The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios

Michele Geronazzo*
Dept. of Architecture, Design
and Media Technology
Aalborg University

Erik Sikström†
Virsabi ApS

Jari Kleimola ‡
Hefio Ltd

Federico Avanzini§
Dept. of Computer Science
University of Milano

Amalia de Götzen, Stefania Serafin¶
Dept. of Architecture, Design
and Media Technology
Aalborg University

Abstract

Achieving a full 3D auditory experience with head-related transfer functions (HRTFs) is still one of the main challenges of spatial audio rendering. HRTFs capture the listener's acoustic effects and personal perception, allowing immersion in virtual reality (VR) applications. This paper aims to investigate the connection between listener sensitivity in vertical localization cues and experienced presence, spatial audio quality, and attention. Two VR experiments with head-mounted display (HMD) and animated visual avatar are proposed: (i) a screening test aiming to evaluate the participants' localization performance with HRTFs for a non-visible spatialized audio source, and (ii) a 2 minute free exploration of a VR scene with five audiovisual sources in a both non-spatialized (2D stereo panning) and spatialized (free-field HRTF rendering) listening conditions. The screening test allows a distinction between good and bad localizers. The second one shows that no biases are introduced in the quality of the experience (QoE) due to different audio rendering methods; more interestingly, good localizers perceive a lower audio latency and they are less involved in the visual aspects.

Index Terms: Human-centered computing—Interaction paradigms—Virtual reality; Human-centered computing—Interaction devices—Sound-based input / output Human-centered computing—Interaction techniques—Auditory feedback

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 scientific literature (see [5] for a reference book in the field, and [32] for a recent review), here we discuss the connection between ability in auditory localization and its relevance in immersive VR experience.

In the field of spatial audio, localization of sound sources is typically the main focus when investigating and comparing technologies based on head-related transfer functions (HRTFs), whereas less attention is devoted to its connection with spatial audio quality, sensation of presence and attention [35]. In particular, psychoacoustic tests demonstrated that personalized HRTFs result in improved localization ability [27, 33] and could be characterized according to vocabularies, such as the Spatial Audio Quality Inventory (SAQI) [23].

It is worthwhile to note that listening with HRTFs (both individual and even more non-individual) exhibits high variability in localization performance, in relation to both differences in acoustic factors due to listener anthropometry [15], and perceptual factors, i.e., the individual ability of encoding directional information [3, 24]. Accordingly, it is very important to have a user characterization in terms of auditory abilities and HRTF usability, which should provide perceptually-relevant guidelines for a personalized design of full 3D spaces within an immersive and multimodal VR context.

Moreover, recent literature on auditory models and spatial hearing [28, 39] suggests a connection between spectral matching abilities and elevation perception due to the tuning processes in low-level auditory cortex. Accordingly, performances in vertical localization could be considered an indirect measure of sensitivity to dynamic spectral changes and auditory plasticity. For such purposes, we used a fast screening test able to investigate localization abilities of each user replacing time- and resource- consuming psychoacoustic tests.

We provided users with personalized HRTFs that were individually selected based on anthropometric data of the external ear (also known as pinna) [12, 16] in order to provide reliable elevation cues for a subsequent free VR exploration. Based on screening results, we defined a criterion which allowed to cluster users into good and bad localizers; we assumed that our personalization method provided reliable spectral cues in the acoustic domain for all users, thus confining the cause of poor/good localization performances in the non-acoustic domain. This hypothesis is strengthened by recent findings which identified a dominance of perceptual factors on acoustical factors for sound localization of virtual sound sources [3]. Accordingly, we focused our analysis on the impact that a perceptual characterization of user has the perceived quality of experience (QoE) for immersive and multimodal scenarios.

From an applicative point of view, it is relevant for our research to address design recommendations for sound in VR environments, allowing interpretation and improvements of the effectiveness in the provided auditory information. In particular, we proposed a initial scenario consisting of an outdoor park scene that participant freely explored with head-mounted display (HMD) and animated visual avatar for 2 minutes. Three audio rendering conditions were considered: a simple 2D stereo panning, a typical 3D HRTF rendering with dummy head acoustics, and a 3D rendering with a state-of-the-art HRTF personalization procedure based on anthropometric data of the pinna [12]. Collected data were analyzed in order to investigate if there were statistically significant differences after experiencing the environment with different audio conditions. A null result in this comparison means that the effects due to user's perceptual characterization could be predominant with respect to acoustic information provided by the audio rendering algorithms.

The remainder of this paper is structured as follows. Sec. 2

*e-mail: mge@create.aau.dk
†e-mail: erik.sikstrom@outlook.com
‡e-mail: jari.kleimola@hefio.com
§e-mail: federico.avanzini@di.unimi.it
¶e-mail: {ago,sts}@create.aau.dk

978-1-5386-7459-8/18/$31.00 ©2018 IEEE
DOI 10.1109/ISMAR.2018.00034

describes previous research concerning spatial audio in VR and HRTF selection procedures. Section 3 reports on the technical implementation of the audiovisual virtual reality applications used in the study. Section 4, describes in detail the two experiments, namely the screening procedure used for investigating the localization ability of the participants, and the subsequent test in which participants could freely explore a virtual environment. The results of both experiments are presented in Sec. 5 and discussed in Sec. 6, while a summary of the paper is given in in Sec.7.

