Audio Engineering Society

Convention e-Brief 433

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

HOBA-VR: HRTF On Demand for Binaural Audio in immersive virtual reality environments

Michele Geronazzo[1], Jari Kleimola[2], Erik Sikstroöm[3], Amalia de Götzen[1], Stefania Serafin[1], and Federico Avanzini[4]

[1]*Dept. of Architecture, Design, and Media Technology - Aalborg University, A. C. Meyers Vænge 15, Copenhagen, 2450, Denmark*
[2]*Hefio Ltd, Otakaari 5 A,Espoo, 02150, Finland*
[3]*Virsabi ApS, Artillerivej 86, Copenhagen, 2300, Denmark*
[4]*Dept. of Computer Science - University of Milano, Via Comelico 39/41, Milano, 20135, Italy*

Correspondence should be addressed to Michele Geronazzo (mge@create.aau.dk)

ABSTRACT

One of the main challenges of spatial audio rendering in headphones is the personalization of the so-called head-related transfer functions (HRTFs). HRTFs capture the listener's acoustic effects supporting immersive and realistic virtual reality (VR) contexts. This e-brief presents the HOBA-VR framework that provides a full-body VR experience with personalized HRTFs that were individually selected on demand based on anthropometric data (pinnae shapes). The proposed WAVH transfer format allows a flexible management of this customization process. A screening test aiming to evaluate user localization performance with selected HRTFs for a non-visible spatialized audio source is also provided. Accordingly, is might be possible to create a user profile that contains also individual non-acoustic factors such as localizability, satisfaction, and confidence.

1 Introduction

Accurate spatial rendering of sound sources for virtual environments has seen an increased interest lately with the rising popularity of virtual reality (VR) and augmented reality (AR) technologies. While the topic of headphone based 3D-audio technology itself has been widely explored in the past, here we propose a framework for personalized immersive VR experiences. The most common approach for spatial audio rendering in VR/AR contexts makes use of dummy head HRTFs for all listeners, avoiding personalization. However, it is well known that listening through dummy ears causes noticeable distortion in localization cues [1]. However, the increase of available HRTF data during the last decade supports the research process towards novel selection processes of non-individual HRTFs.

In our framework, we integrated a Matlab tool which maps anthropometric features into the HRTF domain, following a ray-tracing modeling of pinna acoustics. The main idea is to draw pinna contours on an image loaded into a tool. Distances from the ear canal entrance define reflections on pinna borders generating spectral notches in the HRTF. Accordingly, one can use such anthropometric distances and corresponding notch parameters to choose the best match among a pool of available HRTFs. A screening test in VR able to quickly evaluate (10 minutes) the effectiveness of

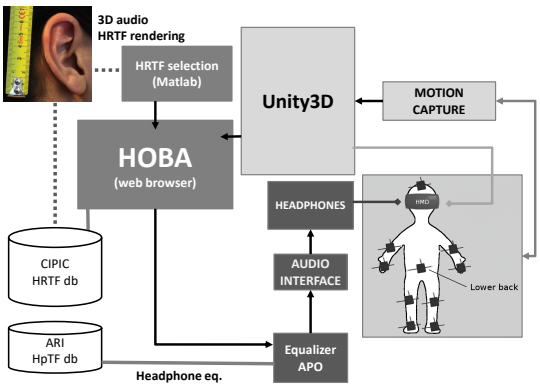

Fig. 1: HOBA-VR system overview.

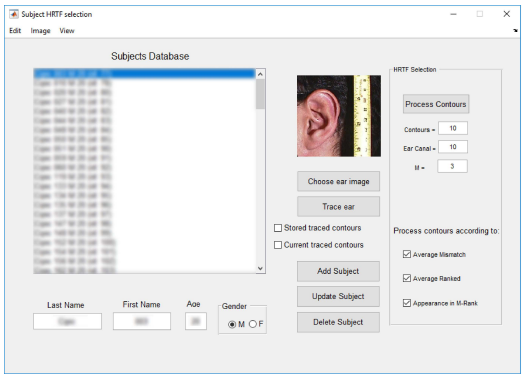

Fig. 2: Tool for HRTF selection with pinna anthropometry: main graphical user interface.

such selection was also developed.

