P-HRTF: Efficient Personalized HRTF Computation for High-Fidelity Spatial Sound

Alok Meshram*, Ravish Mehra†, Hongsheng Yang‡, Enrique Dunn§, Jan-Michael Frahm¶, Dinesh Manocha‖

Department of Computer Science, University of North Carolina, Chapel Hill

ABSTRACT

Accurate rendering of 3D spatial audio for interactive virtual auditory displays requires the use of personalized head-related transfer functions (HRTFs). We present a new approach to compute personalized HRTFs for any individual using a method that combines state-of-the-art image-based 3D modeling with an efficient numerical simulation pipeline. Our 3D modeling framework enables capture of the listener's head and torso using consumer-grade digital cameras to estimate a high-resolution non-parametric surface representation of the head, including the extended vicinity of the listener's ear. We leverage sparse structure from motion and dense surface reconstruction techniques to generate a 3D mesh. This mesh is used as input to a numeric sound propagation solver, which uses acoustic reciprocity and Kirchhoff surface integral representation to efficiently compute an individual's personalized HRTF. The overall computation takes tens of minutes on multi-core desktop machine. We have used our approach to compute the personalized HRTFs of few individuals, and we present our preliminary evaluation here. To the best of our knowledge, this is the first commodity technique that can be used to compute personalized HRTFs in a lab or home setting.

1 INTRODUCTION

The recent interest in the development of low-cost head-mounted displays (HMDs) has sparked consumer and commercial interest in production of immersive virtual reality (VR) experiences. While most of the developments in virtual and augmented reality have focused on the visual modality, the developers of these HMDs and content developers have also been emphasizing the importance of spatial sound and high-quality acoustic effects to increase the user's sense of presence in virtual environments. This is backed by studies that have demonstrated that sound is important in the formation of presence percepts [25], and that rendering of *spatial sound* through headphones increases presence in virtual environments [6, 47].

Head-related transfer functions (HRTFs) are commonly used in spatial sound rendering to model the acoustic filtering of sound by the human head and body. These functions are known to provide important spatial cues to the human auditory system [6, 53]. Sound filtered through HRTFs, when presented over headphones, can be used to simulate free-field (anechoic) listening for users, enabling perception of sound source location [50]. However, HRTFs depend on the geometry of the outer ears (pinnae), head, and torso, meaning that HRTFs vary between individuals. This variation is significant; use of generic or non-personalized HRTFs in spatial sound rendering can lead to localization errors, lateralization artifacts, and unconvincing spatial impressions [25, 49]. However, most VR systems or game engines use a *generic* HRTF for all users. This is because traditional measurement-based techniques to obtain individual HRTFs involve specialized, expensive equipment and settings [5, 54], and it is both difficult and expensive to compute personalized HRTFs for each user.

This has motivated the development of alternative approaches to obtaining personalized HRTFs. One approach is to compute personalized HRTFs by using the 3-D mesh of the head and torso as input to a numerical sound simulation technique to reproduce the HRTF measurement process. These numerical techniques are known to generate HRTFs that match well with measurements. However, they take tens of hours or days to compute an individual's HRTF using a desktop machine.

Generating personalized HRTFs using numerical sound simulation techniques has a key problem: the need for accurate 3D meshes of users or subjects. Traditionally, such meshes are obtained using laser scanners or MRIs, which can be expensive or time-consuming. Image-based 3D modeling, the reconstruction of dense 3D geometry from a set of photographs, offers a passive and cost-effective mechanism for obtaining object mesh models. Given the rapid development of image-based dense modeling techniques in computer vision [46, 21], one can robustly reconstruct a dense 3D object mesh for input images using automated techniques. In particular, it is not uncommon to have commodity cameras (e.g. in smartphones) with 10-20 megapixel resolution, which can be used to generate the input images for our technique. Moreover, the cameras can be easily moved around the object of interest and significantly reduce the intricacies of input capture setup. Accordingly, image-based modeling offers a favorable trade-off between ease of use and modeling accuracy. Many state-of-the-art methods attain very high reconstruction accuracy that are comparable to laser range sensors [46]. In this paper, we explore the use of image-based 3D modeling as a means to streamline the customization of HRTF's for specific subjects.

Main results We present an efficient personalized HRTF computation pipeline that combines a state-of-the-art image-based 3D modeling technique with an efficient numerical simulation pipeline based on the adaptive rectangular decomposition technique [42, 29]. The underlying image-based 3D modeling technique reduces the cost of acquiring 3-D meshes of individuals while generating meshes with good accuracy. These meshes are used by our numerical simulation pipeline which combines the principle of acoustic reciprocity with the adaptive rectangular decomposition-based numerical solver; it efficiently computes the full HRTF in tens of minutes on a multi-core desktop PC. In order to evaluate our pipeline, we use it to generate personalized HRTFs of five subjects and use them for spatial sound rendering in a virtual experience delivered through the Oculus Rift HMD and headphones. We present a preliminary evaluation of these personalized HRTFs by comparing them with the commonly used measured HRTF of the KEMAR dummy head. Overall, our HRTF computation pipeline can generate personalized HRTFs in a relatively small amount of time with-

*e-mail: alok@cs.unc.edu
†e-mail: ravishm@cs.unc.edu
‡e-mail: yhs@cs.unc.edu
§e-mail: dunn@cs.unc.edu
¶e-mail: jmf@cs.unc.edu
‖e-mail: dm@cs.unc.edu
website: http://gamma.cs.unc.edu/HRTF

out using expensive or specialized equipment. To the best of our knowledge, this is the first commodity solution to compute personalized HRTFs for spatial sound rendering.

2 Background

In this section, we provide some background on spatial hearing, sound localization, HRTFs, and their role in rendering spatial sound.

2.1 Spatial Hearing and Sound Localization

Spatial hearing refers to the ability of the human hearing system to perceive spatial aspects of our acoustic environment. This includes sound localization, our ability to associate a spatial position to a sound event or a sound source [6]. Multiple experiments performed to understand the nature of sound localization have related this ability to the presence of spatial cues present in sound signals reaching the ears [8]. We discuss some of these cues.

