuated by the sphere. As such, the microphone that is closest to the acoustic source is the best receiver of the high-frequency content of the source among all microphones in the array. Therefore, when high-frequency content is to be played back, the high-frequency component for each plane wave could be taken from the microphone that is located at the corresponding direction at the sphere and could be added to the playback with an as yet undetermined HRTF filtering approach, or directly mixed in without elevation cues as in [28].

5. VERIFICATION AND TESTING

We describe here the results obtained in simulations and in the real setup. In order to demonstrate the recreation of simulated sound fields from synthetic capture data, we performed several sets of simulations and evaluated the accuracy of the reconstruction. We also performed several recordings in real audio environments and created sample re-synthesized audio renderings.

5.1. Synthetic Verification

The theory presented is formally exact if exact quadrature is available. The implementation of the theory with the approximate quadrature points can be verified using the synthetic data experiment, where the sound field that would have been measured by an array is simulated and then reconstructed using the algorithm presented. The results of this verification for our implementation are presented here.

When the underlying theory is presented to people unfamiliar with the plane-wave basis, the reconstruction of near-field sources using this basis usually leads to some discussion. In the following, we show that it is indeed possible to do that. The first thing we must recognize is what the algorithm is attempting to recreate. It only recreates the projection of the sound field up to a particular bandwidth using the regular spherical eigenfunctions of the Helmholtz equation. We thus take a sound field that is reduced to this bandwidth and compare it with the reconstructed field. This comparison reveals that with Fliege node sampling we are able to achieve the required reconstruction with low error.

In Fig. 3, we show the reconstruction of the sound-field created by a point source. The source is at a distance of 0.2 m from the array center. The first simulation (top row) has a source of frequency 1 kHz and array radius of 0.1 m. This corresponds to $ka = 1.83$ and a predicted p of 3. As it is seen from the plot, reconstruction errors are low in the region where the reconstruction should hold (the neighborhood of the sphere shown as a circular area in this cross-section through the $z = 0$ plane). The 64 Fliege nodes were used. The bottom row shows the same plots for a case where the source frequency is 8 kHz and the array radius is 0.08 m so that $ka = 11.73$ and a predicted $p = 13$. We used 324 Fliege nodes. Again good reconstruction is observed in the region where it should be valid.

5.2. Experimental Verification

Several sets of recording to validate the technique were performed in different conditions (on-street recordings of street traffic, in-room recordings of several speakers, and in-room recordings of a speaker and music). Head-tracked playback of these recordings over headphones is quite convincing. To present the data in a form suitable for the paper, we have selected one representative recording and describe the results of its analysis here.

The recording is made in a typical office room. Two sound sources are presented in the acoustic scene. Both sources are computer speakers; the first plays a speech signal and the second plays music. The recording length is 10 seconds. The locations of these sounds are (elevation, azimuth) $(49^\circ, 103^\circ)$ and $(54^\circ, -32^\circ)$, respec-

