Technical Report:
A spherical microphone array based system for immersive audio scene rendering

Adam O'Donovan, Dmitry N. Zotkin, and Ramani Duraiswami

March 25, 2008

Abstract

For many applications it is necessary to capture an acoustic field and present it for human listeners, creating the same acoustic perception for them as if they were actually present in the scene. Possible applications of this technique include entertainment, education, military training, remote telepresence, surveillance, and others. Recently, there is much interest on the use of spherical microphone arrays in acoustic scene capture and reproduction application. We describe a 32-microphone spherical array based system implemented for spatial audio capture and reproduction. The array embeds hardware that is traditionally external, such as preamplifiers, filters, digital-to-analog converters, and USB interface adapter, resulting in a portable lightweight solution and requiring no hardware on PC side whatsoever other than a high-speed USB port. We provide capability analysis of the array and describe software suite developed for the application.

1 Introduction

An interesting and important problem related to spatial audio is capture and reproduction of arbitrary acoustic fields. When a human listens to an audio scene, a multitude of factors are extracted by the brain from the audio streams, including the number of competing foreground sources, their directions, environmental characteristics, presence of background sources, etc. It would be beneficial for many applications if such an arbitrary acoustic scene could be captured and reproduced with perceptual accuracy. Since audio signals received at the ears change with listener motion, the same effect should be present in the rendered scene, this can be done by the use of a loudspeaker array that attempts to recreate the whole scene in a region or by a head-tracked headphone setup that does it for an individual listener. We focus on headphone presentation in this paper.

The key property required from the acoustic scene capture algorithm is the ability to preserve the directionality of the field in order to render those directional components properly later. Note that the recording of an acoustic field with a single microphone faithfully preserves the variations in acoustic pressure (assuming omnidirectional microphone) at the point where the recording was made; however, it is impossible to infer the directional structure of the field from that recording.

A microphone array can be used to infer directionality from sampled spatial variations of the acoustic field. One of the earlier attempts to do that was the use of Ambisonics technique and the Soundfield microphone [1] to capture the acoustic field and its three first-order derivatives along the coordinate axes. Certain sense of directionality can be achieved with the Ambisonics reproduction; however, the reproduced sound field is only a rough approximation of the original one (to be exact, the Ambisonics reproduction includes only the first-order spherical harmonics, while accurate reproduction would require order of about 10 for the frequencies up to 8-10 kHz). Recently, researchers turned to using spherical microphone arrays [2] [3] for spatial structure preserving acoustic scene capture. They exhibit a number of properties making them especially suitable for this application, including omnidirectionality, beamforming pattern independent of the steering direction, elegant mathematical framework for digital beam steering, and ability to utilize wave scattering off the spherical support to improve directionality. Once the directional components of the field are found, they can be used to present the acoustic field to the listener by rendering those components to appear as arriving from appropriate directions. Such rendering can be done using traditional virtual audio methods (i.e., filtering with the head-related transfer function (HRTF)). For perceptual accuracy, HRTF of a specific listener must be used when the audio scene is rendered for that listener.

There exist other recently published methods for capturing and reproducing spatial audio scenes. One of them is Motion-Tracked Binaural Sound (MTB) [4], where a number of microphones are mounted on the equator of the approximately head-sized sphere and the left and right channels of the headphones worn by user are “connected” to the microphone signals, interpolating between adjacent positions as necessary, based on the current head tracking data. The MTB system successfully creates the impression of presence and responds properly to user motion. Individual HRTFs are not incorporated, and sounds rendered are limited to the equatorial plane only. Another capture and reproduction approach is Wave Field Synthesis (WFS) [5] [6]. In WFS, a sound field incident to a “transmitting” area is captured at the boundary of that area and is fed to an array of loudspeakers arranged similarly on the boundary of a “receiving” area, creating the field in the “receiving” area equivalent to that in the “transmitting” area. This technique is very powerful, primarily because it can reproduce the field in the large area, enabling the user to wander off the reproduction “sweet spot”; however, proper field sampling requires extremely large number of microphones.

We present the results of a recent research project concerning the development of the portable auditory scene capture and reproduction framework. We have developed a compact 32-channel microphone array with direct digital interface to the computer via standard USB 2.0 port. We have also developed a software package to support the data capture from the array and scene reproduction with individualized HRTF and head-tracking. The developed system is omnidirectional and supports arbitrary wavefield reproduction (e.g., with elevated or overhead sources). We describe the theory and the algorithms behind the developed hardware and software, the design of the array, the experimental results obtained, and the capabilities and limitations of the array.

2 Background

In this section, we describe the basic theory and introduce notation used in the rest of the paper.

