Use of 3D Head Shape for Personalized Binaural Audio

Philip J. B. Jackson[1], and Naveen K. Desiraju[1]

[1]*CVSSP, Dept. of Electronic Engineering, University of Surrey, UK*

Correspondence should be addressed to Philip Jackson (p.jackson@surrey.ac.uk)

ABSTRACT

Natural-sounding reproduction of sound over headphones requires accurate estimation of an individual's Head-Related Impulse Responses (HRIRs), capturing details relating to the size and shape of the body, head and ears. A stereo-vision face capture system was used to obtain 3D geometry, which provided surface data for boundary element method (BEM) acoustical simulation. Audio recordings were filtered by the output HRIRs to generate samples for a comparative listening test alongside samples generated with dummy-head HRIRs. Preliminary assessment showed better localization judgements with the personalized HRIRs by the corresponding participant, whereas other listeners performed better with dummy-head HRIRs, which is consistent with expectations for personalized HRIRs. The use of visual measurements for enhancing users' auditory experience merits investigation with additional participants.

1. INTRODUCTION

The use of headphones or earphones for consumption of audio and multimedia content offers many practical advantages, such as increasing the user's sense of immersion, suppressing interference, maintaining privacy and enabling mobility. On the other hand, such direct access to listeners' ears gives sound designers opportunities to create soundscapes to enrich the listening experience. Human perception of sound exploits various monaural and binaural cues for purposes of source localization and spatial awareness, e.g., Interaural Time Difference (ITD), Interaural Level Difference (ILD), and spectral coloration, including the peaks and valleys in acoustical response produced as a result of the listener's torso, head and outer ears (pinnae) [3, 6]. These spectral details can be captured in so-called Head-Related Transfer Functions (HRTFs) or Impulse Responses (HRIRs) [14], which differ enough amongst individuals that significant improvements can achieved using personalized HRIRs compared with those of a dummy head or average person. Therefore, to realize the full benefits of binaural reproduction for immersive gaming, an individual requires accurate estimates of his/her own HRIRs over the audible frequency range.

This paper investigates the use of one person's surface geometry for calculating his HRIRs through acoustical simulation [8]. The ear, head and torso geometries were obtained via 3D video techniques, aligned and converted into a solid 3D mesh, whose acoustical response was computed by the boundary element method (BEM) [9, 7]. Various resolutions of the mesh components were tested. By combining the responses at multiple frequencies, time-domain HRIRs for the left and right ears were obtained and utlized to synthesize audio samples. For comparison, acoustically-measured HRIRs of a dummy head were also employed [5]. These acted as stimuli in subjective listening tests that were conducted to assess localization accuracy, including the impact of personalization.

The rest of this paper is organized as follows. Section 2 outlines the visual capture method, mesh alignment, BEM configuration and acoustical simulation. Section 3 presents the simulated acoustical responses. Section 4 describes the subjective evaluation and gives the results of the assessment. Section 5 concludes.

2. METHOD

The acoustical response of the human ear to incoming sound from a point source in space is characterized by the head related impulse response (HRIR), or the head related transfer function (HRTF), which is the Fourier transform of HRIR [3, 5]. Specifically, the HRTF is defined as the acoustic filter from a sound source to a

defined position in the left or right ear canal, and describes the direction-dependent reflection, distortion and diffusing effects of sound due to an individual's head, torso, and pinna [6]. This paper investigates the use of one person's surface geometry for calculating his HRIRs through boundary element method (BEM) acoustical simulation in the frequency domain [8, 9].

The head and ear geometries were obtained via 3D video techniques using a commercial-off-the-shelf 3dMD face capture system [2], and the torso approximated from Microsoft Kinect depth images [10]. The surface geometries of each component were inspected to eliminate artifacts and aligned manually. MeshLab was employed to connect and complete the components, yielding a solid 3D mesh of the head and torso [1]. The mesh was imported into Matlab for use with the OpenBEM software [7]. HRTFs were computed for frequencies at regular intervals up to 8 kHz to populate a frequency-domain response function and then converted into the time-domain by inverse discrete Fourier transform to give each left and right pair of HRIRs. HRIRs were computed at 5-degree intervals around the head in the horizontal plane and assessed for several resolutions of the geometrical 3D-mesh data.