Interaural Differences
Lord Rayleigh's early work associated sound localization with interaural differences present between the signals received by the two ears [45]. Broadly, interaural time difference (ITD) is defined as the time delay between the arrival of a signal at each of the two eardrums, while interaural intensity difference (IID) is defined as the difference in total signal intensity at the two eardrums within a specific time period or within a particular frequency band. While ITD and IID values vary with source direction relative to the head, multiple source directions can correspond to the same ITD or IID value. As a result, depending solely on ITD and IID values for sound localization can lead to ambiguities such as front-back confusions [51]. Furthermore, the use of ITD and IID alone for spatial sound rendering does not provide the spectral cues that the pinnae produce; these cues are known to play a major role in elevation perception [4].

Spectral Filtering Due to the Pinnae
The pinnae (outer ears) are known to cause direction-dependent modification of the frequency spectrum of the sound signal received at the eardrum. This leads to spectral changes that depend on the relative direction of the source with respect to the head. These changes are known to play a role in sound localization [24]. Note that while ITD and IID are binaural cues that rely on difference between the signals received at the two ears, pinna modification of sound is a monaural cue that depends on the sound received at a single ear. There is evidence that monaural spectral cues are the major cues for determining vertical location of sound sources [33].

2.2 Head-Related Transfer Functions

The aforementioned cues can be thought of as consequences of the sound source's position relative to the listener as well as modification of the sound signal due to the human body before it reaches the ear drum. The *head related transfer function* (HRTF) represents the acoustic filtering, in free-field (anechoic) conditions, of the sound signal from a single source due to the human body (particularly the pinnae, head and torso) as measured at the entrance of the listener's ear canals. Since this acoustic filtering varies with source position and with the geometry of the listener's body, HRTFs are functions of source position and listener's geometry, along with frequency. Usually the dependence on listener geometry is left implicit, and HRTFs are represented only as functions of source positions and frequency. Furthermore, the dependence on source distance from the listener's head is also ignored as for relatively distant sources (1 m or greater), HRTFs are virtually independent of distance compared to direction [9]. For this reason, most HRTF measurements are reported for specific source distance from the head, usually 1 m.

Mathematically, assume that the listener's head is centered at the origin oriented along the positive X-axis, with the interaural axis (the line joining the two ear canals entrances) aligned with the Y-axis and the top of the head pointing along the positive Z-axis. Assuming free-field conditions, let $X_L(\theta,\phi,\omega)$ and $X_R(\theta,\phi,\omega)$ be the Fourier transforms of the signals received at the entrance of the listener's left and right blocked ear canals, respectively, due to a source 1 m away at at azimuth θ and elevation ϕ. Let $X_C(\theta,\phi,\omega)$ be the Fourier transform of the signal received at the origin due to the same source in the absence of the listener, also in free-field conditions. Then the left-ear and right-ear HRTFs can be defined as:

$$H_L(\theta,\phi,\omega)=\frac{X_L(\theta,\phi,\omega)}{X_C(\theta,\phi,\omega)},\quad H_R(\theta,\phi,\omega)=\frac{X_R(\theta,\phi,\omega)}{X_C(\theta,\phi,\omega),} \tag{1}$$

respectively [56]. The time-domain analogue of the HRTF is called the *Head-Related Impulse Response (HRIR)*, and is the inverse Fourier Transform of the corresponding HRTF.

HRTFs and 3-D Audio Systems
Presentation of sound filtered through HRTFs preserves spatial hearing cues, allowing the auditory system to attach spatial attributes to auditory events, just like in natural free-field listening [50]. This is based on the idea that a sound signal filtered with HRTFs, presented through headphones, mimics the free-field transmission of a sound signal from a source to the ears.

Mathematically, let $x(t)$ be the signal played by a sound source at azimuth θ and elevation ϕ. The left- and right-ear headphone signals, $y_l(t)$ and $y_r(t)$, respectively, after HRTF filtering can be expressed as:

$$y_l(t)=h_l(\theta,\phi,t)*x(t),\quad y_r(t)=h_r(\theta,\phi,t)*x(t), \tag{2}$$

respectively, where $*$ represents the convolution operation, and $h_l(\theta,\phi,t)$ and $h_r(\theta,\phi,t)$ are the left- and right- ear HRIRs corresponding to the source direction (θ,ϕ).

Because they are so useful in rendering sound with spatial hearing cues, HRTFs are often used as central components in 3-D audio systems [6].

HRTF Variation and Personalized 3-D Audio
As mentioned above, HRTFs vary from person to person, as the features of the head and the pinnae vary across people. This variation is significant; psychoacoustical experiments have demonstrated that using HRTFs other than a person's own leads to incorrect perception and to localization errors [49]. In other words, it is important to use the personalized HRTFs for a listener to generate high quality spatial sound.

3 Prior Work

In this section, we give a brief overview of prior work in HRTF personalization and in 3D mesh acquisition using image-based 3D modeling.

HRTF Measurements
Physical measurement is commonly used to obtain individual HRTFs in psychoacoustic research [50, 5, 54]. Typically, HRTF measurements are conducted in a large anechoic chamber. A set of high-quality speakers is arranged in a spherical pattern at a fixed distance from a point near the center of the room, with the speakers pointing towards the center. The user is seated so that the center of the line joining his/her ear canals (the inter-aural axis) coincides with the center of the sphere of speakers, and the inter-aural axis is horizontal with the user facing a specific forward direction. A high-quality probe microphone is inserted into each of the user's ears, either inside the ear canal or at the entrance of blocked ear