2.1 Acoustic field representation

Any regular acoustic field in a volume is subject to Helmholtz equation

$$\nabla^2 \psi(k, \mathbf{r}) + k^2 \psi(k, \mathbf{r}) = 0, \tag{1}$$

where k is the wavenumber, $\mathbf{r}$ is a radius-vector of a point within a volume, and $\psi(k, \mathbf{r})$ is an acoustic potential (a Fourier transform of a pressure). In a region with no acoustic sources, the set of elementary solutions for the Helmholtz equation consists of so-called regular basis function $R_n^m(k, \mathbf{r})$ given by

$$R_n^m(k, \mathbf{r}) = j_n(kr) Y_n^m(\theta, \varphi), \tag{2}$$

where (r, θ, φ) are the spherical coordinates of a radius-vector $\mathbf{r}$, $j_n(kr)$ is the spherical Bessel function of the first kind of order n, and $Y_n^m(\theta, \varphi)$ are the spherical harmonics. Similarly to the Fourier transform, any regular acoustic field can be decomposed near the point $\mathbf{r}^*$ over $R_n^m(k, \mathbf{r})$ as follows:

$$\psi(k, \mathbf{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} C_n^m(k) R_n^m(k, \mathbf{r} - \mathbf{r}^*), \tag{3}$$

where $C_n^m(k)$ are the complex decomposition coefficients. In practice, the infinite summation is approximated with the finite series introducing an error term $\varepsilon(p, k, \mathbf{r}, \mathbf{r}^*)$:

$$\psi(k, \mathbf{r}) = \sum_{n=0}^{p} \sum_{m=-n}^{n} C_n^m(k) R_n^m(k, \mathbf{r} - \mathbf{r}^*) + \varepsilon(p, k, \mathbf{r}, \mathbf{r}^*). \tag{4}$$

The parameter p is commonly called the truncation number. It is shown [7] that if $|\mathbf{r} - \mathbf{r}^*| < D$ then setting

$$p = \frac{ekD - 1}{2} \tag{5}$$

results in negligible error term. More accurate estimation of p is possible based on error tolerance; however, this is beyond the scope of this paper.

2.2 Spherical scattering

The potential $\tilde{\psi}(k, \mathbf{s}', \mathbf{s})$ created at a specific point $\mathbf{s}'$ on the surface of the sphere of radius a by a plane wave $e^{ik\mathbf{r}\cdot\mathbf{s}}$ propagating in the direction $\mathbf{s}$ is given by [8]

$$\tilde{\psi}(k, \mathbf{s}', \mathbf{s}) = \frac{i}{(ka)^2} \sum_{n=0}^{\infty} \frac{i^n (2n+1) P_n(\mathbf{s} \cdot \mathbf{s}')}{h_n'(ka)}, \tag{6}$$

where $P_n(\mathbf{s} \cdot \mathbf{s}')$ is the Legendre polynomial of degree n and $h_n'(ka)$ is the derivative of the spherical Hankel function. Note that some authors take $\mathbf{s}$ to be the wave arrival direction instead of propagation direction, in which case the equation is modified slightly. In more general case of an arbitrary incident field given by equation (3), the potential $\tilde{\psi}(k, \mathbf{s}')$ at point $\mathbf{s}'$ is given by

$$\tilde{\psi}(k, \mathbf{s}') = \frac{i}{(ka)^2} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{C_n^m(k) Y_n^m(\mathbf{s}')}{h_n'(ka)}. \tag{7}$$

Equation (6) can actually be obtained from equation (7) by using Gegenbauer expansion of a plane wave [9] and spherical harmonics addition theorem. Both series can be truncated at p given by equation (5) with $D = a$ with negligible accuracy loss.

2.3 Spatial audio perception

Humans derive information about the direction of sound arrival from the cues introduced into the sound spectrum by sound scattering off the listener's anatomical parts, primarily pinnae, head, and torso [10]. Because of asymmetrical shape of pinna, head shadowing, and torso reflections, the spectrum of the sound reaching the eardrum depends on the direction from which the acoustic wave is arriving. A transfer function characterizing those changes is called the head-related transfer function. It is defined as the ratio of potential at the left (right) eardrum $\psi_L(k, \theta, \varphi)$ $(\psi_R(k, \theta, \varphi))$ to the potential at the center of the head $\psi_C(k)$ as if the listener were not present as a function of source direction (θ, φ):

$$H_L(k, \theta, \varphi) = \frac{\psi_L(k, \theta, \varphi)}{\psi_C(k)}, \; H_R(k, \theta, \varphi) = \frac{\psi_R(k, \theta, \varphi)}{\psi_C(k)}. \tag{8}$$

Here the weak dependence on source range is neglected. Also HRTF is often taken to be the transfer function between the center of the head and the entrance to the blocked ear canal (instead of the eardrum). HRTF constructed or measured according to this definition does not include ear canal effects. It follows that a perception of a sound arriving from the direction (θ, φ) can be evoked if the sound source signal is filtered with HRTF for that direction and delivered to the listener's eardrums or to ear canal entrances (e.g., via headphones).

However, due to personal differences in body parts sizes and shapes, the HRTF is substantially different for different individuals. Therefore, an HRTF-based virtual audio reproduction system should be custom-tailored for every particular listener. Various methods have been proposed in literature for performing such tailoring, including measuring HRTF directly by placing a microphone in the listener's ear and playing test signals from many directions in space, selecting HRTF from the HRTF database based on pinna features and shoulder dimensions, fine-tuning HRTF for the particular user based on where he/she perceives acoustic signals with different spectra, and others. Recently, a fast method for HRTF measurement was proposed and implemented in [11], cutting time necessary for direct HRTF measurement from hours to minutes. In the rest of the paper, we assume that the HRTF of a listener is known. If that is not the case, a generic (e.g. KEMAR) HRTF can be used, although one can expect degradation in reproduction accuracy [12].