2 Related works: spatial audio in VR

Previous research has shown that spatial sound has a positive influence on performance in wayfinding tasks [17], and in localization performance in an audio-haptic task [37]. Furthermore, Zhang *et al.* [42] used audio feedback with HRTF based spatialization in an assembly task, and found that providing a combination of visual and auditory cues had a positive effect on efficiency and usability. Concerning the sensation of presence, Hendrix and Barfield [19] observed that the inclusion of spatial audio yielded higher presence ratings after subjects had explored their virtual environment. However, their study did not find any evidence that the spatial audio condition had an influence on the perceived realism of the virtual environment.

Bormann [8] investigated the role of spatial audio in relation to presence when the audio feedback was either task relevant or not. In this study, the virtual environment was presented on a desktop computer. The participants were asked to search an environment for either an object that was also an audio source (a radio playing music), or for another object that was not an audio source. To this, there were additional audio conditions where the audio were either spatialized (using the audio features of the DIVE engine[1]), or spatialized but with the absence of distance attenuation. The findings of the study showed among other things that spatial audio generally had a positive influence on presence scores. However, it was those that used the audio condition without distance attenuation who had the largest increase in presence score compared to the baseline. Also, those participants searching for an object that was also emitting sounds felt less involved with the visual aspects, and more involved with the auditory aspects of the environment compared to those who searched for a non-sounding object.

More recently, Hedke *et al.* [18] investigated the impact of binaural and stereo rendering with headphones and loudspeakers on the QoE of a typical computer game. Their results suggest that differentiations in spatialization techniques are highly task-dependent. In particular, a 2 minute free exploration did not result in a noticeable difference of QoE in the digital game.

2.1 HRTF listening

The measurement of individual HRTFs usually requires special measuring apparatus and a time-consuming procedure, leading to impractical solutions for real-world applications. Alternative methods for HRTF personalization are usually preferred, which look for a delicate trade-off between audio quality and handiness of the personalization procedure [14].

The most common approach for spatial audio rendering in VR/AR contexts makes use of dummy-head HRTFs, or anyway a single set of HRTFs for all listeners, without personalization. Such predefined sets can be taken from databases such as LISTEN [10] and CIPIC [2]. However, it is known that listening through dummy ears causes noticeable distortion in localization cues [40]. During the last decade, the increase of publicly available HRTF data has boosted research on novel approaches to the selection and modeling of non-individual HRTFs. [2]

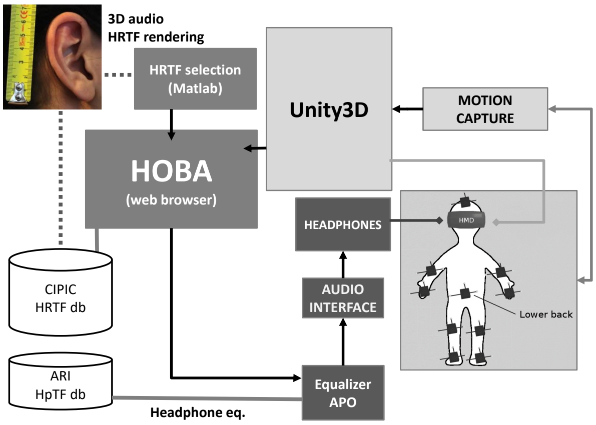


Figure 1: System overview.

Typically, HRTF selection problems can be characterized in terms of three issues:

- **metric domains**: acoustics, anthropometry, and psychoacoustics;
- **spatial ranges**: a subspace around the listener for whom the personalization process results in significant improvements for localization performances, e.g., horizontal or vertical plane only;
- **methods**: computational steps which allow to infer the most appropriate non-individual HRTF set for a listener; pre-processing actions such as *data unification*, *feature extraction* (e.g. the frequency scale factor of Middlebrooks [26]), *dataset reduction* [36], and *dimensionality reduction* [20] can be performed prior the HRTF selection.

For the desired domains and spatial ranges, several approaches can be applied based on anthropometric database matching [43], linear regression models between acoustical and anthropometric features [27], subjective selection [31], or minimization of HRTF differences in the acoustic domain [27]. Once one or a set of best HRTF candidates are identified, listener can also self-tune each HRTF set through spectral manipulations and enhancements [27], and weight adjustments [33]. Finally, in a phase of adaptation to non-individual HRTFs users can obtain multimodal feedback in localization/discrimination tasks and improve their performance [25].

3 Materials and methods

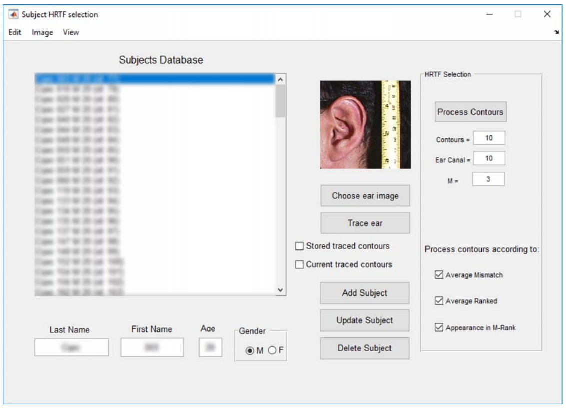
Tests were conducted in an immersive virtual reality environment where participants wore an head mounted display (HMD), headphones, and were equipped with motion tracking markers that in turn provided the information to animate a visual avatar according to the subject's movements.