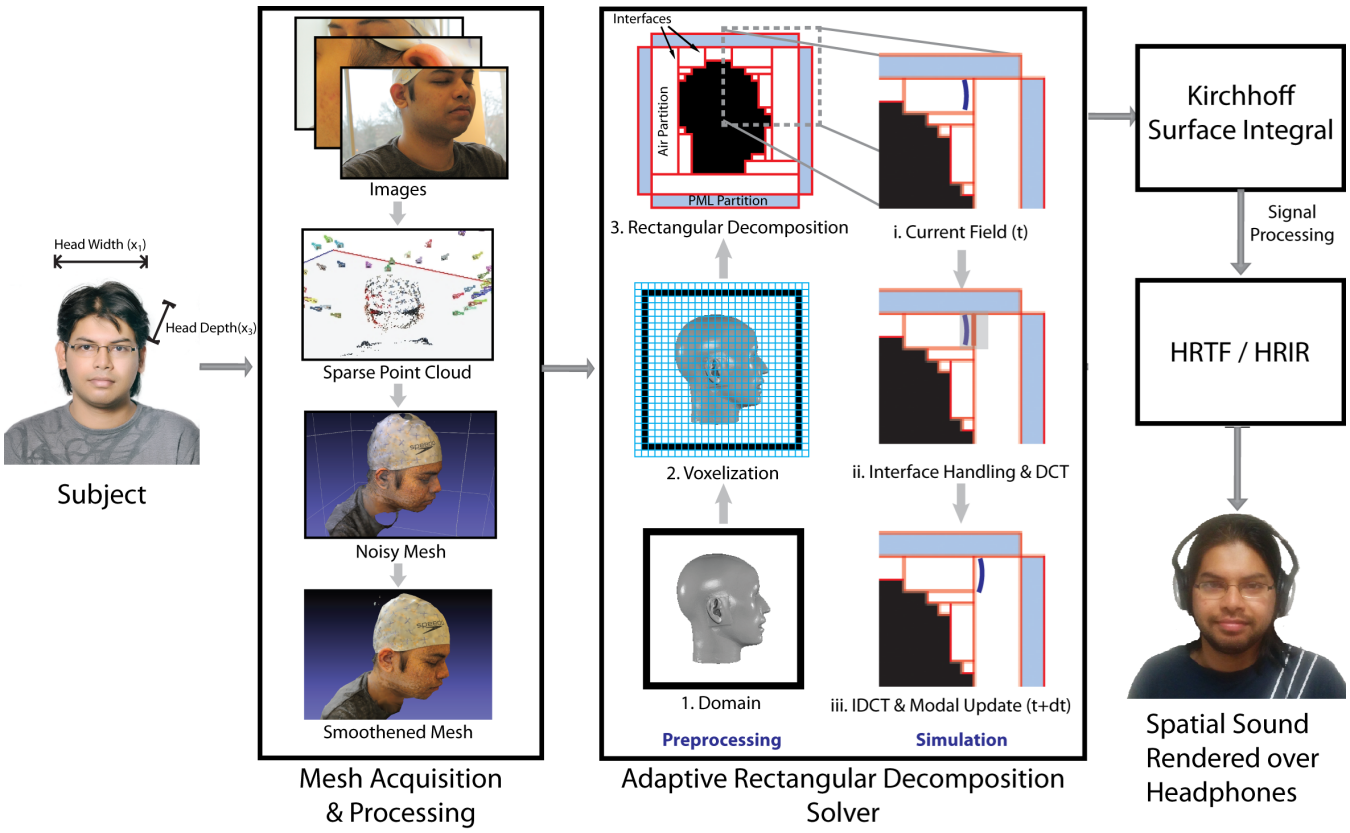

Figure 1: Block diagram showing an overview of our pipeline, with details of our mesh-acquisition technique and our numerical sound simulation technique.

canals. The user must sit still for the duration of the measurements (usually 30 minutes to an hour) during which a specific signal (such as Golay-coded signals) is played one by one from speakers at different directions around the head. The signals arriving at the microphone are recorded and stored. Next, these signals are post-processed to remove the effects of the room, speaker and microphone responses, leaving only the head-related impulse responses, which are then converted into head-related frequency responses by using a Fourier transform. Overall, the focus of most HRTF measurement methods is accuracy. They involve expensive and specialized equipment along with tedious procedures during which user must sit still. As a result, such measurements are conducted mainly for research purposes, and their widespread use is limited.

HRTF Computation Using Analytical Solutions

Over the last century, many researchers have derived exact mathematical solutions of the wave equation for acoustic scattering for shapes that approximate the head and torso. Lord Rayleigh derived the exact solution for the acoustic scattering of a plane wave by a completely rigid sphere [45]. Duda and Martens extended this solution to allow the source to be at variable distance from the sphere [10]. Jo et al. used the solution of acoustic scattering for spheroids, allowing variation of height and radius of the head [22]. Algazi et al. provided solutions for a "snowman" model where both the head and torso were represented by spheres [4]. Given measurements of head radius, head height and torso radius, these models can compute personalized HRTFs very quickly. However, because they approximate the complex geometries of the human body, these techniques generate HRTFs that only match measured HRTFs at low frequencies, missing detailed spectral features at high frequencies, especially due to the pinnae.

HRTF Computation Using Anthropometric Measurements

An alternate approach to generating personalized HRTFs is based on the idea that an individual's HRTF can be correlated to anthropometric measurements, such as head width, pinna height, pinna angles, torso width, etc. These techniques rely on a database of measured HRTFs that also provides anthropometric measurements of subjects. Zotkin et al.'s method uses anthropometric measurements to match a user to a subject in a HRTF database whose measurements are closest to the user's based on a distance metric. The subject's measured HRTF is then used for the user [55]. Another set of techniques computes new HRTFs for each user based on the correlation established between anthropometric measurements and HRTFs. Hu et al. use a back-propagation artificial neural network, trained using an HRTF database, to generate an individual's HRTF [20]. A very recent approach due to Bilinski et al. treats HRTF computation as a problem of finding a sparse representation of a subject's anthropometric features, based on a database containing anthropometric measurements and HRTF measurements. The database was newly generated by the authors, and for the purposes of anthropometric measurements they relied mostly on automatic computation from 3D models of the subjects. A new subject's measurements are used to generate a sparse representation which in turn is used to compute the HRTF for that subject [7].

HRTF Computation Using Numerical Simulation

Numerical sound simulation techniques have been used as an

approach for computing individual HRTFs. This approach requires as input the accurate 3-D mesh of an individual's body, in particular the head and torso. The simulation mimics the HRTF measurement process described above; the resultant pressure signals are then processed to compute the HRTFs. Kahana et al. used the boundary element method (BEM) to compute HRTFs of the KEMAR dummy head with six different pinna models as well as for an individual [23]. Gumerov et al. described a parallel fast multipole accelerated boundary element method (FMM-BEM) based technique to compute HRTFs [17]. They used the technique to compute HRTFs for the Fritz and KEMAR dummy heads and presented qualitative comparisons between the measured HRTFs and the computed HRTFs. This technique required tens of hours on a desktop machine. Mokhtari et al. presented an HRTF computation technique that combined the finite difference time domain (FDTD) method with the Kirchhoff-Helmholtz integral equation (KHIE); they used this technique to compute HRTFs for the KEMAR dummy head [35] as well as human subjects [34]. They presented qualitative as well as quantitative comparisons of the computed and measured HRTFs.