3 Spatial Scene Recording and Playback

In summary, the following steps are involved in capturing and reproducing the acoustic scene:

- Record the scene with the spherical microphone array;
- Decompose the scene into components arriving from various directions;
- Dynamically render those components for the listener as coming from their respective directions.

As a result of this process, the listener would be presented with the same spatial arrangement of the acoustic energy (including sources and reverberation) as there it was in the original sound scene. Note that it is not necessary to model reverberation at all with this technique; it is captured and played back as part of the spatial sound field.

Below we describe these steps in greater details.

3.1 Scene recording

To record the scene, the array is placed at the point where the recording is to be made and the raw digital acoustic data from 32 microphones is streamed to the PC over USB cable. In our system, no signal processing is performed at this step and data is stored on the hard disk in raw form.

3.2 Scene decomposition

The goal of this step is to decompose the scene into the components that arrive from various directions. Several decomposition methods can be conceived, including spherical harmonics based beamforming [3], field decomposition over plane-wave basis [13], and analysis based on spherical convolution [14]. While all methods can be related to each other theoretically, it is not clear which of these methods is practically "best" with respect to the ability to isolate sources, noise and reverberation tolerance, numerical stability, and ultimate perceptual quality of the rendered scene. We are currently undertaking a study comparing the performance of those methods using real data collected from the array as well as simulated data. For the described system, we implemented spherical harmonic based beamforming algorithm originally described in [3] and improved in [15], [16], and [17], among others.

To perform beamforming, the raw audio data is detrended and is broken into frames. The processing is then done on a frame-by-frame basis, and overlap-and-add technique is used to avoid artifacts arising on frame boundaries. The frame is Fourier transformed; the field potential $\psi(k, \mathbf{s}'_i)$ at microphone number i is then just the Fourier transform coefficient at wavenumber k. Assume that the total number of microphones is L_i and the total number of beamforming directions is L_j. The weights $w(k, \mathbf{s}_j, \mathbf{s}'_i)$ that should be assigned to each microphone to achieve a regular beampattern of order p for the look direction $\mathbf{s}_j$ are [3]

$$w(k, \mathbf{s}_j, \mathbf{s}'_i) = \sum_{n=0}^{p} \frac{1}{2i^n b_n(ka)} \sum_{m=-n}^{n} Y_n^{m*}(\mathbf{s}_j) Y_n^m(\mathbf{s}'_i), \tag{9}$$

where

$$b_n(ka) = j_n(ka) - \frac{j'_n(ka)}{h'_n(ka)} h_n(ka) \tag{10}$$

and quadrature coefficients are assumed to be unity (which is the case for our system as the microphones are arranged on the truncated icosahedron grid). As noted by many authors, the magnitude of $b_n(ka)$ decays rapidly for n greater than ka, leading to numerical instabilities (i.e., white noise amplification). Therefore, in practical implementation the truncation number should be varied with the wavenumber. In our implementation, we choose $p = \lceil ka \rceil$. Equation (5) can also be used with $D = a$.

The maximum frequency supported by the array are limited by spatial aliasing; in fact, if L_i microphones are distributed evenly over the sphere of radius a, then the distance between microphones is approximately $4aL_i^{-1/2}$ (this is slight underestimate) and spatial aliasing occurs at $k > (\pi/4a)\sqrt{L_i}$. Accordingly, the maximum value of ka is about $(\pi/4)\sqrt{L_i}$ and is independent of the sphere radius. Therefore, one can roughly estimate maximum beamforming order p achievable without distorting the beamforming pattern as $p \sim \sqrt{L_i}$, which is consistent with results presented earlier by other authors. This is also consistent with estimation of number of microphones necessary for forming quadrature of order p over the sphere given in [13] as $L_i = (p+1)^2$. From these derivations, we estimate that with 32 microphones $p = 5$ order should be achievable at higher end of useful frequency range. It is important to understand that these performance bounds are not hard in a sense that the processing algorithms do not break down completely and immediately when constraints on k and on p are violated; rather, these values signify soft limits, and the beampattern start to degrade gradually when those are crossed. Therefore, the constraints derived should be considered approximate and are useful for rough estimate of array capabilities only. We show experimental confirmation of these bounds in the later section.

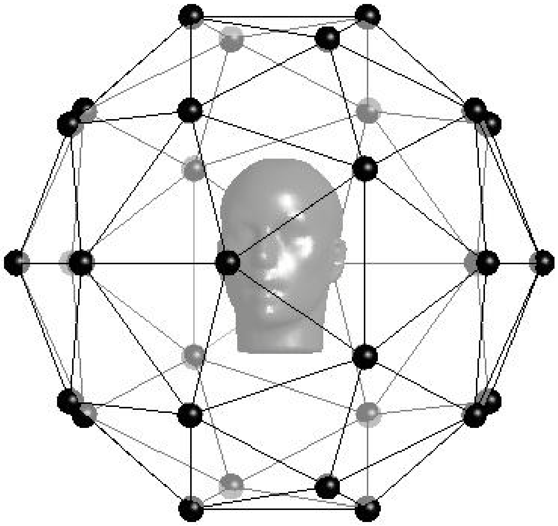

Figure 1: The 32-node beamforming grid used in the system. Each node represents one of the beamforming directions as well as virtual loudspeaker location during rendering.