3.1 Apparatus for immersive virtual reality

The computer equipment and software used for the two studies are presented in Fig. 1. The graphics rendering, audio and motion tracking software was running on a Windows 7 PC computer (Intel i7-4470K 3.5GHz CPU, 16 GB RAM and a MSI Gaming X GeForce

[1] DIVE, Distributed Virtual Environment, by the Swedish Institute Of Computer Science (SICS), version 3.3 (1999)

[2] See, for instance, the official website of the Spatially Oriented Format for Acoustics (SOFA) project, http://sofaconventions.org



Contours = 10
Ear Canal = 10
M = 3

□ Stored traced contours
□ Current traced contours

Process contours according to:
☑ Average Mismatch
☑ Average Ranked
☑ Appearance in M-Rank

● M ○ F

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

GTX 1070 graphics card). The HMD used was a nVisor SX with a FOV of 60 degrees with a screen resolution 1280x1024 pixels in each eye. The audio feedback was delivered through a RME Fireface 800 interface with a pair of Sennheiser HD600 headphones. The motion-tracking was done with a Naturalpoint Optitrack motion-tracking system with 12 cameras of the model V100:R2 and with 10 three-point trackables attached on the subjects. In particular, the head-tracking had a latency of 10 ms which did not compromise an effective rendering for auralization [29]. The virtual environments was developed with Unity3D v4.6[3].

3.2 Spatial audio rendering

3.2.1 HRTF selection tool

We adopted the Matlab tool developed by Geronazzo *et. al* [12] that implements the method of mapping anthropometric features into the HRTF domain, following a ray-tracing modeling of pinna acoustics [15].[4] The main idea is to draw pinna contours on an image loaded into the tool (see Fig. 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 [16]. 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 was performed in the CIPIC database [2] that provided HRTFs of 43 human subjects at 25 different azimuths and 50 different elevations, to a total of 1250 directions.

Assume that N estimates of the C_1 contour and K estimates of the ear canal entrance have been traced 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 pinna contour of the subject, and F_0 are the notch frequencies of the HRTF, estimated with an *ad-hoc* algorithm [13]. The pair (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 $[-45^\circ, +45^\circ]$ 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.

In this study, we set $N = K = 10$, and C_1 contours and ear canal entrances were traced manually on the left-side pinna image of each participant by the experimenter, following the guidelines in [12]. Then the HRTF sets in the CIPIC database were automatically ranked in order of similarity with the participant, according to the mismatch function of Eq. (1). The final best non-individual HRTF set was selected using the "top-3" metric defined in [12]: in short this metric counts the number of times (for all the $N \times K$ pairs) in which an HRTF appears in the first 3 positions of the ordered mismatch list, and was proven to be the robust against measurement errors.

3.2.2 HOBA framework

The runtime software environment is distributed into two loosely connected subsystems. 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. The audio subsystem relies on the HRTFs On-demand for Binaural Audio (HOBA) rendering framework for web browsers; in this work, spatial audio rendering was performed in the web browser. HOBA extends W3C Web Audio API with support for i) remote soundscape, ii) spherical coordinate system, and most importantly, iii) custom HRTFs in spatial audio rendering.

The subsystems are interconnected via a network socket, using the Open Sound Control (OSC) content format [41] 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. An overview of the technical description of the HOBA framework has been published recently [11] and the git repository is available at the following link: `https://github.com/hoba3d`.

3.2.3 Headphone equalization

Sennheiser HD600 headphones were equalized using their headphone impulse responses (HpIRs) measured over more than 100 human subjects from the Acoustic Research Institute of the Austrian Academy of Sciences;[5] data are available in SOFA format [7] and were used to compute compensation filters to remove the average acoustic headphone contribution, and thus to reduce spectral coloration. Equalization filters were loaded in Equalizer APO software [6] 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.

4 Experiments

4.1 Screening test

The aim of this experiment was to conduct a screening of the subject pool's abilities of accurately locating spatialized sounds in both azimuth and elevation; auditory stimuli were presented using individually selected HRTFs, following the selection method described in Sec. 3.2.1, which is able to provide reliable vertical localization cues [15, 16].

The experimental design focused on a short execution-time and a comfortable VR experience (10 minutes maximum) in such a

[3] https://unity3d.com/

[4] https://github.com/msmhrtf/sel

[5] http://sofacoustics.org/data/headphones/ari

[6] https://sourceforge.net/projects/equalizerapo/

Type	*Behavior*	*Visibility*	*Level*
Old transistor radio	Static - positioned at a table while playing a static radio noise	Medium	45.5 dBA
Fireplace	Static - placed at ground level, playing a looped fire recording	Very clear, with animated fire and flickering light sources	35.3 dBA
Bird	Static - placed in a tree at approximately head height, playing a loop of birdsong with twittering heard at regular intervals	Hard to spot, due to poor lighting conditions	50 dBA
Street lamp (malfunctioning)	Irregular - placed high up on a pole, with a lamp that is humming and flickering. Every time the lamp goes off, the hum pauses. When the lamp is lit again, a faint "clink" is heard	Clear. The subjects need to look up to see it, but it is not discreet as it has a flickering light source as well as the sound source	36.6 dBA
Grasshopper	Static - positioned at ground level in a tuft of grass at the side of the path	Hard to spot, as it hasn't got a visual representation itself, but is hidden in high grass	32.7 dBA

Table 1: Audiovisual sound sources used in the free exploration task. Loudness were measured using an AZ9822 digital sound level meter located at the center of the right headphone cup while stationed inside an anechoic chamber. Each sound stimulus was virtually placed at 1 m distance to the right of the avatar ear position in the VR environment for all audio-rendering conditions.