Mesh Acquisition

Computer vision 3D reconstruction is the process of mapping a set of input images to a 3D representation (e.g a 3D mesh) of their contents. The process by which this mapping is achieved is generally divided into two stages: structure from motion and dense shape estimation. Structure from motion (SfM) takes a set of images as input in order to estimate a scene 3D reference system, the viewing parameters of the cameras capturing the images (i.e. relative poses, individual focal length, etc.) as well as a sparse (feature-based) representation of the scene. Detailed reviews of the underlying mechanisms in SfM techniques can be found in [18, 27, 11]. Most state-of-the-art SfM systems [48, 52, 2, 12] share a common processing pipeline: 1) detect and match features between input images to construct 2D tracks; 2) incrementally estimate the set of viewing parameters and the structure that best explain the image measurements; 3) perform global model refinement through non-linear bundle adjustment optimization. In this work we use an SfM framework in the spirit of [52] and achieve highly accurate local registrations of the cameras around the user's head.

Dense reconstruction using multi-view stereo assumes a known camera registration to estimate the depth of the scene for every pixel in each image of the image set. The attained image-specific depth-fields are integrated into a single reference coordinate system and fused to generate a surface model. In our proposed method, we use the estimated 3D mesh representation to generate personalized HRTFs. There are a wide variety of dense multi-view stereo methods available from real-time multi-view stereo [14, 30] that perform high-accuracy offline stereo estimation [46]. In practice, depthmaps estimation commonly achieves accuracies within 2% of the scene's depth. In the case of our close-up additional sampling (10cm for the ear), this means our method will consistently provide estimates within 2mm of the ground truth. A detailed discussion pertaining the relationship between depth error, camera baseline and image resolution for stereo can be found in the work of Gallup et al. [13]. To overcome the noise in any particular depth map, a range of local stereo fusion methods have been developed for improved modeling [31, 15]. Given that we aim for a globally optimal solution of the 3D mesh used to compute the user's HRTFs, we leverage the global stereo fusion method of Jancosek and Pajdla [21] for producing a dense 3D mesh model of the head.

Overall, for the purposes of 3D reconstruction, our technique relies on a set of camera images which capture the user's head densely all around. These images are utilized by SfM, followed by dense reconstruction. Note that while in general we could use any 3D modeling technique that provides high fidelity 3D models with sufficient accuracy, there were some considerations that guided our choices. We highlight these considerations by discussing some recent approaches. The Dense Tracking and Mapping in real time (DTAM) technique by Newcombe et al. [37] is one such approach. DTAM inherently uses the redundancy of video data, which would require a significantly higher spatial sampling of the user's head resulting in at least an order of magnitude higher number of frames. As we demonstrated in our experiments, to achieve high quality meshes a much higher resolution ($>$ 6 MB) than HD video resolution (2MB) is required to obtain accurate meshes. While the DTAM paper did not evaluate the quality of the estimated 3D model, the resolution and accuracy relation will be similar. Another recent approach is Monofusion, by Pradeep et al. [40]. Monofusion is similar to our method (in terms of leveraging depth fusion) but it also leverages video data, requiring an order of magnitude more frames. As the authors point out, their method is optimized for speed rather than robustness. We opted for a robust global approach as the achieved model quality is critical and there is no real-time requirement for our approach.

4 Efficient HRTF Computation

Figure 1 provides a broad overview of our HRTF computation pipeline, showing its various components along with details of our mesh acquisition and numerical sound simulation techniques. In this section we provide descriptions of these components and their roles in the pipeline.

4.1 Mesh Acquisition Technique

To generate personalized HRTFs, it is critical to obtain accurate head and ear geometry of the user. To allow easy acquisition and a highly accurate model, we propose the use of commodity digital cameras for the acquisition of the head and ear geometry of the user. In all our experiments, we used images captured by an off-the-shelf digital SLR camera (Canon 60D) with image resolution (3456×2304). This allowed us to observe details of the skin texture, which were leveraged by multi-view stereo estimation modules to determine reliable dense correspondences.

To model the area of the head behind the ear, a critical area for computation of personalized HRTFs, the user wears a swim cap to hide his or her during the data capture. For precise modeling, the user's head was densely captured all around with samples at roughly every 15 degrees (see Figure 2). The selected angular separation between captures gives us at least three samples within a 30 degree range, which enables both robust feature matching and precise geometric triangulation. Moreover, this sampling provides us with sufficient overlap between the views to enable high-accuracy multi-view stereo estimation. Empirically, we found that sampling intervals larger than 15 degrees introduced severe aberrations into the resulting 3D model. To increase the model resolution around the ear, we specifically captured 20+ convergent close-up shots for each ear. Figure 3 depicts the effects of additional close-up redundant sampling to enhance the level of detail. From the captured images we calculated and matched SIFT features [26] for each image with its top K appearance nearest neighbors, as measured by the GIST descriptor [38]. Using these matches, we leverage a structure from motion algorithm in the spirit of VisualSFM [52] to perform the incremental structure from motion and bundle adjustment using the cameras internal calibration as provided by the EXIF data of the images. This step provides us the camera registration needed for the dense modeling of the scene.