An important practical question is how to choose the beamforming grid (how large L_j should be and what should be the directions $\mathbf{s}'_j$). Obviously the beamformer resolution is finite and is decreasing as p decreases; therefore, it does not make sense to beamform at a grid finer than the beamformer resolution. Paper [14] suggests that the angular width of the beampattern main lobe is approximately $2\pi/p$, so the width at half-maximum is approximately half of that, or π/p. At the same time, note that if p^2 microphones are distributed evenly over the sphere, the angular distance between neighboring microphones is also π/p. Thus, with the given number of microphones on the sphere the best beampattern that can be achieved has the width at half-maximum roughly equal to the angular distance between microphones. This is confirmed by experimental data (shown later in the paper). Based on that, we select the beamforming grid to be identical to the microphone grid; thus, from 32 signals recorded at microphones, we compute 32 beamformed signals in 32 directions coinciding with microphone directions (i.e., vectors from the sphere center to the microphone positions on the sphere). Figure 1 shows the beamforming grid relative to the listener.

Note that the beamforming can be done very efficiently assuming the microphone positions and the beamforming directions are known. The frequencydomain output signal $y_j(k)$ for direction $\mathbf{s}_j$ is simply

$$y_j(k) = \sum_i w(k, \mathbf{s}_j, \mathbf{s}'_i)\psi(k, \mathbf{s}'_i), \tag{11}$$

where weights can be computed in advance using equation (9), and time-domain signal is obtained by doing inverse Fourier transform. It is interesting to note that other scene decomposition methods (e.g., fitting-based plane-wave decomposition) can be formulated in exactly the same framework but use weights that are computed differently.

3.3 Playback

After the beamforming step is done, L_j acoustic streams $y_j(k)$ are obtained, each representing what would be heard if a directional microphone were pointed at the corresponding direction. These streams can be rendered using traditional virtual audio techniques (see e.g. [18]) as follows. Assume that the user is placed at the origin of the virtual environment and is free to move and/or rotate; user's motion are tracked by a hardware device, such as Polhemus tracker. Place L_j virtual loudspeakers in the environment far away (say at range of 2 meters). During the rendering, for the current data frame, determine (using the head-tracking data) the current direction (θ_j, φ_j) to the j^{th} virtual loudspeaker in user-bound coordinate frame and retrieve or generate the pair of HRTFs $H_L(k, \theta_j, \varphi_j)$ and $H_R(k, \theta_j, \varphi_j)$ that would be most appropriate to render the source located in direction (θ_j, φ_j). This can be a pair of HRTFs for the direction closest to (θ_j, φ_j) available in the measurement grid or HRTF generated on the fly using some interpolation method. Repeat that for all virtual loudspeakers and generate total output stream for the left ear $x_L(t)$ as

$$x_L(t) = IFFT(\sum_j y_j(k) H_L(k, \theta_j, \varphi_j))(t), \tag{12}$$

and similarly for the right ear $x_R(t)$. Note that for online implementation equations (11) and (12) can be combined in a straightforward manner and simplified to go directly (in one matrix-vector multiplication) from time-domain signals acquired from individual microphones to time-domain signals to be delivered to listener's ears.

If a permanent playback installation is possible, the playback can also be performed via a set of 32 physical loudspeakers fixed in the proper directions in accordance with the beamformer grid with the user being located at the center of the listening area. In this case, neither head-tracking nor HRTF filtering is necessary because sources are physically external with respect to the user and are fixed in the environment. In this way, our designed spherical array and beamforming package can be used to create virtual auditory reality via loudspeakers, similarly to the way it is done in high-order Ambisonics or in wave field synthesis [19].

4 Hardware Design

The motivation for the array design was our dissatisfaction with some aspects of our previously developed arrays [20] [21]. They both had 64 channel and had 64 cables – one per each microphone – that had to be plugged into two bulky 32-channel preamplifiers, which were connected in turn to two data acquisition cards sitting in a desktop PC. Street scenes recording was complicated due to the need to bring all the equipment out and keep it powered; furthermore, connection cables were coming loose quite often. In addition, occasionally microphones were failing and it was challenging to replace a microphone in a tangle of 64 cables. So in a nutshell the design goal was to have portable solution requiring no external hardware, having microphones easily replaceable, and connecting with one cable instead of 64.

The physical support of the new microphone array consists of two polycarbonate clear-color hemispheres of radius 7.4 cm. Figure 2 shows the array and some of its internal components. 16 holes are drilled in each hemisphere arranging a total of 32 microphones in truncated icosahedron pattern. Panasonic WM-61A speech band microphones are used. Each microphone is mounted on a miniature (2 by 2 cm) printed circuit board; those boards are placed and glued into the spherical shell from the inside so that the microphone appears from the microphone hole flush with the surface. Each miniature circuit board contains an amplifier with the gain factor of 50 on TLC-271 chip, a number of resistors and capacitors supporting the amplifier, and two connectors – one for microphone and one for power connection and signal output. A microphone is inserted into the microphone connector through the microphone hole so that it can be pulled out and replaced easily without disassembling the array.

Three credit-card sized boards are stacked and placed in the center of the array. Two of these boards are identical; each of these contains 16 digital low-pass filters (TLC-14 chips) and one 16-channel sequential analog-to-digital converter (AD-7490 chip). The digital filter chip has programmable cutoff frequency and is intended to prevent aliasing. ADC accuracy is 12 bits.