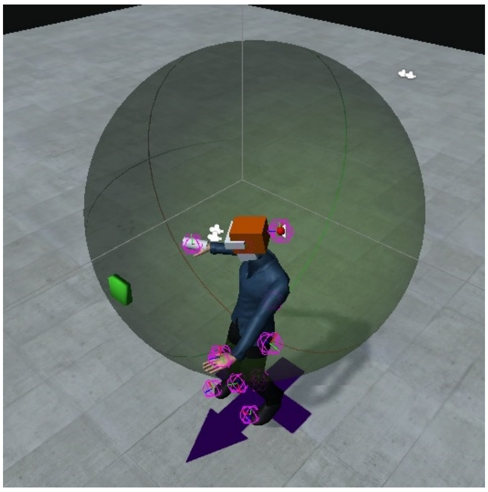


Figure 3: Outside view of the screening test.

way to be used as screening test with the same efficacy of a longer psycho-acoustic evaluation.[7]

A sound source localization task was implemented in an immersive VR environment consisting of a textured plane, on which the subject is standing, and the interior of a semi-transparent sphere with a 1m radius. The sphere was also equipped with lines indicating the horizontal, median and traversal planes. An illustration of the virtual environment and the avatar is presented in Fig. 3.

The basic auditory stimulus was a train of noise bursts, presented at 60 dBA level [16] when measured from the earphone cup. Directional filtering through HRTFs rendered all the combinations of the following angles (spherical coordinate system):

- azimuths, θ: -180° (behind), -120°, -60°, 0° (straight ahead), 60°, 120°;
- elevation, ϕ: -28.125°, 0° (at the horizon), 28.125°, 56.250°, 90° (above); these coordinates follow the CIPIC HRTF database resolution of 5.625° in order to avoid localization biases that may arise due interpolations in non-available spatial locations;

These values led to a total of 6 (azimuths) $\times$ 4 (elevations) $+$ 1 (elevation 90°) spatial locations; we identified such positions in order to achieve a limited loss of accuracy in the characterization. We followed two distinct assumptions for θ and ϕ selections. Azimuth errors have a slight increase in variability towards lateral angles [6], thus leading to consider directly front/back positions and towards directly left/right. Position dependent elevation errors could be connected to pinna resonance modes [4] occurring at the horizon or in the frontal region (within $\pm 45°$), and in a wide range above the head ($\phi > 45°$).

At the start of each session, subject head was located at the origin of the coordinate system. The distance of the sound sources was set to 1 m, which corresponds to the dimensions of the sphere in the visual environment. This sphere radius was chosen to correspond to the sound source distance in the CIPIC HRTF measurements: with this choice, no rendering compensations due to distance mismatch were needed. The presentation order of these locations was randomized; test locations were presented once per audio-rendering condition.

At the start of each condition, the center of the visual sphere and the locations of the sound sources were set approximately to the height and position of the subject's head, while she/he was told to look straight forward. In order to have a coherent ear position between participants and avatar's height, a generic ear position was measured from placing the head mounted display on a Brüel & Kjær 4128 head and torso simulator (HATS). [8] 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 with the head-tracker. Once the trial was started, the subjects were allowed to move their head and look for the sound source (note that no visual information of its location was given). Participants were not allowed to move their feet during each session, but were allowed to turn their upper body around as much as they wanted.

Before starting each trial they were instructed to look straight ahead, i.e. towards the direction azimuth $\theta = \phi = 0°$. A green cube in such position was used as visual reference point helping participants to find this starting point. A game controller with a virtual representation of a laser pointer was implemented using motion

[7] A formal validation of this short screening test is included in a manuscript which is currently in preparation.

[8] https://www.bksv.com/en/products/transducers/ear-simulators/head-and-torso

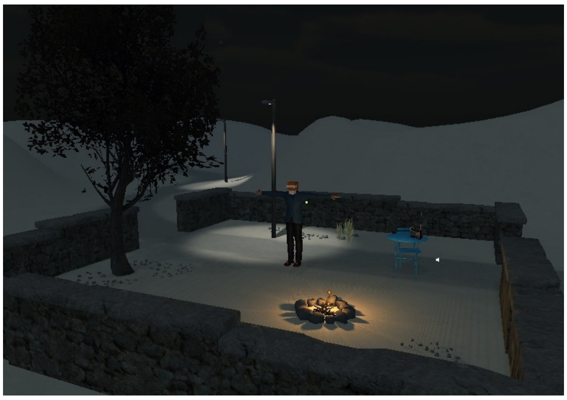


Figure 4: Outside view of the virtual environment used in the free exploration task.

ID	θ	ϕ	*Slope*	*p*
1	26.71, ±26.15	18.72, ±14.86	0.77	***
2	30.77, ±44.12	33.75, ±16.17	0.64	***
3	26.29, ±35.42	33.1, ±28.22	0.59	**
4	10.58, ±19.08	28.22, ±19.87	0.28	.07
5	12.86, ±16.69	32.97, ±18.67	0.54	**
6	35.98, ±45.32	25.71, ±23.76	0.32	*
7	20.68, ±38.88	27.76, ±19.2	0.07	.23
8	3.67, ±2.97	15.58, ±12.47	0.63	***
9	22.85, ±37.43	32.4, ±22.5	0.005	.5
10	6.18, ±4.57	31.15, ±18.75	0.2	.13
11	6.26, ±7.61	17.6, ±14.77	0.73	***
12	33, ±34.21	30.27, ±17.9	0.01	.77

Table 2: The mean-values, standard deviations for azimuth and elevation errors in degrees, slope-values, and p-value on the linear regression obtained during the screening test. Good localizers are marked with $**$ and $***$ (see discussion on Sec. 4.1).

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 was held in the right hand allowing the subjects to point at the location where they perceived the sound was coming from. By pressing the left button, the software logged the location of the pointer into a text file. The logging of the perceived position was also accompanied with auditory feedback (a "click" sound and the silencing of the noise bursts). Pressing the right mouse button initialized the next trail.