Next, we perform the dense modeling of the user's head to obtain the desired mesh model required to compute personalized HRTFs. We opt for using a two tier computation that first estimates two-view depths maps [19]. Besides limited accuracy from two view depth maps, highlights on the users skin occur naturally. These are then eliminated by the next step the depth map fusion [21], which

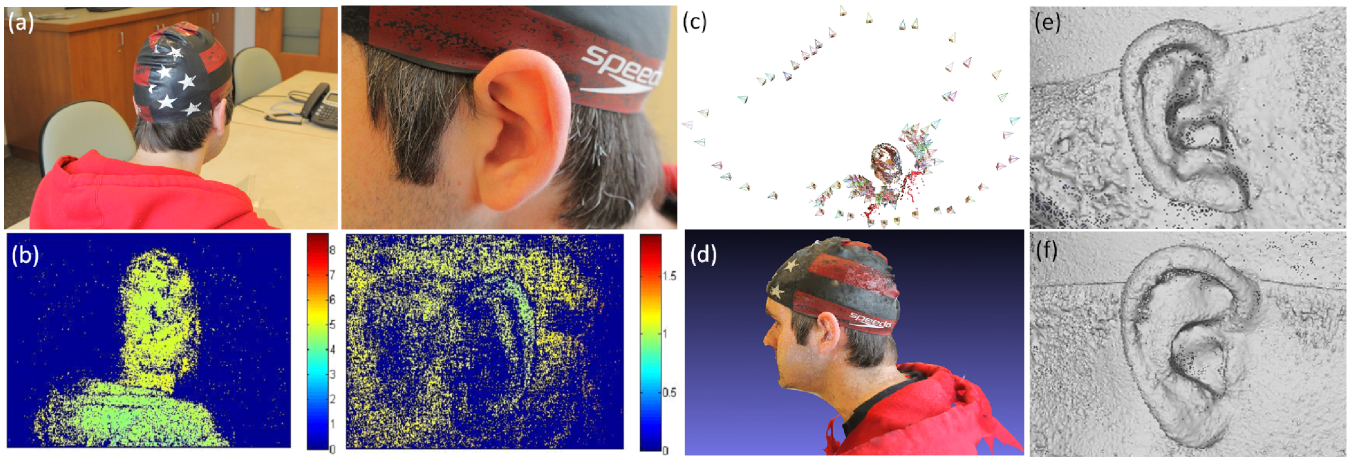

Figure 2: (a) Input example photos from different viewpoints and (b) their calculated relative dense depthmaps; (c) Registered camera trajectory and sparse 3D point clouds; (d) Visualization of resulting dense head mesh; (e),(f) Close-ups of pinna part from our two tested users.

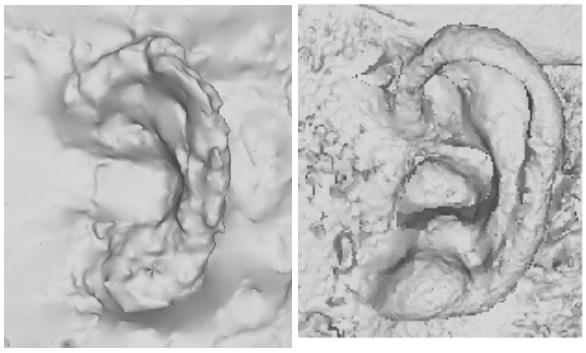

Figure 3: Effect of sampling configuration on the level of detail of the attained 3D model. Left: Detail of ear structure using uniform sampling. Right: Detail of enhanced ear structure attained with additional convergent close-up sampling over the ear region.

rejects the erroneous geometry resulting from highlights. To execute this two tier modeling, we leveraged the software of Jancosek and Pajdla [21]. The resulting mesh is then produced by applying a 3D Delaunay triangulation of dense point clouds and the construction of a graph based on the tetrahedrons from the Delaunay triangulation with weights set according to camera-vertex ray visibility. Then it further refines the graph's t-edge weights and obtains a water-tight dense surface mesh by using a graph-cut based labeling optimization to label each tetrahedron as inside or outside.

Before the generated surface mesh is used for the next step in our pipeline, we perform some mesh cleanup steps. First, since the generated mesh may not be to scale with the subject, we collect the subject's head width and head depth (anthropometric measurements x_1 and x_3 as described in [5]) and use it to scale the generated mesh. Next, we remove stray vertices and triangles from the main head mesh. Following this, we perform hole-filling using standard techniques to cover the holes in the mesh. Finally, we align and orient the mesh to match the alignment of the head during HRTF measurements (as described in section 2.2) and place the head mesh at the center of a cubical simulation domain. The cubical simulation domain is used as input for our next step.

4.2 Adaptive Rectangular Decomposition (ARD)

In the next step of our pipeline, we use an efficient numerical sound simulation technique called ARD [43] to compute the scattering of sound waves due to the mesh obtained at the end of the previous step. A related work describing the use of ARD for efficient HRTF computation for dummy heads using their scanned 3D models can be found in [32]. In the following sections we provide a brief overview of this sound simulation pipeline.

ARD performs sound propagation simulation by solving the acoustic wave equation. Like finite difference based methods, ARD divides the simulation domain into grid cells and computes sound wave pressure at each of those grid cells at each time step. However, compared to finite difference based methods, ARD has much less numerical dispersion error and is two orders of magnitude faster for homogeneous media. The principle behind ARD's efficiency and accuracy is the use of the exact numerical solution of the wave equation within rectangular (cuboidal) domains consisting of an isotropic, homogeneous, dissipation-free medium. As this solution is composed of cosines, ARD uses efficient Fast Fourier Transform (FFT) algorithms to compute propagation within the rectangular region. Note that for the purposes of computing HRTFs, the medium requirement is met as HRTFs are defined in free-field conditions with air as the medium.

The ARD block in Figure 1 shows the different stages of the ARD technique. In the preprocessing stage (left column in the ARD block in figure 1), ARD simulation generates a rectangular (cuboidal in 3D) decomposition of the computation domain. This decomposition is generated in a series of steps. First, the domain is voxelized to generate a grid of voxels. Next, the voxels containing the isotropic, homogeneous, dissipation-free medium (air) are grouped together to form rectangular regions called air partitions. Finally, absorbing boundary conditions are applied by using perfectly matched layer (PML) partitions at the boundary to simulate free-field conditions, as required by the HRTF definition. The simulation stage (right column in the ARD block in figure 1) consists of two updates: interface handling and mode update. The interface handling step is used to propagate sound across two adjacent partitions, which can be either air-air or air-PML partitions. These interface updates are based on finite difference stencils, as discussed in [43, 29]. The mode update step propagates sound within each air partition by updating FFT mode coefficients, using an update equation derived from the acoustic wave equation. Recently, a parallel version of ARD designed for large CPU clusters has been developed [36].

While ARD simulation can be used to achieve an exact replication of HRTF measurement (as described in section 2.1), this approach can result in large computational costs, requiring days to compute full HRTFs. We next describe the reasons for these costs and how we reduce them.