The third board is an Opal Kelly XEM3001 USB interface kit based on Xilinx Spartan-3 FPGA. The USB cable connects to the USB connector on XEM3001 board. There is also a power connector on the array to supply power to the ADC boards and to amplifiers. All boards in the system use surface-mount technology. We have developed custom firmware that generates system clocks, controls ADC chips and digital filters, collects the sampled data from two ADC chips in parallel, buffers them in FIFO queue, and sends the data over USB to the PC. Because of the sequential sampling nature, phase correction is implemented in beamforming algorithm to account for skew in channel sampling times. PC side acquisition software is based on FrontPanel library provided by Opal Kelly. It simply streams the data from the FPGA and saves it to the hard disk in raw form.

In the current implementation, the total sampling frequency is 1.25 MHz, resulting in the per-channel sampling frequency of 39.0625 kHz. Each data sample consists of 12 bits with 4 auxiliary ”marker” bits attached; these can potentially

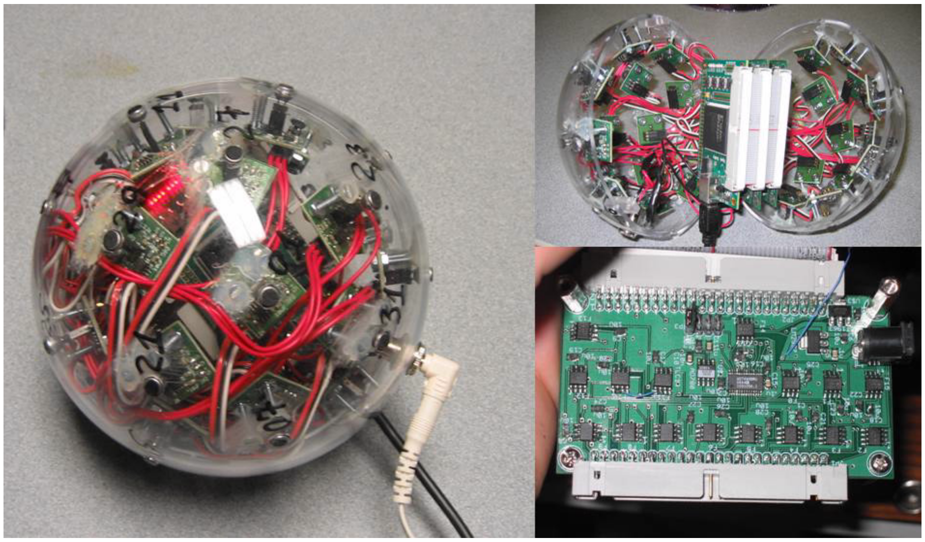

Figure 2: Left: Assembled spherical microphone array. Top right: Array pictured open; a large chip seen in the middle is the FPGA. Bottom right: A close-up of an ADC board.

be stripped on FPGA and data be repacked to reduce data rate but we don't do it. As such, the rate of data transfer from the array is about 2.5 MBytes per second, which is significantly below the maximum USB 2.0 bandwidth. The cut-off frequency of the digital filters is set to 16 kHz. However, these frequencies can be changed easily in software, if necessary. Our implementation also consumes very little of available FPGA processing power. In future, we plan to implement parts of signal processing on the FPGA as well; modules performing FIR/IIR filtering, Fourier transform, multiply-and-add operations, and other basic signal processing blocks are readily available for FPGA. Ideally, the output of the array can be dependent on the application (e.g., in an application requiring visualization of spatial acoustic patterns the firmware computing spatial distribution of energy can be downloaded and the array could send images showing the energy distribution, such as plots presented in the later section of this paper, to the PC).

The dynamic range of 12-bit ADC is 72 dB. We had selected the gain of the amplifiers so that the signal level of about 90 dB would result in saturation of ADC, so the absolute noise floor of the system is about 18 dB. Per specification, the microphone signal-to-noise ratio is more than 62 dB. In practice, we observed that in a recording done in a silence in soundproof room the self-noise of the system spans the lowest 2 bits of the ADC range. Useful dynamic range of the system is then about 60 dB, from 30 dB to 90 dB.

The beamforming and playback are implemented as separate applications.