BEM is a numerical method for solving linear partial differential equations (PDE) reformulated as a set of integral equations, one defining the boundary and another relating the solution at the boundary to all points in the domain [9]. The BEM is derived by discretization of these integral equations. Compared with the Finite Element Method (FEM) and the Finite Difference Method (FDM), BEM is more computationally efficient and generally much easier to use [4]. However, BEM is restricted to homogeneous and linear PDEs. The major advantage is that in BEM, only the boundary of the domain of the PDE requires discretization (or sub-division) to produce a boundary mesh. For problems where the domain exists exterior to a boundary, such as a problem involving an acoustic field generated by a sound source, the size of the domain is infinite, but it has a limited boundary. In such cases, as here, the BEM is particularly helpful, as the problem reduces to solving a series of Helmholtz equations over a limited boundary surface.

2.1. Mesh generation

The first task in this study was to capture the geometry of the head, ears and torso of an individual user. The face capturing session was performed in the Visual Media Lab in the Centre for Vision Speech & Signal Processing (CVSSP). The apparatus used for this procedure was the 3dMDface System [2], which has been widely used for many years in the medical and dental industry for anthropomorphic analysis and monitoring of patients. Figure 1 shows the set-up for the 3dMDface system. It enables the user to record accurately an individual's natural head position and details of his/her facial features. It employs two camera pods, on either side of the head, each of which has three cameras with a 50 mm Navtar Raptar lenses. Of these six cameras, four have filters for near infrared placed on them; the remaining two are normal colour cameras. The rig includes two Opti Solar 250 projectors with infrared filters that project a speckle pattern. This system provides coverage of the face from ear to ear; it does not capture hair. Simultaneous images from all the cameras are combined to reconstruct a consolidated mesh representing the overlapped regions of the face that are contained in the images.

Fig. 1: Face capture rig with lighting in the lab.

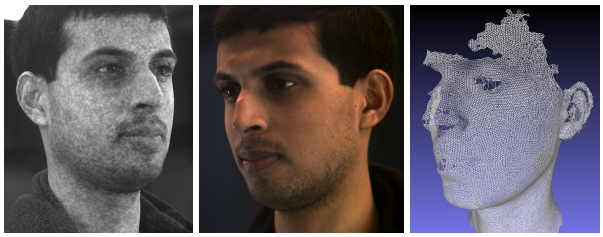

Fig. 2: Infrared (left) and color (center) images, and output mesh (right) from the face capture system.

Examples of the face capture are shown in Figure 2. The meshes output by the 3dMDface system have certain characteristics. Meshes only capture the skin and do not capture the top or back of the head, so only a partial mesh is obtained for the head. Meshes are most detailed near the cameras' centre of focus. The capture region is lim-

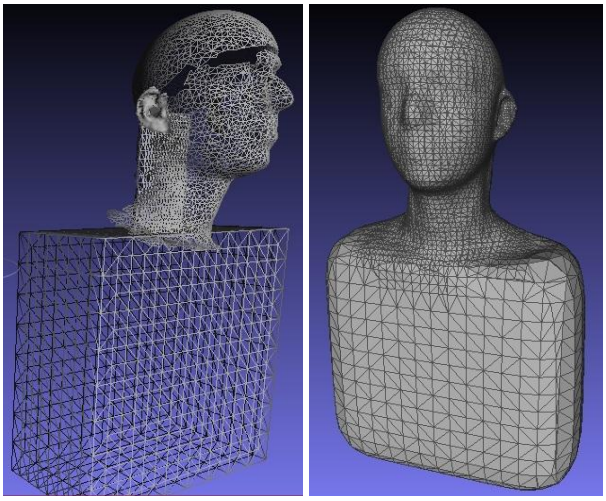

Fig. 3: Head and torso construction (left) and completed mesh (right).

ited to the front of the face and the top of the shoulders. Each face mesh is composed of approximately 40 000 triangular elements, connecting 20 000 nodes. The mesh resolution is defined by the distance between adjacent nodes, which was 1.7 mm on average for frontal face. To maximize resolution in the vicinity of the ears, separate ear meshes were captured, which yielded an average 1.3 mm inter-node distance.

A larger complementary mesh of the entire head and torso was obtained with a Kinect [10] with four times the number of elements and nodes, and a maximum 3.4 mm inter-node distance. This mesh facilitated the creation of synthetic mesh patches in the MeshLab utility and acted as a guide for their integration [1]. The face and ear meshes from the 3dMDface system were manually aligned and stitched together with the synthetic meshes by Poisson reconstruction, as shown in Figure 3.