4.2 Free exploration in VR

This experiment aimed at evaluating the effect of auditory localization abilities during free exploration of a multimodal and complex virtual scene.

A night scene with a partially lit path in an area of sand dunes was designed to accommodate this experiment (see Fig. 4 for an illustration of the virtual park environment). Motivations behind such a choice were: i) a plausible setting for an acoustic environment without any background sounds (at night), and ii) free-field listening condition, no room reverberation among the sand dunes. Additionally, it was arbitrarily chosen to include five audiovisual sound sources with distinct features to provide variation between stimuli. These audiovisual sources are described in Table 1.

The area in which the sound sources were placed was surrounded by a stone wall to visually remind the participants not to attempt to walk away from the scene. Additionally, invisible colliders were also added to wall objects. Since the present study did not take into account room acoustics, we did not render audible reflections from the walls. In order to avoid HRTF dependencies and biases based on acoustic factors and their usage by participants, three different audio-rendering conditions were tested:

- *Stereo:* 2D audio condition using Unity3D's built-in audio engine; head orientation guided stereo panning to synthesize sound sources in lateral directions;
- *Generic HRTF:* 3D audio with HOBA loading a dummy-head generic HRTF set;
- *Custom HRTF:* 3D audio with HOBA loading an individually selected HRTF set with the tool described in Sec. 3.2.1.

The order of the conditions was randomized and placement of the audiovisual sources in the environment were randomly switched between three pre-defined configurations, where the placement of each audiovisual source were moved around within the walled area. However, the locations were chosen to be plausible, such that the street lamp was for example always placed somewhere by the path leading through the walled area. The subjects were allowed to freely explore the scene for approximately two minutes. The interactive locomotion and navigation features was implemented using a walking-in-place locomotion technique, using an algorithm described in [34]. The choice of a walking-in-place was to provide an ecological navigation solution, and real walking was not possible as the area of the scene were larger than what the motion capture system could track.

The experiment involved three trials, one for each audio condition, in randomized order. After each trial, a break was issued and the subjects were asked to fill in a questionnaire with questions regarding the level of presence experienced, spatial audio quality (adapted from SAQI [23]) and attention [8]. The questionnaire items were the following:

- Q1: Externalization - Was the sound source perceived inside or outside the head? (More internalized - More externalized)
- Q2: Responsiveness - To what extent did you experience that there were delayed reactions in the sound reproduction system? (Lower delay - Higher delay)
- Q3: Naturalness - How natural (close to real life) did you find the sound reproduction? (Lower naturalness - Higher naturalness)
- Q4: Presence - To what degree did you experience a sense of "being in the space"? (Lower - Higher)
- Q5: Attention audio - How much did the auditory aspects of the environment involve you? (Very little - Very much)
- Q6: Attention visual - How much did the visual aspects of the environment involve you? (Very little - Very much)
- Q7: How realistic did the virtual world seem to you? (Less realistic - More realistic)
- Q8: Did you perceive elevation? (Yes - No)

Questionnaire items Q1 to Q7 were presented along with seven-point rating scales.

5 Results

Twelve subjects participated voluntarily in the study. Seven of the subjects were female and five were male (age M = 32.75, SD = 5.56) and the experiment had a duration of approximately one hour in total. The participants all reported normal hearing, and all of them were right handed.

Item	*Global*	*good*	*bad*
Q1 Externalization	5.5, ±1.25	5.39, ±1.5	5.61, ±0.98
Q2 Responsiveness	2.42, ±1.48	2 ±1.33	2.83, ±1.54
Q3 Naturalness	5.33, ±1.01	5.33, ±1.08	5.33, ±0.97
Q4 Presence	5.22, ±1.12	4.94, ±1.3	5.5, ±0.86
Q5 Att. audio	5.31, ±1.69	4.89, ±2	5.72, ±1.23
Q6 Att. visuals	4.17, ±1.48	3.6, ±1.69	4.78, ±0.94
Q7 Realistic	4.5, ±1.38	4.28, ±1.36	4.72, ±1.41

Table 3: The mean and standard deviation values of the responses to the questionnaire items from the second experiment (seven-point ratings), for all audio conditions *(Global)* and the better and worse localizers subsets.

5.1 Screening test

Data acquired from the screening test included error angles in both elevation and azimuth calculated from the actual position of the sound source and its perceived position, i.e. the logged coordinates from the virtual laser pointer.

It is known from the literature that performances in vertical localization vary remarkably among individuals more than horizontal localization [24]. Based on this evidence, a linear regression analysis was performed on the elevation errors only, in order to group participants in terms of vertical localization abilities with the personalized HRTFs. We thus defined two categories: good and bad vertical localizers. This discrimination was based on the statistical significance of the linear regression within an α-level equals to 1% and 0.1%; the motivation behind this choice was to limit Type I errors, i.e. *false positive* finding of good localizers, introduced by our short localization test. Moreover, we provided a psycho-physical interpretation: slope threshold value of .35 corresponds to an angular error 15° which is comparable to the average localization blur in the median plane estimated in previous literature [9].[9]

This criterion led us to identify these two groups:

- subjects with the highest slope-values (all with a slope value $>$ .35) were considered good elevation localizers; six subject IDs: 1, 2, 3, 5, 8, and 11.
- subjects with low slope-values (all with a slope value $\leq$ .35) were considered bad elevation localizers; six subjects IDs: 4, 6, 7, 9, 10, and 12.

A summary of the screening test data is presented in table 2.