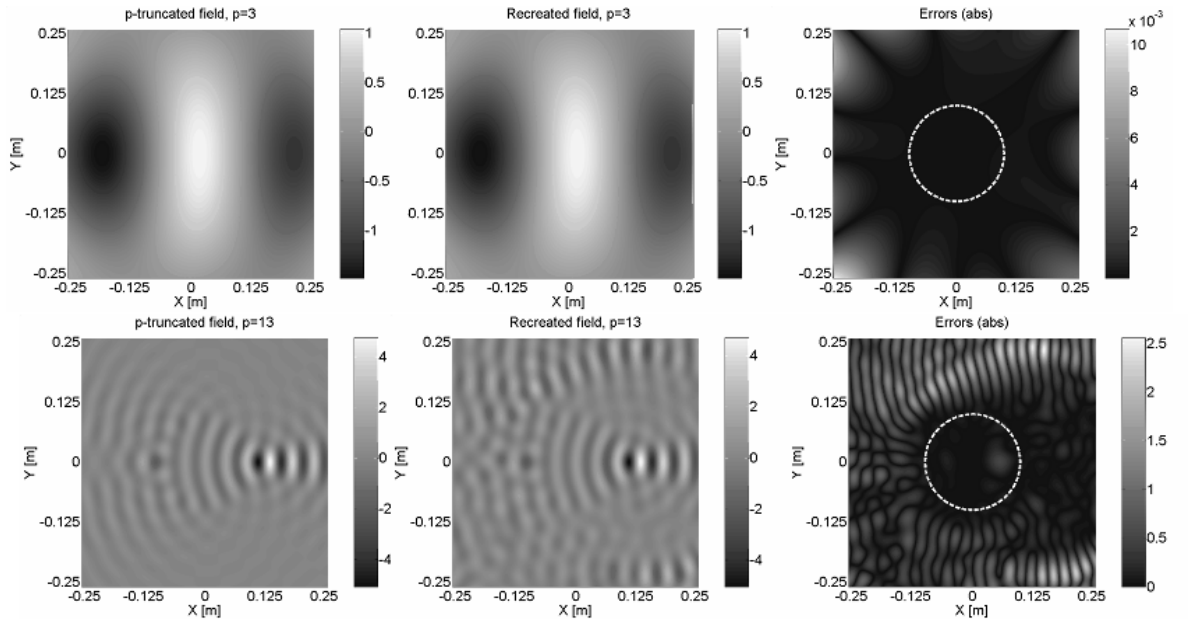

Fig. 3: Reconstruction of the field due to a point source located at a distance of $0.2m$ from the center of a spherical microphone array of radius a. The top row shows the reconstruction for a 1 kHz source and $a = 0.1m$, whereas the bottom row shows the reconstruction for a 8 kHz source and $a = 0.08m$. Good reconstruction in the region of the array is achieved in both cases.

tively, in the spherical coordinate system. The recorded data is processed through the reconstruction algorithm described above.

While the raw plot of the sound field intensity measured by the microphones on the sphere shows a mixed and unclear picture, the reconstruction in terms of plane-waves presents a very clear separation of the resulting audio streams (see Fig. 4). During playback with our HRTFs, notable spatial separation is obtained in the resulting audio stream, especially when comparison is made with the audio stream recorded at any given microphone at the sphere surface, where a confused mix of sources is perceived. In the HRTF enabled playback, the positions of sources are perceived as external and stable, and no position drift or jitter is noticeable.

Fig. 4 corresponds to a 2000-sample data frame at 6.758 s instant of the recording and a frequency of 975 Hz, which corresponds to $k = 17.8$ and $ka = 1.8$. The left part of the figure shows the acoustic field on the sphere surface as measured by microphones. The positions of the dots correspond to the positions of the microphones on the sphere, and dot gray level represents the magnitude of the measured pressure Fourier coefficient. Two crosses are placed in the plot at the directions of two sound sources. Because of the complex interference pattern between two acoustic waves depicted in the plot, there is no agreement between the positions of the pressure peaks and the actual source positions, and it is hard to infer the positions of the acoustic sources in the scene from the plot. After processing the data with the algorithm described above, we plot the magnitudes of $\mu_{in}(\mathbf{s}_l)$ in the right part of Fig. 4. Two peaks corresponding to two sources can now clearly be seen at the correct locations. Rendering of the computed plane waves with the corresponding coefficients and with HRTF imposed produces the final audio stream.

6. CONCLUSION

We have presented a generalized approach that, akin to holography, builds a spatial representation of the sound field locally from spherical array recordings. Our approach is shown to work well both in simulations and