2 The HOBA-VR framework

The runtime software environment is distributed into two loosely connected subsystems (see Fig. 1 for a schematic representation). The master subsystem contains the main logics, 3D object models, graphics rendering, and user position/pose tracking. This part was implemented in the Unity3D game engine. Spatial audio rendering was performed in the Firefox web browser. The subsystems are interconnected via a network socket, using the Open Sound Control (OSC) content format [2] as messaging payload. A simple Node.js hub was additionally required to bridge the UDP socket and WebSocket compatible endpoints together. The master subsystem initializes the remote soundscape with sound objects. It can thereafter dynamically alter the 3D positions of the remote sound objects using OSC. Listener position and pose are controlled in a similar manner. The audio subsystem relies on a robust HRTF selection algorithm and the HRTFs On-demand for Binaural Audio (HOBA) rendering framework for web browsers.

2.1 HRTF selection tool

We adopted the Matlab tool developed by Geronzzo *et. al* [3, 4] which is publicly available at the following link: `https://github.com/msmhrtf/sel`. This method implements the mapping of anthropometric features into the HRTF domain, following a ray-tracing modeling of pinna acoustics [5, 6]. The main idea is to draw pinna contours on an image loaded into the tool (see Figure 2 for software GUI). Distances from the ear canal entrance define reflections on pinna borders generating spectral notches in the HRTF. Accordingly, one can use such anthropometric distances and corresponding notch parameters to choose the best match among a pool of available HRTFs [7, 6]. In particular, it has been shown that the first and most prominent notch in the HRTF is typically associated to the most external pinna contour on the helix border (the "C_1" contour hereafter).

HRTF selection can be performed in the CIPIC database [8] that provided HRTFs of 43 subjects (plus two dummy-head KEMAR HRTF sets with small and large pinnae, respectively) at 25 different azimuths and 50 different elevations, to a total of 1250 directions. This approach can be easily extended to other publicly available HRTF sets.

Now, assume that N estimates of the C_1 contour and K estimates of the ear canal entrance have been traces on a $2D$ picture of the pinna of a subject (the meaning of N and K is explained later). One can define the basic notch distance metric in the form of a mismatch function between the corresponding notch frequencies, and the notch frequencies of a HRTF:

$$m_{(k,n)} = \frac{1}{N_\varphi} \sum_{\varphi} \frac{|f_0^{(k,n)}(\varphi) - F_0(\varphi)|}{F_0(\varphi)}, \qquad (1)$$

where $f_0^{(k,n)}(\varphi) = c/[2d_c^{(k,n)}(\varphi)]$ are the frequencies extracted from the image and contours of the subject, and F_0 are the notch frequencies extracted from the HRTF with an *ad-hoc* algorithm developed in [9]; (k,n) with

$(0 \leq k < K)$ and $(0 \leq n < N)$ refers to a one particular pair of traced C_1 contour and ear canal entrance; φ spans all the $[-45°, +45°]$ elevation angles for which the notch is present in the corresponding HRTF; N_φ is the number of elevation angles on which the summation is performed. Extracted notches need to be grouped into a single track evolving through elevation consistently, a labeling algorithm (e.g. [5]) performed such computation within adjacent HRTFs in elevation.

The default value has been set to $N = K = 10$ and C_1 contours and ear canal entrances were traced manually on the pinna image of each participant by the experimenter that followed the guidelines in [3]; then the HRTF sets in the CIPIC database were automatically ranked in order of similarity with the participant. The final best non-individual HRTF set was selected taking into account the following mismatch function:

- **Top-**3 **appearance:** for each HRTF, a similarity score is assigned according to the number of times (for all the (k,n) pairs) in which that HRTF ranks in the first 3 positions.

It has to be noted that the algorithm assigned a similarity rank list corresponds exactly to increasing values of the mismatch function calculated with Eq. (1) (for a single (k,n) pair). Then, at each HRTF is assigned a similarity score that is an integer corresponding to its ranked position taken from the previous mismatch list (for a single (k,n) pair).

2.2 HOBA framework

HOBA extends W3C Web Audio API with support for

1. remote soundscape;
2. spherical coordinate system;
3. custom HRTFs in spatial audio rendering.

Details of the initial HOBA release can be found in [10]; source code of our framework is freely available in the following repo: `https://github.com/hoba3d` under MIT license. A brief technical description follows below.

The remote soundscape extension implements sound objects and the listener singleton with *SpatialSource* and *AudioListener* classes. The former renders audio buffers through *SpatialPanner* instance, which implements the custom rendering algorithm as discussed below. Coordinate system extension enables conversions between cartesian and spherial presentations.