5.2 Free exploration in VR

For the multimodal experiment, which involved free exploration of a park environment with five audio-visual objects, the three audio rendering conditions (*Generic, Custom, Stereo*) were evaluated using the questionnaire described in Sec.4.2.

Due to non-normality of data distribution, non-parametric tests were performed: Friedman's test and repeated Wilcoxon signed-rank tests with Bonferroni correction. Tests were also performed using two groupings derived from the screening test. No statistically significant differences were found between the audio rendering conditions on the questionnaire items, in any of the above mentioned approaches. The mean and standard deviations from all questionnaire items, of all the audio conditions combined, are presented in Table 3 (*Global* column). Four out of twelve subjects reported that there were no elevation cues heard while exploring the VR scene with the *Stereo* condition, while one subject reported that there were no elevation cues when doing the same with the *Custom* condition. This subject, subject 7, was one of those who had been included in the subset of bad localizers due to low slope-values from the screening tests.

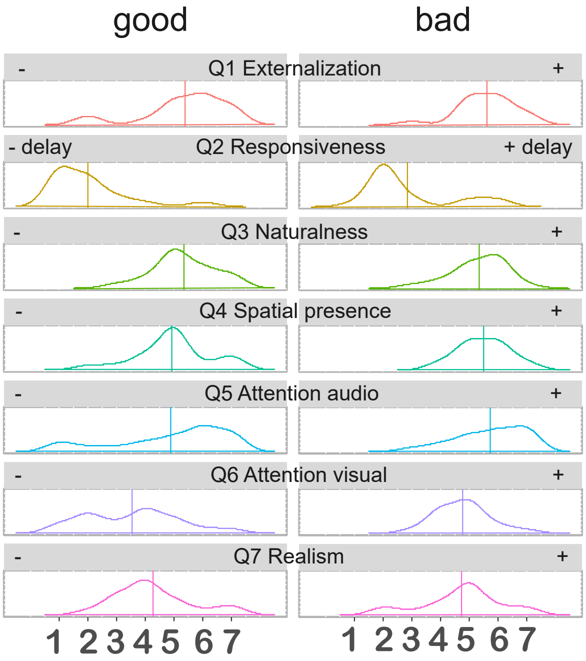


Figure 5: Response distributions for Likert questions Q1-7, grouped by localization ability, i.e. good vs. bad.

Additionally, a statistical analysis using the same tools were conducted with the two HRTF conditions combined versus the stereo condition to investigate if there was at all an effect of the 3D audio rendering. However, even here there were no significant differences found among the questionnaire responses.

Figure 5 depicts a graphical representation of data densities using the `likert` R-package[10], grouped by localization ability. The distributions of responses suggested a general trend in comparing the two groups: good localizers were less involved in the VR experience and gave lower scores to naturalness, presence, and realism, in comparison to bad localizers. The statistical analysis on these two groups were computed using Friedman's test and repeated Wilcoxon rank sum tests with Bonferroni correction. Since audio conditions did not show statistical differences, we could not consider such a distinction in our analysis. The results showed that the good localizers had an overall lower rating on *Q2: Responsiveness*. The good localizers (M = 2, Mdn = 2, SD = 1.33) rated the sound reproduction system as being faster (W = 224, p-value = 0.038) than the bad localizers (M = 2.83, Mdn = 2, SD = 1.54, see 6.a).

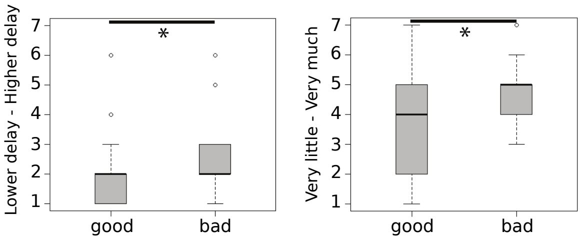
Additionally, for the responses to *Q6: Attention visual - How much did the visual aspects of the environment involve you?*, the statistical analysis (W = 236.5, p-value = 0.016) showed that the good localizers (M = 3.56, Mdn = 4, SD = 1.69) responded that they were less involved with the visual aspects of the scene than the bad localizers (M = 4.78, Mdn = 5, SD = 0.94, see 6.b).

6 General Discussion

Our test allowed to investigate short-term attention and QoE in VR. In the proposed questionnaire, we verified that there were no sta-

[9] The localization blur identifies the average margin of angular error in the human auditory system, and it can be expressed in terms of minimum audible angle (MAA).

[10] `https://cran.r-project.org/package=likert`, likert.density.plot with default values.



(a) Responsiveness

(b) Att. to visual

Figure 6: (a) Responses to Q2 for the good and bad localizers [1 = Lower delay, 7 = Higher delay]. (b) Responses to Q6 for the good and bad localizers [1 = Very little, 7 = Very much]

tistically significant differences among audio conditions for any of the 8 items, thus stressing the visual dominance over different audio rendering techniques for a short listening experience. Possible reasons explaining why the participants did not notice any difference between the audio conditions could likely be related to visual dominance in spatial localization (within the visual field of view) [38], and to the divided attention on interactive tasks and audio quality evaluations. Previous research on the influence of interactive tasks involved subjects either actively playing a computer game, or passively watching it, while the audio tracks were exposed to degradations (using low-pass filters, drop-outs in multichannel systems, audiovisual asynchrony) [21].

One of the main outcomes of this study is that users in the active conditions were more tolerant to degradations. Some of these results go partly against previous work, for example Barfield and Hendrix's study on spatialized audio [19], who did observe higher presence ratings in their spatial audio condition. The same can also be said when comparing the present results with those of Bormann [8]. However, there are differences between their experiments and those of the present study: there were less interactivity, less immersion and fewer audiovisual sources without any animations. In this sense, our approach has an important ecological validity.