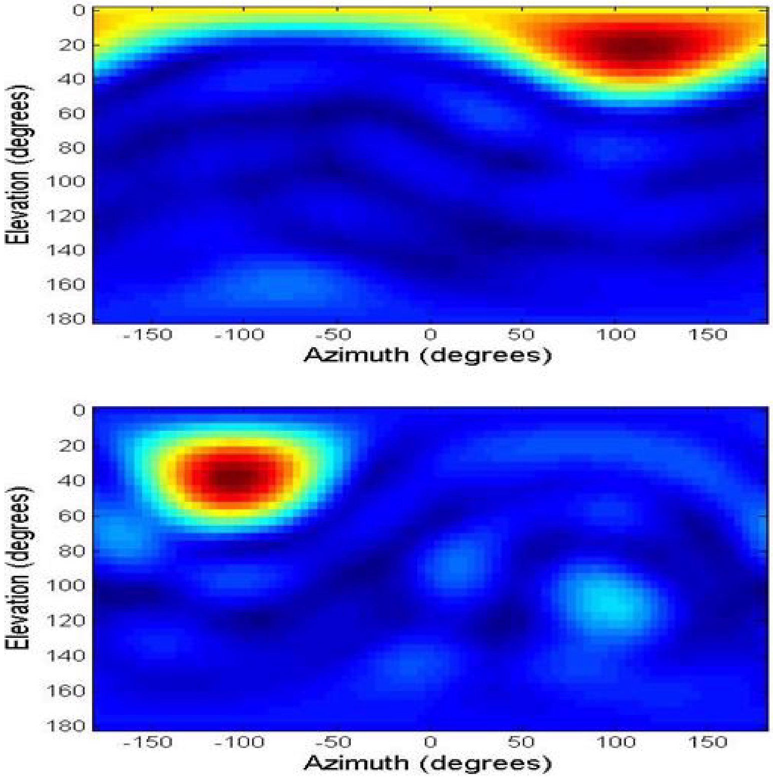

Figure 3: Steered beamformer response power for speaker 1 (top plot) and speaker 2 (bottom plot). Clear peaks can be seen in each of these intensity images at the location of each speaker

Beamforming application processes the raw data, forms 32 beamforming signals using the described algorithms, and stores those on disk in intermediate format. Playback application renders the signals from their appropriate directions, responding to the data sent by head-tracking device (currently supported are Polhemus FasTrak, Ascension Technology Flock of Birds, and Intersense InertiaCube) and allowing for import of individual HRTF for use in rendering. According to preliminary experiments, combined beamforming and playback from raw data can be done in real time but is not currently implemented.

5 Results and Limitations

To test the capabilities of our system, we performed a series of experiments in which recordings were made containing multiple sound sources. During these experiments, the microphone array was suspended from the ceiling in a large

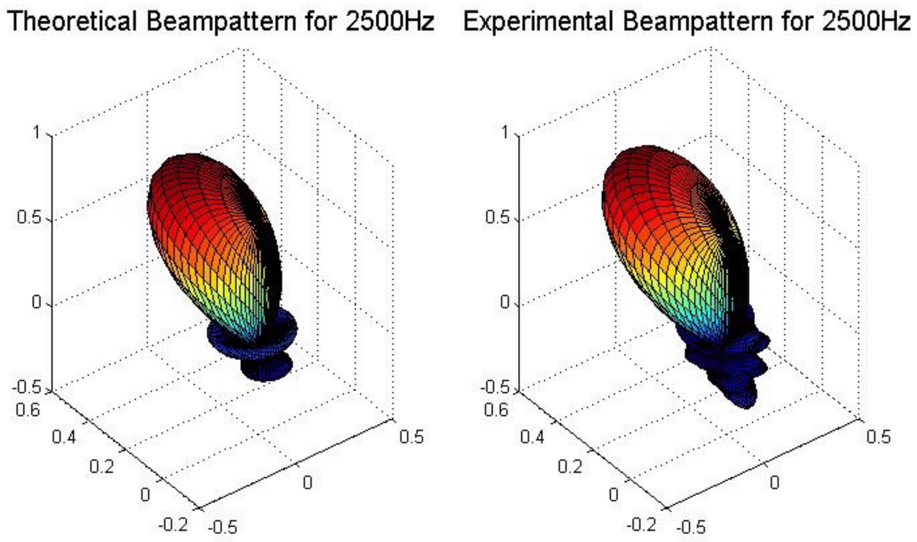

Figure 4: A comparison of the theoretical beampattern for 2500 Hz and the actual obtained beampattern at 2500 Hz. Overall the achieved beampattern agrees quite well with theory, with some irregularities in side lobes.

reverberant environment (a basketball gym) at approximately 1 meter above the ground, and conversations taking place between two persons standing each about 1.5 meters from the array were recorded. Speaker one (S_1) was located at approximately $(20, 140)$ degrees (elevation, azimuth) and speaker two (S_2) was located at $(40, -110)$. We plotted first the steered beamformer response power at the frequency of 2500 Hz over the whole range of directions (Figure 3). The data recorded was segmented into fragments containing only a single speaker. Each segment was then broken into 1024-sample long frames, and the steered power response was computed for each frame and averaged over the entire segment. Figure 3 presents the resulting power response for S_1 and S_2. As can be seen, the maximum in the intensity map is located very close to the true speaker location.

In plots in Figure 3, one can actually see the “ridges” surrounding the main peak waving throughout the plots as well as the “bright spot” located opposite to the main peak. In Figure 4, we re-plotted the steered response power in three dimensions to visualize the beampattern realized by our system in reverberant environment and compared this experimentally-generated beampattern (Figure 4, left) with the theoretical one (Figure 4, right) at the same frequency of 2500 Hz (at that frequency, $p = 4$). It can be seen that the plots are substantially similar. Subtle differences in the side lobe structure can be seen and are due to the environmental noise and reverberation; however the overall structure of the