The extended HRTF support loads custom HRTF datasets from local or internet URLs in the proposed WAVH transfer format. WAVH extends the canonical RIFF/WAVE file format with a LIST chunk, which can contain any number of custom head-related impulse response (HRIR) chunks.[1] Each HRIR chunk encodes data pertaining to a single HRIR, consisting of an 'info' part (describing azimuth, elevation, distance and delay attributes) and of a 'data' part (containing binaural HRIR data). The RIFF structure then becomes

$$RIFF\{WAVE\{fmt\,list\{hrir\,hrir\,hrir\,...\}\}\}.$$

with

$$hrir ::= \{info\,data\}$$

The backend codebase is able to convert SOFA format files into the proposed WAVH format. HOBA framework decodes the WAVH representation into an array of HRIRs. The array is then organized into a 3D mesh, based on azimuth, elevation and distance of each HRIR. The mesh is finally Delaunay-triangulated, and centered on *AudioListener* coordinate space. The orientation of the mesh is synchronized to the pose of the listener. Listener's position and the positions of *SpatialSource* objects are bound to master subsystem's world coordinate space.

The real-time spatial audio rendering algorithm operates as follows. A vector drawn from *SpatialSource* position to the origin of *AudioListener* intersects one of the mesh triangles. The vertices of that triangle denote three HRIR components from the loaded HRTF dataset. *SpatialPanner* interpolates the three HRIRs into a single composite impulse response, which finally convolves the dry audio source signal. Interpolation is repeated for each change in spatial source position (and naturally, for each position or pose update of the listener). Position updates are cross-faded with short (50 ms) linear ramps to avoid audio artifacts.

2.2.1 Headphone equalization

At the moment, our framework supports Sennheiser HD600 headphones that were equalized using their

[1] HRTF is the Fourier transform of the HRIR.

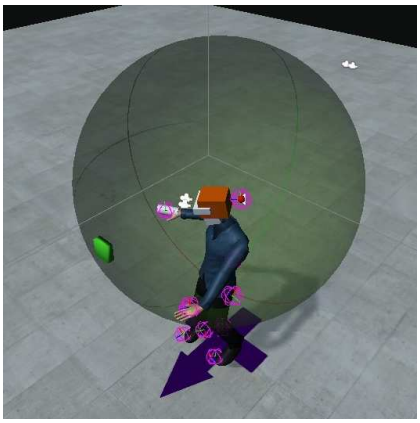

Fig. 3: Outside view of the localization screening test.

headphone impulse responses (HpIRs) measured over more than 100 human subjects from the Acoustic Research Institute of the Austrian Academy of Sciences; [2] data are available in SOFA format [11] helping the computation of compensation filters able to remove the average acoustic headphone contribution, and thus to reduce spectral coloration while listening with Sennheiser HD600 [12]. Nothing prevents future integration with other headphones for which HpIRs are available.

Equalization filters were loaded in Equalizer APO software [3] which is able to perform low-latency convolution between an arbitrary impulse response (i.e. the FIR equalization filters) and the streaming audio played back from HOBA framework.

3 Localization screening test

One of the relevant features of the proposed framework is the build-in screening test that aims at providing a user profile of the abilities in locating spatialized sounds. The design of this test followed this key requirement: keeping the execution quick and comfortable for participants (10 minutes maximum) in such a way to be used as a monitoring test replacing time- and resource- consuming psychoacoustic tests. Accordingly, it is necessary to evaluate HRTF-based spatialization prior to the VR experience, testing specific HRTF sets and user localization ability.

In this first version, we adopt a typical sound source localization task and the test is implemented in an immersive virtual reality environment consisted of a textured plane on which the subject is standing and the inside of a semi-transparent sphere with a 1 m radius. The sphere is also equipped with lines indicating the horizontal, median and traversal planes (see Fig. 3 for screenshot of the VR environment).

The auditory stimulus is played through HOBA framework. The stimuli - a train of noise bursts, presented at 60 dBA level [7] when measured from the earphone cup; directional filtering through HRTFs render all the combinations of the following angles (spherical coordinate system):

- azimuths: -180° (behind), -120°, -60°, 0° (straight ahead), 60°, 120°;
- elevation: -28.125°, 0° (at the horizon), 28.125°, 56.250°, 90° (above);

These values lead to a total of 6 (azimuths) $\times$ 4 (elevations) $+$ 1 (elevation 90°) spatial locations; at the start of each session, user head should be located at the origin of the coordinate system. The presentation order of these locations is always randomized, and test locations presented once.