On the other hand, when investigating the two different subsets of participants based on their performance in the screening test, an analysis of the questionnaire data with all the audio conditions combined showed that there were statistically significant differences on two items: *Q2: Responsiveness* and *Q6: Attention visual*. In particular, spatial localization abilities and spectral sensitivities could be related to individual perception of the *end-to-end spatial audio system latency* (SASL) [30], which accounts for *Q2: Responsiveness*. Moreover, good localizers were also to some extent paying less attention to visual aspects rather than to auditory aspects of the environment; however, responses to the questionnaire item asking that very question *Q5: Attention audio*, did not show any significant differences.

When comparing these results with previous research, it can be noted that Bormann [8] also observed a lower level of involvement with the visual aspects in their experiment, but this result was only observed for the condition where subjects were actively searching for the audiovisual source. Hence one could argue that localizing audiovisual sources with acoustic cues was more task relevant (as the subjects in that study were asked to actively locate that object within a maze) than our proposed free exploration.

It is worthwhile to note that our results provided an example of the importance in adopting pre-experimental screenings for VR experiences using spatial audio rendering, and particularly HRTF-based. Further research is still needed in order to characterize listeners perceptually, and related influences on the multi-modal perception of immersive virtual environments. In particular, the screening test did not include repetitions of each position in the localization task and it did not contain a detailed individual characterization of e.g. spectral sensitivities, or non-acoustic factors such as satisfaction and confidence. A longer screening procedure would provide a more accurate user profile at the expense of a lightweight procedure. Additionally, future research could also adopt a gamification-based approach in evaluating the localization ability of individual subjects for integrating screening tests in even more ecological use-cases.

Moreover, a longer VR exploration, a more complex scene, or a task dependent activity (e.g. a simple searching task) could possibly reveal different aspects and a new level of accessibility to audio materials. Accordingly, configuration of the VR scenes could also reveal some content-related aspects that might be integrated in good design practices.

Finally, in order to limit acoustical factors, we decided to avoid simulation of room acoustics. One can perform a similar experiment in a reverberant virtual space might yielding different results due to additional dynamic localization cues, i.e. early/late reverberations, and direct-to-reverberant energy ratio [22].

7 Conclusions

The proposed pilot study aimed to investigate perceptual differences in the experience of HRTF-based spatial audio rendering with headphones. The first experiment acted as screening procedure for identification of good and bad localizers according to a psycho-acoustic motivated threshold on localization blur in elevation. Using personalized HRTF selection based on the shape of each participant's external ear shape, the discrimination between good and bad localizers relied on elevation performance data of a simplified localization task in VR. The second experiment aimed at studying differences in VR experience of a free exploration task.

Visual dominance was modulated by perceptual abilities in decoding spectral HRTF features and localization, when all conditions were not considered as meaningful factor in the analysis; our findings suggest that good and bad localizers perceived differently audio latency in the auralization (good localizers perceived a lower latency) and that there was a modulation in the involvement of visual aspects (good localizers were less involved with the visuals). These results could be considered in the design principles of immersive VR experience in order to increase the relevance of salient stimuli and supporting users' attention in the processing of low-level features for perceptual learning [1].

Future research should further investigate how the auditory/non-acoustic aspects of users characterization influences their experience of audio in VR contexts, in order to train listeners to binaural audio with HRTFs; experimental validation with massive participation of human subjects will be highly relevant for the applicability of our findings to different VR scenarios and audio rendering techniques. It is worthwhile to note that our experimental methodology and the software implementation of our system which is based on HOBA and Unity, is technologically-ready for a widespread application in mobile VR devices.

Acknowledgments

This study was supported by the internationalization grant of the 2016-2021 strategic program "Knowledge for the World" awarded by Aalborg University to Michele Geronazzo.

References

[1] M. Ahissar and S. Hochstein. Attentional control of early perceptual learning. *Proc Natl Acad Sci U S A*, 90(12):5718–5722, June 1993.

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

[3] G. Andéol, S. Savel, and A. Guillaume. Perceptual factors contribute more than acoustical factors to sound localization abilities with virtual sources. *Front. Neurosci*, 8:451, 2015.

[4] D. W. Batteau. The Role of the Pinna in Human Localization. *Proc. R. Soc. London. Series B, Biological Sciences*, 168(11):158–180, Aug. 1967.

[5] D. R. Begault. *3-D sound for virtual reality and multimedia*. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[6] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, USA, 1983.

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

[8] K. Bormann. Presence and the utility of audio spatialization. *Presence: Teleoperators and Virtual Environments*, 14(3):278–297, 2005.

[9] P. Damaske and B. Wagener. Directional Hearing Tests by the Aid of a Dummy Head. *Acta Acustica united with Acustica*, 21(1):30–35, Jan. 1969.

[10] G. Eckel. Immersive Audio-Augmented Environments - the LISTEN Project. In *Proc. 5th IEEE Int. Conf. Info. Visualization (IV'01)*, pages 571–573, Los Alamitos, CA, USA, July 2001.

[11] M. Geronazzo, J. Kleimola, E. Sikström, A. De Götzen, S. Serafin, and F. Avanzini. HOBA-VR: HRTF On Demand for Binaural Audio in immersive virtual reality environments. In *Proc. 144th Conv. Audio Eng. Society*, Milano, May 2018.

[12] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini. 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)*, pages 397–404, Edinburgh, UK, Sept. 2017.