3. SIMULATIONS

The frequency range of reliable simulations is limited by the resolution of the mesh, which provides discrete sampling of the geometry. It has been found that the upper frequency limit for which a mesh is valid determined by a critical wavelength that is 1/6th of the smallest inter-node distance [8]. However, acoustical simulation is computationally intensive with the memory usage and processor time directly linked to the number of el-

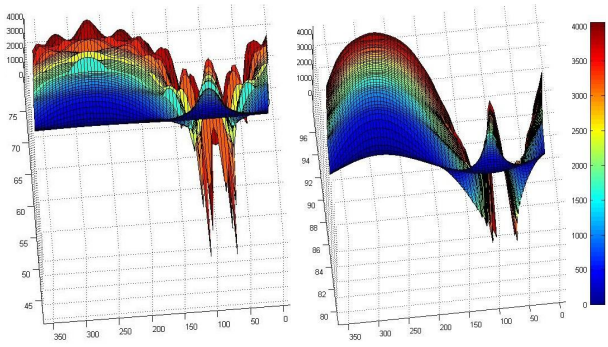

Fig. 4: Frequency response for spherical meshes at 270°, coarse (left) and fine (right): sound pressure magnitude (dB) versus azimuth (degree) and frequency (Hz).

ements in the mesh. Therefore, a series of experiments was conducted for the range 0 Hz–4 kHz to compare the simulated effects of resolution of the ear, the head and the torso. Each HRTF simulation took 200–400 hours to compute on a PC. From the original, full resolution meshes, lower resolution versions were made by merging nodes. Tests were performed on ear meshes discretized at 1 mm, 2 mm and 5 mm resolution, on head meshes with 5 mm, 8 mm and 10 mm resolution, and with/without the 10-mm-resolution torso mesh. In all cases, the elements at the entrance of ear canals were identified for analysis.

Given the arbitrary closed geometry of the test meshes, the 3D-BEM package was applied treating the mesh as a rigid body with zero particle velocity normal to the boundary [7]. For the sakes of observing the effect of mesh resolution and of comparison, we first ran simulations with spherical meshes.

3.1. Pilot test with spherical meshes

Simulation results were obtained for spherical meshes with various element sizes. The sphere diameter was 0.2 m. To represent the left ear and right ear positions, two points were selected at 270° and 90°, respectively. Figure 4 shows the 'left-ear' results up to 4 kHz for those with average inter-element distances of 33 mm (coarse) and 8 mm (fine). The results demonstrate a clear relationship between mesh resolution and the cut-off frequency of reliable HRIR simulations. According to the critical wavelength, it is expected that the coarse mesh will have discretization artifacts above 1.7 kHz, whereas the fine mesh is would not encounter such effects until 7.6 kHz.

Compared to the smooth HRTF obtained for the finely

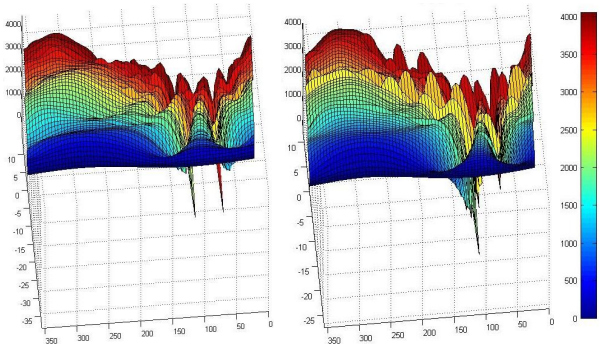

Fig. 5: Left-ear HRTFs of captured mesh with (left) and without (right) the torso: sound pressure magnitude (dB) versus azimuth (degree) and frequency (Hz). Resolution: head 8 mm, ear 2 mm, torso 10 mm.

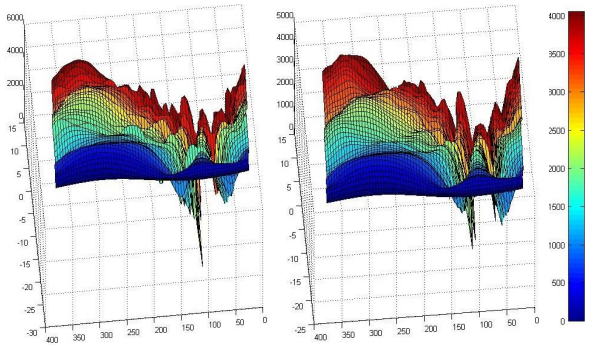

Fig. 6: Left-ear HRTFs of captured mesh with coarse (left) and fine (right) detail of the head: sound pressure magnitude (dB) versus azimuth (degree) and frequency (Hz). Resolution: head 10 mm and 5 mm, ear 2 mm, torso 10 mm.

sampled sphere in Fig. 4 (right), the cyan peaks at azimuths of 270°, 150° and 30° in Fig. 4 (left), and the yellow, orange and red ones at higher frequencies and additional azimuths are indications of specular artifacts resulting from the discretization. It is concluded that practical specification of the mesh size determines the validity of the acoustical simulations and provides a frequency limit on the HRIR estimation.