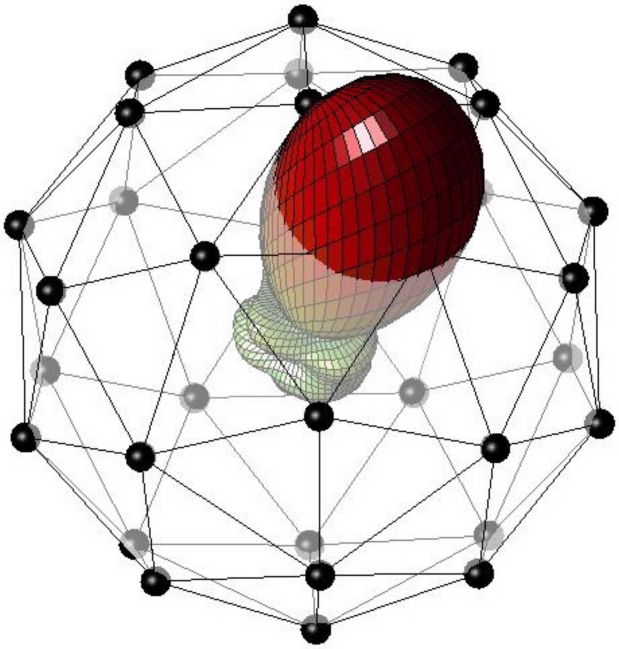

Figure 5: Beampattern overlaid with the beamformer grid (which is identical to the microphone grid).

beam is faithfully retained.

Another plot that provides insights to the behavior of the system is presented in Figure 5. It was predicted in section 3.2 that the beampattern width at half-maximum should be comparable to the angular distance between microphones in the microphone array grid; in this plot, the beampattern is actually overlaid with the beamformer grid (which is in our case the same as the microphone grid). It is seen that this relationship holds well and it indeed does not make much sense to beamform at more directions than the number of microphones in the array.

Using experimental data, we also looked at the beampattern shape at frequencies higher than the spatial aliasing limit. Using derivations in section 3.2, we estimate the spatial aliasing frequency to be approximately 2900 Hz. In Figure 6, we show the experimental beamforming pattern for frequencies higher than this limit for the same data fragment as in the top panel of Figure 3. As Figure 6 shows, beyond the spatial aliasing frequency spurious secondary peaks begin to appear, and at about 5500 Hz they surpass the main lobe in intensity. It is important to notice that these spatial aliasing effects are gradual. According to these plots, we can estimate "soft" upper useful array frequency to be about 4000 Hz.

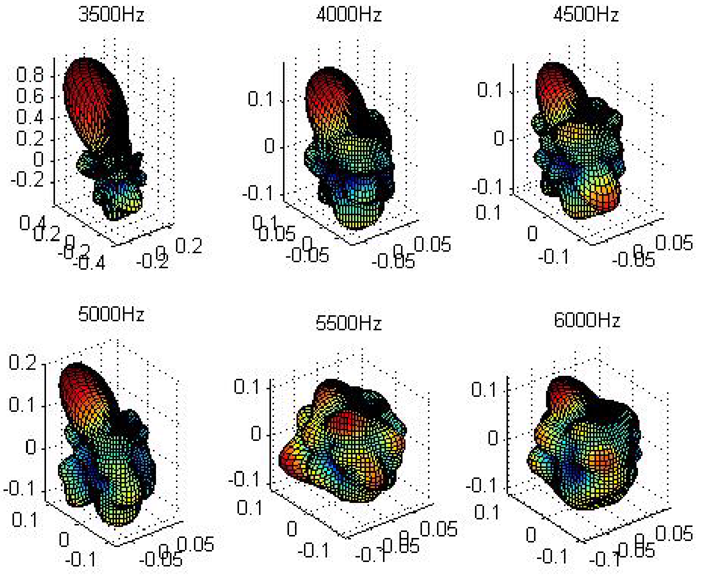

Figure 6: The effect of spatial aliasing. Shown from top left to bottom right are the obtained beampatterns for frequencies above the spatial aliasing frequency. As one can see, the beampattern degradation is gradual and the directionality is totally lost only at 5500 Hz..

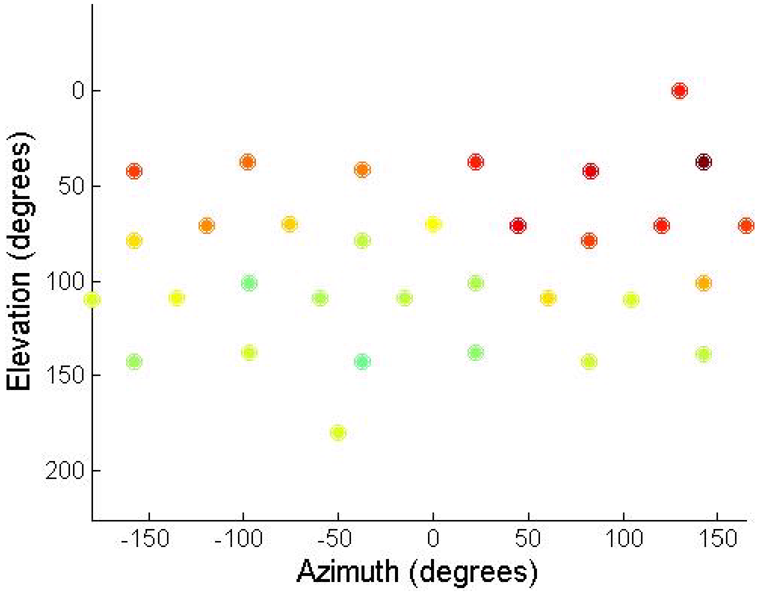

Figure 7: Cumulative power in [5 kHz, 15 kHz] frequency range in raw microphone signal plotted at the microphone positions as the dot color. A peak is present at the speaker's true location.