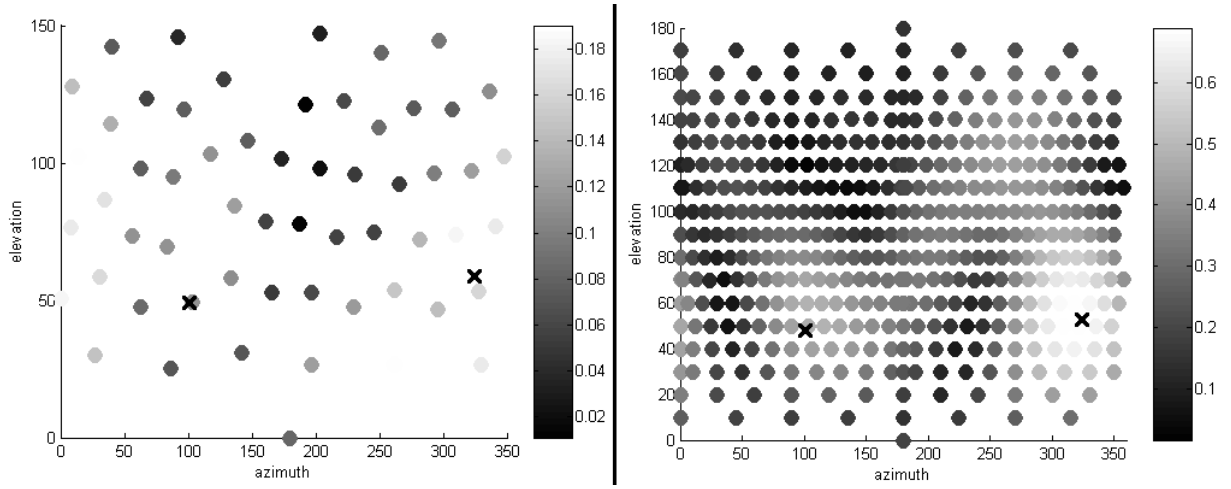

Fig. 4: Two sound sources were placed approximately 1.5 m from the array at an (elevation, azimuth) respectively of $(49^\circ, 103^\circ)$ (for speech) and $(54^\circ, -32^\circ)$ (for music) in a normal office room. Measured magnitudes of the sound field at a time instant at the different microphones in the array are shown on the left; dots represent microphone positions on the sphere surface, and dot gray level represents magnitude. Source locations are indicated by X. The sound field does not show any spatial structure. On the right, the reconstructed plane-wave coefficient magnitudes μ_{in} are shown at the coordinates corresponding to the HRTF grid. Clear spatial structure is visible, and two peaks at the locations corresponding to known source positions can be seen.

in limited practical testing. Key to our approach is an analytical apparatus that constructs the representation of the sound field using plane wave expansions, which are shown to be identical to conventional near-field expansions up to a specified order in the spherical eigenfunctions. While the present paper focused on presenting the theory and sample verification results, future work will focus on software implementation details and on presentation of a reference implementation of the work. Key theoretical area of work is the development of the microphone array layouts and associated quadrature weights that, at least approximately, achieve accurate integration of spherical harmonics over the sphere for practical microphone arrays (post-construction). The theory can also be used to objectively compare many surround-sound encodings that have been proposed. Other work may include extensions that would allow building representations from multiple arrays, thus capturing larger scenes.

7. REFERENCES

[1] W. M. Hartmann (1999). "How we localize sound", Physics Today, vol. 52, no. 11, pp. 24-29.

[2] F. L. Wightman and D. J. Kistler (1989). "Headphone simulation of free-field listening, II: Psychophysical validation", J. Acoust. Soc. Am., vol. 85, no. 2, pp. 868-878.

[3] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman (1993). "Localization using nonindividualized head-related transfer functions", J. Acoust. Soc. Am., vol. 94, no. 1, pp. 111-123.

[4] B. G. Shinn-Cunningham (2001). "Creating three dimensions in virtual auditory displays", in Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality, ed. by M. Smith, G. Salvendy, D. Harris, and R. Koubek, Lawrence Erlbaum Associates, NJ, pp. 604-608.

[5] J. B. Allen and D. A. Berkley (1979). "Image method for efficiently simulating small-room acoustics", J. Acoust. Soc. Am., vol. 65, no. 5, pp. 943-950.

[6] J. Borish (1984). “Extension of the image model to arbitrary polyhedra”, J. Acoust. Soc. Am., vol. 75, no. 6, pp. 1827-1836.

[7] M. Kompis and H. Dillier (1993). “Simulating transfer functions in a reverberant room including source directivity and head-shadow effects”, J. Acoust. Soc. Am., vol. 93, no. 5, pp. 2779-2787.

[8] J.-M. Jot, L. Cerveau, and O. Warusfel (1997). “Analysis and synthesis of room reverberation based on a statistical time-frequency model”, Proc. AES 103th conv., New York, NY, preprint 4629.

[9] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, and A. Ngan (2004). “A beam tracing method for interactive architectural acoustics”, J. Acoust. Soc. Am., vol. 115, no. 2, pp. 739-756.

[10] W. G. Gardner (1995). “Efficient convolution without input-output delay”, J. Audio Eng. Soc., vol. 43, pp. 127-136.