3.2. Effect of torso

To investigate the influence of the torso on the HRTF, tests were conducted on a head-and-torso mesh and an isolated-head mesh. The sound source was placed at a radial distance of 1.4 m from the origin in the center of the head.

The results, plotted in Figure 5, show significant modulations of the HRTFs above 900 Hz, which were not seen for the sphere. The detailed pattern of the peaks and valleys differs between the with and without torso cases, especially in the range 2.5 kHz–2.7 kHz in Fig. 5 (right), although the primary and secondary ridges either side of 90° share some common features.

3.3. Ear mesh resolution

Since the outer ear regions are small relative to the head and torso, it is possible to employ a much finer resolution on the pinnae without significantly affecting the computation time. Results were obtained by simulating ear meshes with 1 mm, 2 mm and 5 mm resolution. These revealed only subtle differences in the shape of the HRTFs (not shown here), but some influence on the

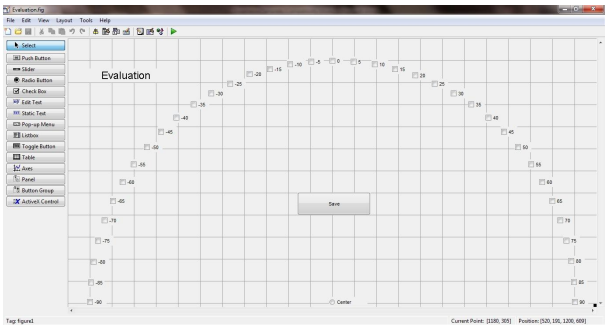

Fig. 7: GUI for evaluation of perceived azimuth angle.

overall magnitude. This is not surprising since the critical frequencies for these resolutions (56 kHz, 28 kHz and 11 kHz, respectively) are well above 4 kHz. In the subjective evaluation described in Section 4, the finest resolution was selected.

3.4. Head mesh resolution

Comparing the results of HRTF simulation with different head resolutions, a significant reduction in artifacts was observed at the finest resolution, even though the critical frequency was above 4 kHz in all cases. The results in Figure 6 compare the 5-mm (fine) and 10-mm (coarse) resolution meshes. The results are broadly in agreement. Yet, for the coarse mesh (Fig. 6 (left)) there are some (cyan) artifacts at approximately 1.6 kHz, which can be clearly seen at 220° and 120°. Once again, the fine resolution was chosen for the subjective evaluation below.

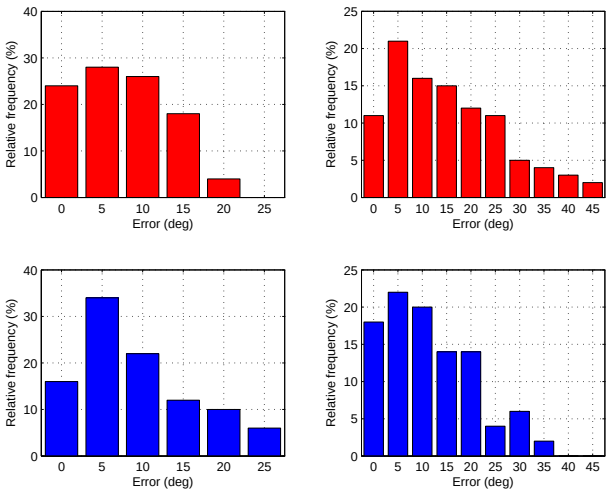

Fig. 8: Histograms of azimuthal error magnitude from lateralization tests: responses by meshed subject for (upper, left) BEM simulated and (lower, left) measured KEMAR HRIRs, responses by other participants for (upper, right) simulated and (lower, right) measured HRIRs.

4. SUBJECTIVE EVALUATION

The HRTFs were simulated with 5-mm head resolution, 1-mm ear resolution and the 10-mm resolution torso. The frequency range was extended up to 8 kHz to provide a wide-band listening evaluation. Binaural audio samples were generated by convolving a number of monaural speech recordings with the computed HRIRs, at azimuths ranging from 90-degrees to the left to 90-degrees to the right in 5-degree steps. The stimuli were formed from 6 selected recordings of English speech by native speakers. For comparison, samples were similarly generated using acoustically-measured HRIRs of a KEMAR dummy head [5]. Participants in the listening tests were required to localize the sound source from the binaurally-presented stimuli by choosing one of the 37 intended directions in a Matlab-based GUI (see Fig. 7). Stimuli were assessed in listening tests over headphones alongside those generated from dummy-head HRTFs with randomized presentation, by a number of listeners: the person whose geometry was captured, and 10 other participants.