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

[14] M. Geronazzo, S. Spagnol, and F. Avanzini. Mixed Structural Modeling of Head-Related Transfer Functions for Customized Binaural Audio Delivery. In *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*, pages 1–8, Santorini, Greece, July 2013.

[15] M. Geronazzo, S. Spagnol, and F. Avanzini. Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 26(7):1243–1256, July 2018.

[16] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini. 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)*, pages 4496–4500, Florence, Italy, May 2014.

[17] R. Gunther, R. Kazman, and C. MacGregor. Using 3d sound as a navigational aid in virtual environments. *Behaviour & Information Technology*, 23(6):435–446, 2004.

[18] T. Hedke, J. Ahrens, J. Beyer, and S. Mller. Impact of Spatial Audio Presentation on the Quality of Experience of Computer Games. In *Proc. of Jahrestagung fr Akustik (DAGA'17)*, page 4, Kiel, 2017. German Acoustical Society (DEGA).

[19] C. Hendrix and W. Barfield. Presence in virtual environments as a function of visual and auditory cues. In *Virtual Reality Annual International Symposium, 1995. Proceedings.*, pages 74–82. IEEE, 1995.

[20] S. Hwang, Y. Park, and Y.-s. Park. Modeling and Customization of Head-Related Impulse Responses Based on General Basis Functions in Time Domain. *Acta Acustica united with Acustica*, 94(6):965–980, 2008.

[21] R. Kassier, S. K. Zielinski, and F. Rumsey. Computer games and multichannel audio quality part 2 - evaluation of time-variant audio degradations under divided and undivided attention. In *Audio Engineering Society Convention 115*, 2003.

[22] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Atten Percept Psychophys*, pages 1–23, Nov. 2015.

[23] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl. A spatial audio quality inventory (saqi). *Acta Acustica united with Acustica*, 100(5):984–994, 2014.

[24] P. Majdak, R. Baumgartner, and B. Laback. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Front Psychol*, 5:1–10, Apr. 2014.

[25] C. Mendonça, G. Campos, P. Dias, and J. A. Santos. Learning Auditory Space: Generalization and Long-Term Effects. *PLoS ONE*, 8(10):e77900, Oct. 2013.

[26] J. C. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *The Journal of the Acoustical Society of America*, 106(3):1480–1492, 1999.

[27] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan. Psychophysical customization of directional transfer functions for virtual sound localization. *The Journal of the Acoustical Society of America*, 108(6):3088–3091, Dec. 2000.

[28] A. J. V. Opstal, J. Vliegen, and T. V. Esch. Reconstructing spectral cues for sound localization from responses to rippled noise stimuli. *PLOS ONE*, 12(3):e0174185, Mar. 2017.

[29] G. D. Romigh, D. S. Brungart, and B. D. Simpson. Free-Field Localization Performance With a Head-Tracked Virtual Auditory Display. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):943–954, Aug. 2015.

[30] N. Sankaran, J. Hillis, M. Zannoli, and R. Mehra. Perceptual thresholds of spatial audio update latency in virtual auditory and audiovisual environments. *The Journal of the Acoustical Society of America*, 140(4):3008–3008, Oct. 2016.

[31] B. U. Seeber and H. Fastl. Subjective selection of nonindividual head-related transfer functions. In *Proc. 2003 Int. Conf. Auditory Display (ICAD 2003)*, pages 259–262, Boston, MA, USA, July 2003.

[32] S. Serafin, M. Geronazzo, N. C. Nilsson, C. Erkut, and R. Nordahl. Sonic Interactions in Virtual Reality: State of the Art, Current Challenges and Future Directions. *IEEE Computer Graphics and Applications*, 38(2):31–43, 2018.

[33] K. H. Shin and Y. Park. Enhanced Vertical Perception through Head-Related Impulse Response Customization based on Pinna Response Tuning in the Median Plane. *IEICE Trans. Fundamentals*, E91-A(1):345–356, Jan. 2008.

[34] E. Sikström, M. H. Laursen, K. S. Pedersen, A. De Götzen, and S. Serafin. Participatory amplitude level adjustment of gesture controlled upper body garment sound in immersive virtual reality. In *Audio Engineering Society Convention 136*, 2014.

[35] L. S. R. Simon, A. Andreopoulou, and B. F. G. Katz. Investigation of Perceptual Interaural Time Difference Evaluation Protocols in a Binaural Context. *Acta Acustica united with Acustica*, 102(1):129–140, Jan. 2016.

[36] R. H. So, B. Ngan, A. Horner, J. Braasch, J. Blauert, and K. L. Leung. Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: cluster analysis and an experimental study. *Ergonomics*, 53(6):767–781, 2010.

[37] M. Stamm and M. Altinsoy. Assessment of binaural–proprioceptive interaction in human-machine interfaces. In *The Technology of Binaural Listening*, pages 449–475. Springer, 2013.

[38] G. J. Thomas. Experimental study of the influence of vision on sound localization. *Journal of Experimental Psychology*, 28(2):163, 1941.

[39] R. Trapeau and M. Schnwiesner. The encoding of sound source elevation in the human auditory cortex. *J. Neurosci.*, pages 2530–17, Mar. 2018.

[40] 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(1):111–123, 1993. 00940.

[41] M. Wright. Open Sound Control: An Enabling Technology for Musical Networking. *Org. Sound*, 10(3):193–200, Dec. 2005.

[42] Y. Zhang, T. Fernando, H. Xiao, and A. R. L. Travis. Evaluation of auditory and visual feedback on task performance in a virtual assembly environment. *Presence: Teleoperators and Virtual Environments*, 15(6):613–626, 2006.

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