To account for this limitation we implement a fix for properly rendering higher frequencies similarly to how it is done in MTB system [4]. For a given beamforming direction, we perform beamforming only up to the spatial aliasing limit or slightly above. We then find the closest microphone to this beamforming direction and high pass filter the actual signal recorded at the microphone using the same cutoff frequency. The two signals are the combined to form a complete broadband audio signal. The rationale for that decision is that at higher frequencies the effects of acoustic shadowing from the solid spherical housing are significant, so the signal at microphone located at direction $\mathbf{s}'$ should contain mostly the energy for the source(s) located in the direction $\mathbf{s}'$. Figure 7 shows a plot of the average intensity at frequencies from 5 kHz to 15 kHz for the same data fragment as in the top panel of Figure 3. As can be seen, a fair amount of directionality is present and the peak is located at the location of the actual speaker.

Informal listening experiments show that it is generally possible to identify locations of the sound sources in the rendered environment and to follow them along as they move around. The rendered sources appear stable with respect to the environment (i.e., stay in the same position if the listener turns the head) and externalized with respect to the listener. Without the high-frequency fix, elevation perception is poor because the highest frequency in the beamformed signal is approximately 3.5 kHz and cues creating the perception of elevation are very weak in this range. When high-frequency fix is applied, elevation perception is restored successfully, although the spatial resolution of the system is inevitably limited by the beampattern width (i.e., by the number of microphones in the array). We are currently working on gathering more experimental data with the array and on further evaluating reproduction quality.

6 Conclusions and Future Work

We have developed and implemented a 32-microphone spherical array system for recording and rendering spatial acoustic scenes. The array is portable, does not require any additional hardware to operate, and can be plugged into a USB port on any PC. Spherical harmonics based beamforming and HRTF based playback software was also implemented as a part of complete scene capture and rendering solution. In test recordings, system capabilities agree very well with theoretical constraints. A method for enabling scene rendering at frequencies higher than the array spatial aliasing limit was proposed and implemented. Future work is planned on investigating other plane-wave decomposition methods for the array and on using array-embedded processing power for signal processing tasks.

7 Acknowledgements

This work was supported by a project entitled "Evaluation of a Virtual Reality Holographic Sound System" awarded by the U. S. Department of Veterans Affairs Atlanta Medical Center to the University of Maryland, College Park. We specifically thank David A. Ross of VA Atlanta Medical Center, Decatur, GA for his contribution to this project through support and discussions.

References

[1] R. K. Furness (1990). “Ambisonics – An overview”, Proc. 8th AES Intl. Conf., Washington, D. C. pp. 181-189.

[2] 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, vol. 2, pp. 1949-1952.

[3] 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.

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

[5] A. J. Berkhout, D. de Vries, and P. Vogel (1993). "Acoustic control by wave field synthesis", J. Acoust. Soc. Am., vol. 93, no. 5, pp. 2764-2778.

[6] H. Teutsch, S. Spors, W. Herbordt, W. Kellermann, and R. Rabenstein (2003). "An integrated real-time system for immersive audio applications", Proc. IEEE WASPAA 2003, New Paltz, NY, October 2003, pp. 67-70.

[7] N. A. Gumerov and R. Duraiswami (2005). "Fast multipole methods for the Helmholtz equation in three dimensions", Elsevier, The Netherlands.

[8] R. O. Duda and W. L. Martens (1998). "Range dependence of the response of a spherical head model", J. Acoust. Soc. Am., vol. 104, no. 5, pp. 3048-3058.

[9] M. Abramowitz and I. Stegun (1964). "Handbook of mathematical functions", Government Printing Office.

[10] W. M. Hartmann (1999). "How we localize sound", Physics Today, November 1999, pp. 24-29.

[11] 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, no. 4, pp. 2202-2215.

[12] 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, no. 1, pp. 111-123.

[13] R. Duraiswami, Z. Li, D. N. Zotkin, E. Grassi, and N. A. Gumerov (2005). "Plane-wave decomposition analysis for the spherical microphone arrays", Proc. IEEE WASPAA 2005, New Paltz, NY, October 2005, pp. 150-153.

[14] 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.

[15] B. Rafaely (2005). "Analysis and design of spherical microphone arrays", IEEE Trans. Speech and Audio Proc., vol. 13, no. 1, pp. 135-143.

[16] H. Teutsch and W. Kellermann (2006). "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays", J. Acoust. Soc. Am., vol. 120, no. 5, pp. 2724-2736.

[17] Z. Li and R. Duraiswami (2007). "Flexible and optimal design of spherical microphone arrays for beamforming", IEEE Trans. Speech, Audio, and Language Proc., vol. 15, no. 2, pp. 702-714.

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

[19] J. Daniel. R. Nicol, and S. Moreau (2003). "Further investigation of high order Ambisonics and wavefield synthesis for holophonic sound imaging", Proc. AES 114th Conv., Amsterdam, The Netherlands, preprint #5788.

[20] 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 Conv., New York, NY, preprint #6540.

[21] Z. Li and R. Duraiswami (2005). "Hemispherical microphone arrays for sound capture and beamforming", Proc. IEEE WASPAA 2005, New Paltz, NY, pp. 106-109.