The localization error was defined for each stimulus as the difference between the azimuth of the HRTF and the response recorded from the participant. Histograms of the magnitude of the errors are shown in Figure 8. Analysis of the results reveals that the majority of the audio signals were identified within a range of 10° either side of the original signal direction. The dispersion shows that it was difficult to obtain the exact localisation of the sound sources. These errors can be attributed to the limited resolution of the mesh and its approximation as rigid body. The BEM-simulated HRTFs provided better performance for the meshed participant as compared to the KEMAR HRTFs: only ∼20% of the responses were outside the range of 10° for the simulated HRTF dataset, as compared to ∼30% for the measured dataset. The mean error was reduced from 9.2° to 7.5° through HRTF individualization. For the other participants, the mean error increased from 11.5° to 15.1° in comparison of the measured dummy-head HRTFs and those simulated for a different individual.

In summary, a preference is exhibited for the personalized HRTF by the corresponding participant (N=1), but for the dummy-head HRTFs by the other listeners (N=10). This finding confirms expectations [13, 12, 11], and provides a positive indication for the use of stereo image processing with BEM simulation for personalization.

5. CONCLUSION

This paper presents a procedure to obtain HRIRs for an individual utilizing commercially-available 3D vision systems to acquire ear, head and torso geometry and BEM acoustical simulation. Simulation tests with captured geometry investigated the effects of torso, ear and head meshes at various resolutions. With an appropriate resolution for these components, the HRIRs were computed from the integrated mesh and convolved with speech signals to create stimuli for subjective evaluation. Listening tests provided validation of the simulated HRIRs in confirming that the participant whose geometry was meshed could locate sources more accurately with them than with measured dummy-head HRIRs. In contrast, other participants recorded better accuracy with the measured HRIRs, as expected. Further validation is warranted against acoustical measurements of individual HRIRs and with multiple personalized meshes. It would be interesting to investigate means of extending the simulations' frequency range and reducing computational requirements, and, from an applications perspective, to quantify the effect on users' presence and immersion during game play.

6. ACKNOWLEDGMENTS

Thanks to CVSSP for the use of its visual media and computing facilities; to Martin Klaudiny and Charles Malleson for assistance and provision of mesh data with the 3dMDface and Kinect systems; to all the participants in the listening tests, especially Mohit Garg, Nicolas Gaedes and Giulia Falgari.

7. REFERENCES

[1] 3D-CoForm. *MeshLab*, 2012. http://meshlab.sourceforge.net/, accessed Oct. 2012.

[2] 3dMD. *3dMDface System*. UK, 2012. http://www.3dmd.com/3dMDface/, accessed Aug. 2012.

[3] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 2nd edition, 1996.

[4] M. Costabel. *Principles of Boundary Element Methods*. Finite Elements in Physics. Technische Hochschule Darmstadt, Germany, 1986.

[5] William G. Gardner and Keith D. Martin. HRTF measurements of a KEMAR. *J. Acoust. Soc. Amer.*, 97(6):3907–3908, 1995. http://sound.media.mit.edu/resources/KEMAR.html.

[6] H. Hu, L. Zhou, H. Ma, and Z. Wu. HRTF personalization based on artificial neural network in individual virtual auditory space. *J. Applied Acoustics*, 69(2008):163–172, 2007.

[7] P. Juhl and V.C. Henriquez. *OpenBEM: Open source Matlab codes for the Boundary Element Method*. Denmark, 2012. Accessed August 2012.

[8] Brian F. G. Katz. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *J. Acoust. Soc. Amer.*, 110(5):2440–2448, 2001.

[9] S. Kirkup. *The Boundary Element Method in Acoustics*. Integrated Sound Software. Todmorden, UK, 2nd edition, 2007.

[10] Microsoft. *Introducing Kinect for Xbox 360*, 2012. http://www.xbox.com/en-GB/KINECT, accessed Oct. 2012.

[11] John C. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *J. Acoust. Soc. Amer.*, 106(3):1493–1510, 1999.

[12] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Amer.*, 94(1):111–123, 1993.

[13] Frederic L. Wightman and Doris J. Kistler. Headphone simulation of free-field listening. ii: Psychophysical validation. *J. Acoust. Soc. Amer.*, 85(2):868–878, 1989.

[14] Wen Zhang, Mengqiu Zhang, R.A. Kennedy, and T.D. Abhayapala. On high-resolution head-related transfer function measurements: An efficient sampling scheme. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(2):575–584, 2012.