At the start of each condition, the center of the visual sphere and the locations of the sound sources should be set approximately to the height and position of the subject's head, while user should look straight forward. Rather than tracking the exact ear position for each participant, a generic ear position was measured from placing the head mounted display on a Bruel and Kjaer 4128 head and torso simulator (HATS). [4] By measuring distances from the three-point trackable on the display to the ear canal of the HATS, an approximate and generic position for the ears was acquired allowing a coherent rendering in the virtual world.

A game controller with a virtual representation of a laser pointer was implemented using motion capture data, a USB mouse, and a ray-casting method combined with a narrow angle red spotlight attached to the avatar's right hand. The controller should hold in the dominant hand allowing the subjects to point at the location they perceived the sound was coming from. By pressing the left button, an ad-hoc software logged the location of the pointer into a text file. The logging of the perceived position is also accompanied with auditory feedback (a "click" sound and the silencing

[2] http://sofacoustics.org/data/headphones/ari
[3] https://sourceforge.net/projects/equalizerapo/
[4] https://www.bksv.com/en/products/transducers/ear-simulators/head-and-torso

of the noise bursts). Pressing the right mouse button initialized the next trail.

4 Conclusions

In this paper, we briefly described the HOBA-VR framework: a proof-of-concept for the future of 3D-media environments and broadcasting. Immersive experiences require high spatial fidelity in an individual sound field reproduction (personalized HRTFs) and data format (WAVH format). Future research on the auditory side of user characterization and its influence on audio experience in VR could benefit from our proposed open-source framework; experimental validation with massive participation of human subjects will be highly relevant for the applicability of HRTF selection procedures in different VR scenarios. It is worthwhile to notice that our software implementation which is based on HOBA (Web Audio API) and Unity, is technologically-ready for a widespread application in mobile VR devices.

Acknowledgments

This study was supported by the 2016-2021 strategic program "Knowledge for the World" awarded by Aalborg University to MG.

References

[1] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L., "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, 94(1), pp. 111–123, 1993, doi: 10.1121/1.407089, 00940.

[2] Wright, M., Freed, A., et al., "Open SoundControl: A New Protocol for Communicating with Sound Synthesizers." in *ICMC*, 1997.

[3] Geronazzo, M., Peruch, E., Prandoni, F., and Avanzini, F., "Improving elevation perception with a tool for image-guided head-related transfer function selection," in *Proc. of the 20th Int. Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK, 2017.

[4] Geronazzo, M., Peruch, E., Prandoni, F., and Avanzini, F., "Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization," *J Audio Eng Soc*, 2018.

[5] Spagnol, S., Geronazzo, M., and Avanzini, F., "On the Relation between Pinna Reflection Patterns and Head-Related Transfer Function Features," *IEEE Trans. Audio, Speech, Lang. Process.*, 21(3), pp. 508–519, 2013.

[6] Geronazzo, M., Spagnol, S., and Avanzini, F., "Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric," *IEEE/ACM Trans. Audio, Speech, Lang. Process. - accepted for publication*, 2018.

[7] Geronazzo, M., Spagnol, S., Bedin, A., and Avanzini, F., "Enhancing Vertical Localization with Image-guided Selection of Non-individual Head-Related Transfer Functions," in *IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP 2014)*, pp. 4496–4500, Florence, Italy, 2014.

[8] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C., "The CIPIC HRTF Database," in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, pp. 1–4, New Paltz, New York, USA, 2001.

[9] Geronazzo, M., Spagnol, S., and Avanzini, F., "Estimation and Modeling of Pinna-Related Transfer Functions," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, pp. 431–438, Graz, Austria, 2010.

[10] Geronazzo, M., Kleimola, J., and Majdak, P., "Personalization Support for Binaural Headphone Reproduction in Web Browsers," in *Proc. 1st Web Audio Conference*, Paris, France, 2015.

[11] Boren, B. B., Geronazzo, M., Majdak, P., and Choueiri, E., "PHOnA: A Public Dataset of Measured Headphone Transfer Functions," in *Proc. 137th Conv. Audio Eng. Society*, Audio Engineering Society, 2014.

[12] Boren, B., Geronazzo, M., Brinkmann, F., and Choueiri, E., "Coloration Metrics for Headphone Equalization," in *Proc. of the 21st Int. Conf. on Auditory Display (ICAD 2015)*, pp. 29–34, Graz, Austria, 2015.