4.2.1 Acoustic Reciprocity

HRTFs are functions of source position, requiring multiple separate recordings of the signal at the ears due to different sound sources placed around the listener. Replicating this process through simulation requires multiple separate simulations, one for each source position (usually in the hundreds). We avoid this cost by making use of the acoustic reciprocity principle, which states that the acoustic response remains the same if we reverse the sense of source and receiver [39]. We therefore place *sources* at the receiver positions, inside the ears, and place *receivers* at the various source positions used in HRTF measurement. This reduces the required number of simulations to only two, one for each ear.

4.2.2 Kirchhoff Surface Integral Representation

As mentioned in section 2.2, HRTFs are measured at a distance of around 1 m, which is much greater than the typical size of the head. To compute pressure values at this distance, direct simulation requires a large, mostly-empty simulation domain. As the computational cost of ARD simulation increases cubically with simulation domain size, reducing the size of this simulation domain would result in significantly lower costs. We do this by making use of the Kirchoff surface integral representation (KSIR), which can be used to compute the pressure value at a point outside a simulation domain using pressure values on a cuboidal surface closely fitting the mesh [44]. Only pressure values at this surface need to be computed by ARD, thus significantly reducing the size of the domain as well as the computational cost.

4.3 Signal Processing

Since we use Gaussian impulse sources in our ARD simulations, the output of the KSIR calculation is a set of responses that correspond to the head mesh's scattering of Gaussian impulse sound. In order to convert these Gaussian impulse responses to HRIRs, we use a digital signal processing script that implements equation 1. Specifically, the frequency response of the Gaussian impulse signal at the center of the head in the absence of the head ($X_C(\theta, \phi, \omega)$ in equation 1) is removed from the head responses by this script in the frequency domain, and the HRIR is obtained by performing an inverse Fourier transform.

4.4 Spatial Sound Rendering

In order to perform spatial sound rendering using HRTFs, three steps need to be performed: (a) compute direction of incoming sound field at listener position; (b) model scattering of sound around the listeners head using HRTFs; (c) incorporate listeners head orientation. To compute the direction of the incoming sound field at the listener position, we use the plane wave-decomposition approach proposed by Mehra et al. [28]. This approach uses high-order derivatives of the pressure field at the listener position to compute the plane wave-decomposition of the sound field at interactive rates. This plane wave-decomposition is further represented in the spherical harmonic basis as discussed in their work. Scattering of sound around the head is modeled using the personalized HRTFs computed by our technique. Similar to Rafaely et al. [41], we also convert our HRTFs into spherical harmonic basis. By doing this, the listeners head rotation can be easily modeled using standard spherical harmonic rotation techniques [16]. Finally, as described in Rafaely et al. [41], the spatial sound for each ear can be computed as a simple dot product of the spherical harmonic coefficients of the plane-wave decomposition and the HRTF. This enables us to generate spatial sound at interactive rates.

We have integrated our spatial sound rendering pipeline with the Oculus Rift HMD and the Source SDK game engine. Our system uses the head orientation provided by the HMD to generate spatial sound and renders it over the headphones. The users position is controlled through an XBox 360 wireless controller. Figure 4 shows a typical use-case scenario of our system. Therefore, our system allows a user to play the VR game with 3D spatial sound, rendered using personalized HRTFs.

5 Evaluation

In this section we present a preliminary qualitative evaluation of personalized HRTFs computed by our pipeline. We compare the HRTFs computed for five subjects with the measured HRTF of KEMAR, a dummy head whose HRTFs are commonly used in spatial sound rendering. Before we present these results, we discuss some pertinent parameters used to generate the HRTFs presented in this evaluation.

5.1 Pipeline Parameters and Details

The images used by our image-based 3D modeling technique were taken using a Canon 60D DSLR camera with an image resolution of 3456×2304. Subjects were asked to wear a swimming cap to cover their hair. This was done to avoid problems in the 3D mesh due to hair, especially near the ears. Head width and head depth were measured for the subjects, and these measurements were used to scale the meshes. Table 5.1 presents the head dimensions of the five subjects, along with the number of images captured to generate their 3D meshes and the complexity of the generated meshes (given in number of triangles).

Once computed, the meshes were manually processed to remove the torso portion below the neck. This step was performed because torso data was often incomplete; the focus of our images was the subject's head and ears.

ARD simulations were run with a grid cell size of 2 mm, which was chosen to adequately represent the curves of the pinna and the head after voxelization. The absorption coefficient of the mesh surface was set to 0.02 to correspond to that of the human skin, as reported by Ackerman et al. [1]. The simulations were run to generate impulse responses of 4.5 ms duration to match the duration of the measured KEMAR HRTFs used as part of our test [5].

Timing results are as follows. Image capture of a subject's head requires $2-3$ minutes, while camera pose estimation requires $5-10$ minutes and dense 3-D reconstruction & mesh generation takes about 2 hours on a single core machine.

In terms of our sound simulation pipeline, the simulation time for computing the HRTF for all subjects was about 20 minutes on a desktop machine with an 8 core, 3.40 GHz CPU.

5.2 Qualitative Comparison

Figure 5 presents a qualitative comparison, within the horizontal plane, between the measured left-ear HRTF of the KEMAR dummy head and the computed left-ear HRTFs of the 5 subjects. This comparison is presented in the frequency range of 700 Hz - 14 kHz, as outside this range the measured data is considered unreliable [17].

At low frequencies, the wavelength of sound waves is comparable to the overall size of the head, and only the broad spheroidal shape of the head is responsible for the acoustic scattering characterized by the HRTF; the effects of the pinnae start occurring at around 3 kHz [3]. This results in similar features in HRTFs at low frequencies, as individual differences in head geometry, which are at a smaller scale, don't contribute. This effect can be observed in the HRTFs computed by our pipeline in the approximate frequency range of 700 Hz - 3 kHz in figure 5. Note that the finger-like projections between the azimuth range of 40 to 145 degrees are found

Figure 4: Typical use-case scenario for the personalized HRTFs computed by our pipeline. The user is being delivered a virtual experience through the HMD and through spatial sound, rendered using the personalized HRTF computed for them by our technique.

in the measured HRTF of KEMAR and in the HRTFs computed by our technique.