[11] G. Garcia (2002). “Optimal filter partition for efficient convolution with short input/output delay”, Proc. AES 113th conv., Los Angeles, CA, preprint 5560.

[12] W. G. Gardner (1997). “Head tracked 3-d audio using loudspeakers”, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 1997), New Paltz, NY.

[13] C. Kyriakakis (1998), “Fundamental and technological limitations of immersive audio”, Proc. of the IEEE, vol. 86, no. 5, pp. 941-951.

[14] C. Kyriakakis, P. Tsakalides, and T. Holman (1999). “Surrounded by sound”, IEEE Signal Processing Magazine, vol. 16, no. 1, pp. 55-66.

[15] A. J. Berkhout, D. de Vries, and P. Vogel (1993). “Acoustic control by wave field synthesis”, J. Acoust. Soc. Am., vol. 93, pp. 2764-2778.

[16] A. J. Berkhout, D. de Vries, and J. J. Sonke (1997). “Array technology for acoustic wave field analysis in enclosures”, J. Acoust. Soc. Am., vol. 102, pp. 2757-2770.

[17] H. Teutsch, S. Spors, W. Herbordt, W. Kellermann, and R. Rabenstein (2003). “An integrated real-time system for immersive audio applications”, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2003), New Paltz, NY.

[18] M. Gerzon (1985). “Ambisonics in multichannel broadcasting and video”, J. Audio Eng. Soc., vol. 33, no. 11, pp. 859.

[19] J. Daniel, R. Nicol, and S. Moreau (2003). “Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging”, Presented at the 114th AES Convention, Amsterdam, 2003 March

[20] D. N. Zotkin, R. Duraiswami, and L. S. Davis (2004). “Rendering localized spatial audio in a virtual auditory space”, IEEE Transaction on Multimedia, vol. 6, no. 4, pp. 553-564.

[21] N. Tsingos, E. Gallo, and G. Drettakis (2004). “Perceptual audio rendering of complex virtual environments”, Proc. SIGGRAPH 2004, Los Angeles, CA.

[22] B. D. van Veen and K. M. Buckley (1998). “Beamforming: A versatile approach to spatial filtering”, IEEE ASSP Magazine, vol. 5, pp. 4-24.

[23] M. Brandstein and D. Ward, eds. (2001). “Microphone Arrays: Signal Processing Techniques and Applications”, Springer Verlag.

[24] T. D. Abhayapala and D. B. Ward (2002). “Theory and design of high order sound field microphones using spherical microphone array”, Proc. IEEE ICASSP 2002, Orlando, FL, May 2002, vol. 2, pp. 1949-1952.

[25] 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, May 2002, vol. 2, pp. 1781-1784.

[26] B. Rafaely (2004). “Plane-wave decomposition of the sound field on a sphere by spherical convolution”, J. Acoust. Soc. Am., vol. 116, no. 4, pp. 2149-2157.

[27] B. Rafaely (2005). “Analysis and design of spherical microphone arrays”, IEEE Transactions on Speech and Audio Processing, vol. 13, no. 1, pp. 135-143.

[28] V. R. Algazi, R. O. Duda, and D. M. Thompson (2004). “Motion-tracked binaural sound”, Proc. AES 116th conv., Berlin, Germany, preprint 6015.

[29] V. Rokhlin (1993). “Diagonal forms of translation operators for the Helmholtz equation in three dimensions”, Appl. and Comp. Harmonic Analysis, vol. 1, no. 1, pp. 82-93.

[30] M. Abramowitz and I. A. Stegun (1964). “Handbook of mathematical functions”, National Bureau of Standards, Washington, D.C.

[31] W.-C. Chew, J.-M. Jin, E. Michielssen, and J. Song (2001). “Fast and efficient algorithms in computational electromagnetics”, Atrech House.

[32] M. Taylor (1995). “Cubature for the sphere and the discrete spherical harmonic transform”, SIAM Journal of Numerical Analysis, vol. 32, no. 2, pp. 667-670.

[33] J. Fliege and U. Maier (1999). “The distribution of points on the sphere and corresponding cubature formulae”, IMA Journal on Numerical Analysis, vol. 19, no. 2, pp. 317-334.

[34] N. A. Gumerov and R. Duraiswami (2005). “Fast multipole methods for the Helmholtz equation in three dimensions”, Elsevier Science, Netherlands.