At high frequencies, individual differences in pinna geometry and in head geometry start contributing to acoustic scattering, resulting in wider variation of HRTF features. This is also reflected in our computed HRTFs. Especially above $8-9$ kHz, significant variation between individual HRTFs can be observed.

Overall, the qualitative comparison in figure 5 shows that the HRTFs computed by our pipeline fall into ranges similar to those of a commonly used measured HRTF, showing expected matches at low frequencies and expected variation at high frequency.

6 Conclusion and Future Work

We presented a pipeline that efficiently computes personalized HRTFs. Our pipeline uses a set of images captured using a commodity camera to generate high-quality meshes to be used for personalized HRTF computation. Furthermore, our approach takes tens of minutes of simulation, on contrast to the tens of hours or days required by previous simulation-based techniques.

We also presented a preliminary evaluation of personalized HRTFs computed using our method for five subjects. At low frequencies, we observed expected feature matches with a commonly used measured HRTF. However, this comparison does not present an evaluation of the high-frequency components of the computed HRTFs. Evaluation of high-frequency components of the computed HRTFs would require measured HRTFs of the same subjects. Therefore, our first goal for future work is to use our pipeline to compute HRTFs for subjects whose measurements are available, thereby enabling a more thorough evaluation of our approach.

Furthermore, as one of the primary goals of personalized HRTFs is to render spatial sound, another future direction for our pipeline involves conducting a user study to surveys qualitative as well as quantitative responses of subjects to spatial sound rendered through the personalized HRTFs generated for them by our pipeline.

Acknowledgements

This research was supported by Link Foundation Fellowship in Advanced Simulation and Training, ARO Contracts W911NF-10-1-0506, W911NF-12-1-0430, W911NF-13-C-0037, and the National Science Foundation grants 0917040, 1320644, and 1349074.

References

[1] E. Ackerman and F. Oda. Acoustic absorption coefficients of human body surfaces. Technical report, DTIC Document, 1962.

[2] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski. Building Rome in a Day. *Comm. ACM*, 2011.

[3] V. R. Algazi, C. Avendano, and R. O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122, 2001.

[4] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112:2053, 2002.

[5] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102. IEEE, 2001.

[6] D. R. Begault. *3-D Sound for Virtual Reality and Multimedia*. Academic Press, 2000.

[7] P. Bilinski, J. Ahrens, M. R. Thomas, I. J. Tashev, and J. C. Platt. HRTF magnitude synthesis via sparse representation of anthropometric features. ICASSP, 2014.

[8] J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

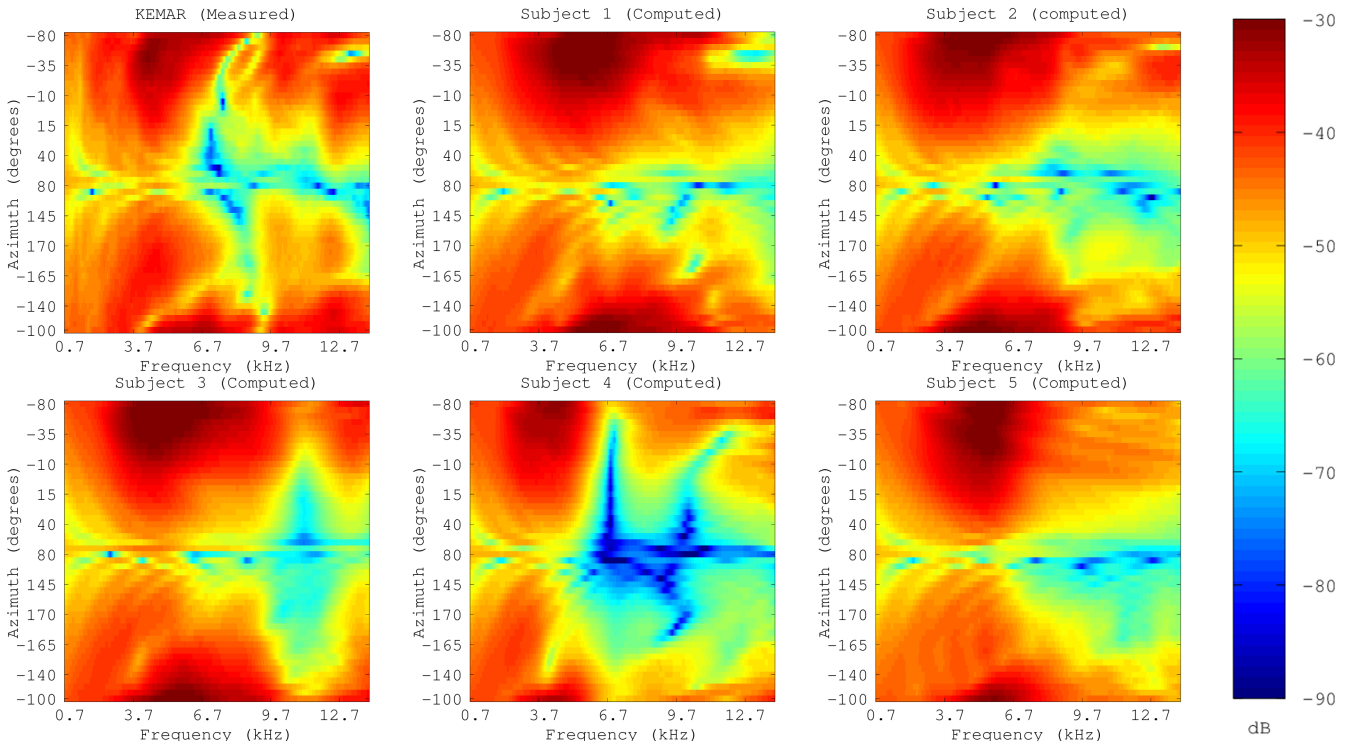

Figure 5: Plots showing the measured left-ear HRTF of KEMAR as well as the computed left-ear HRTFs of five subjects at 50 azimuths in the horizontal plane. 0 degree azimuth is in front of the listener, with positive azimuth to the right, and negative azimuth to the left of the listener. At low frequencies, HRTFs generally have similar features as individual differences in head and pinna geometry are at a smaller scale, which primarily affect high frequencies. This is observed in the HRTFs generated by our technique where features at low frequencies (below 3 kHz), such as the finger-like projections around 80 degree azimuth, can be observed in all computed HRTFs as well as the measured HRTF of KEMAR.

[9] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106:1465, 1999.

[10] R. O. Duda and W. L. Martens. Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104:3048, 1998.

[11] D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[12] J. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. *ECCV*, 2010.

[13] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys. Variable baseline/resolution stereo. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[14] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[15] D. Gallup, M. Pollefeys, and J.-M. Frahm. 3d reconstruction using an n-layer heightmap. In *Pattern Recognition*, pages 1–10. Springer, 2010.

[16] R. Green. Spherical Harmonic Lighting: The Gritty Details. *Archives of the Game Developers Conference*, Mar. 2003.

[17] N. A. Gumerov, A. E. ODonovan, R. Duraiswami, and D. N. Zotkin. Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation. *The Journal of the Acoustical Society of America*, 127:370, 2010.

[18] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[19] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):328–341, Feb 2008.

[20] H. Hu, L. Zhou, H. Ma, and Z. Wu. HRTF personalization based on artificial neural network in individual virtual auditory space. *Applied Acoustics*, 69(2):163–172, 2008.

[21] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[22] H. Jo and Y. S. Park. Signal processing: Aproximation of head related transfer function using prolate spheroidal head model. *Proceedings of the 15th International Congress on Sound and Vibration*, 2008(1):2963–2970, 2008.

[23] Y. Kahana and P. A. Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *Journal of sound and vibration*, 300(3):552–579, 2007.

[24] E. H. Langendijk and A. W. Bronkhorst. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112:1583, 2002.

[25] P. Larsson, A. Väljamäe, D. Västfjäll, A. Tajadura-Jiménez, and M. Kleiner. Auditory-induced presence in mixed reality environments and related technology. In *The Engineering of Mixed Reality Systems*, pages 143–163. Springer, 2010.

[26] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[27] Y. Ma. *An invitation to 3-d vision: from images to geometric models*, volume 26. springer, 2004.

[28] R. Mehra, L. Antani, S. Kim, and D. Manocha. Source and listener directivity for interactive wave-based sound propagation. *Visualization and Computer Graphics, IEEE Transactions on*, 20(4):495–503, April 2014.

[29] R. Mehra, N. Raghuvanshi, L. Savioja, M. C. Lin, and D. Manocha. An efficient GPU-based time domain solver for the acoustic wave

Subject No.	Head Width (x_1) in mm	Head Depth in mm(x_3)	No. of Images Used in Estimated Dense Model	No. of Triangles in Generated Mesh
1	148	178	115	3,105,365
2	164	196	95	6,935,830
3	170	179	108	3,671,849
4	153	196	129	5,256,111
5	135	176	85	4,006,730

Table 1: Relevant parameters used for generating head meshes for numerical sound simulation for five subjects.

equation. *Applied Acoustics*, 73(2):83–94, 2012.

[30] X. Mei, X. Sun, M. Zhou, H. Wang, X. Zhang, et al. On building an accurate stereo matching system on graphics hardware. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 467–474. IEEE, 2011.

[31] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[32] A. Meshram, R. Mehra, and D. Manocha. Efcient HRTF computation using adaptive rectangular decomposition. In *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society, to appear.

[33] J. C. Middlebrooks and D. M. Green. Sound localization by human listeners. *Annual review of psychology*, 42(1):135–159, 1991.

[34] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Computer simulation of HRTFs for personalization of 3d audio. In *Universal Communication, 2008. ISUC'08. Second International Symposium on*, pages 435–440. IEEE, 2008.

[35] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Computer simulation of KEMAR's head-related transfer functions: verification with measurements and acoustic effects of modifying head shape and pinna concavity. *Principles and Applications of Spatial Hearing*, pages 179–194, 2010.

[36] N. Morales, R. Mehra, and D. Manocha. MPARD: A parallel ARD-based wave simulator for distributed memory architectures. Technical report, Department of Computer Science, UNC Chapel Hill, 2014.

[37] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011.

[38] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[39] A. D. Pierce. *Acoustics: An Introduction to Its Physical Principles and Applications*. Acoustical Society of America, 1989.

[40] V. Pradeep, C. Rhemann, S. Izadi, C. Zach, M. Bleyer, and S. Bathiche. MonoFusion: Real-time 3d reconstruction of small scenes with a single web camera. In *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, pages 83–88. IEEE, 2013.

[41] B. Rafaely and A. Avni. Interaural cross correlation in a sound field represented by spherical harmonics. *The Journal of the Acoustical Society of America*, 127(2):823–828, 2010.

[42] N. Raghuvanshi, R. Narain, and M. C. Lin. Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):789–801, 2009.

[43] N. Raghuvanshi, R. Narain, and M. C. Lin. Efficient and accurate sound propagation using adaptive rectangular decomposition. *Visualization and Computer Graphics, IEEE Transactions on*, 15(5):789–801, 2009.

[44] O. M. Ramahi. Near-and far-field calculations in FDTD simulations using kirchhoff surface integral representation. *Antennas and Propagation, IEEE Transactions on*, 45(5):753–759, 1997.

[45] L. Rayleigh. XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

[46] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528, June 2006.

[47] R. D. Shilling and B. Shinn-Cunningham. Virtual auditory displays. Technical report, DTIC Document, 2000.

[48] N. Snavely. Bundler: Structure from motion (sfm) for unordered image collections. http://www.cs.cornell.edu/ snavely/bundler.

[49] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94:111, 1993.

[50] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85:858, 1989.

[51] F. L. Wightman and D. J. Kistler. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105:2841, 1999.

[52] C. Wu. Visualsfm: A visual structure from motion system. http://ccwu.me/vsfm/, 2011.

[53] B. Xie. *Head-related transfer function and virtual auditory display*. Plantation, FL: J. Ross Publishing, 2013.

[54] B. Xie, X. Zhong, D. Rao, and Z. Liang. Head-related transfer function database and its analyses. *Science in China Series G: Physics, Mechanics and Astronomy*, 50(3):267–280, 2007.

[55] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis. HRTF personalization using anthropometric measurements. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 157–160. Ieee, 2003.

[56] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Rendering localized spatial audio in a virtual auditory space. *Multimedia, IEEE Transactions on*, 6(4):553–564, 2004.