VIRTUAL REALITY EXPLORATION WITH DIFFERENT HEAD-RELATED TRANSFER FUNCTIONS

Erik Sikström
Virsabi ApS
erik.sikstrom@outlook.com

Michele Geronazzo
Dept. of Architecture, Design,
and Media Technology
Aalborg University
mge@create.aau.dk

Jari Kleimola
Hefio Ltd
jari.kleimola@hefio.com

Federico Avanzini
Dept. of Computer Science
University of Milano
federico.avanzini@di.unimi.it

Amalia de Götzen, and Stefania Serafin
Dept. of Architecture, Design,
and Media Technology
Aalborg University
[ago,sts]@create.aau.dk

ABSTRACT

One of the main challenges of spatial audio rendering in headphones is the crucial work behind the personalization of the so-called head-related transfer functions (HRTFs). HRTFs capture the listener's acoustic effects allowing a personal perception of immersion in virtual reality context. This paper aims to investigate the possible benefits of personalized HRTFs that were individually selected based on anthropometric data (pinnae shapes). Personalized audio rendering was compared to a generic HRTF and a stereo sound condition. Two studies were performed; the first study consisted of a screening test aiming to evaluate the participants' localization performance with HRTFs for a non-visible spatialized audio source. The second experiment allowed the participants to freely explore a VR scene with five audiovisual sources for two minutes each, with both HRTF and stereo conditions. A questionnaire with items for spatial audio quality, presence and attention was used for the evaluation. Results indicate that audio rendering methods made no difference on responses to the questionnaire in the two minutes of a free exploration.

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 discuss its applications and relevance in immersive VR experiences [1].

Previous research has shown that spatial sound has a positive influence on performance in wayfinding tasks [2], and in localization performance in an audio-haptic task [3]. Furthermore, Zhang et al. [4] used audio feedback with head-related transfer functions (HRTFs) 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 observed that the inclusion of spatial audio yielded higher presence-questionnaire ratings after their subjects had explored their virtual environment [5]. 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 investigated the utility of spatial audio in relation to presence when the audio feedback was both task relevant or not [6]. In Bormann's study, the virtual environment was presented on a desktop computer. The participants were asked to search an environment for either an object that also was an audio source (a radio playing music), or search 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.

In this paper, we aim to continue this line of work by investigating the possible contributions of HRTF-based spatialization to perceived spatial audio quality and sensation of presence and attention within an immersive virtual reality context. In particular, we are interested in comparing HRTFs that are individually selected based on anthropometric data of the pinna, against a generic dummy-head HRTF, and a $2D$ stereo condition. To our knowledge, per-

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

sonalized HRTFs selection has not been evaluated in studies within immersive VR before, nor native VR engines support a personalization stage. However, it has been previously shown that individually selected HRTFs result in improved localization ability for elevation cues within psychophysical tests [7, 8].

It is worthwhile to notice that listening with non-individual HRTFs exhibits high variability in localization performance related to differences in acoustic factors due to listener anthropometry, and to perceptual factors, i.e., the individual ability of encoding directional information [9]. Accordingly, it is necessary to evaluate HRTF-based spatialization prior to the VR experience, designing a fast pre-experiment screening test able to to investigate the localization ability of each subject replacing time- and resource- consuming psychoacoustic tests.

The remainder of this paper is structured as follows. Section 2 describes previous research concerning HRTF selection procedures. Section 3 describes the technical implementations of the audiovisual virtual reality applications that were used in the current study. The experiments themselves are described in detail in Section 4, where Study 1 details a screening procedure used for investigating the localization ability of the participants while using generic or customized HRTFs. Additionally Study 2 allowed the participants to explore a virtual environment using generic and customized HRTFs, as well as with a stereo condition. The results of both studies are presented in Sec. 5. Section 6 discusses the findings of the experiments, and Sec. 7 summarizes the paper and the final conclusions.

2. RELATED WORKS ON HRTF SELECTION

The measurement of individual HRTFs usually requires a special measuring apparatus in a time-consuming procedure, leading to unpractical solutions for listeners involved in every-day applications. Alternative methods for HRTF personalization are usually preferred looking for a delicate trade-off between audio quality and handiness of the personalization procedure.

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 [10]. However, the increase of available HRTF data during the last decade supports the research process towards novel selection processes of non-individual HRTFs. [2]

Typically, HRTF selection problems are characterized by:

- **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 [11]), *dataset reduction* [12], and *dimensionality reduction* [13] can be performed prior the HRTF selection.

Accordingly, for the desired domains of action and spatial ranges, one can adopt several approaches such as anthropometric database matching, exploiting linear regression models between acoustical and anthropometric features, relying on subjective selection, or minimizing differences between HRTFs in the acoustic domain [14]. Once one or a set of best HRTF candidate are identified, listener can self-tune each HRTF set acting on spectral manipulations and enhancement [7], and adjusting weights [15]; moreover, a period of adaptation to non-individual HRTFs can be characterized by multimodal feedback to correct answer of localization/discrimination tasks [16].

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 provided the information to animate a visual avatar according to the subject's movements.

3.1 Apparatus for immersive virtual reality

The system used for the two studies are presented in Fig. 1. The graphics rendering, audio and motion tracking software was running on one Windows 7 PC computer (Intel i7-4470K 3.5GHz CPU, 16 GB RAM and a MSI Gaming X GeForce GTX 1070 graphics card). The HMD used was a nVisor SX with a FOV of 60 degrees with a screen resolution 1280x1024 pixel in each eye. The audio feedback was delivered through a RME Fireface 800 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. Virtual environments was developed with Unity3D v4.6 [3] .

[3] https://unity3d.com/

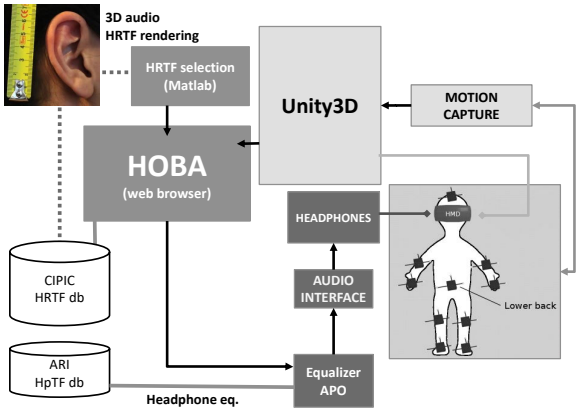

Figure 1. System overview.

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

3.2 Spatial audio rendering

3.2.1 HRTF selection tool

We adopted the Matlab tool developed by Geronzzo *et. al* [8, 17] that implements the method of mapping anthropometric features into the HRTF domain, following a ray-tracing modeling of pinna acoustics [14]. 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 available HRTFs that were considered from the CIPIC database in this study. [4]

From [14], we know that one can consider only the first and most prominent notch associated with the most external pinna contour on the helix border (the "C_1" contour hereafter); thus N estimates of C_1 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; (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.

In this study, we set $N = K = 10$ 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 [17]; 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 equally the 1^{st} ranking positions of the following three mismatch functions:

- **Mismatch:** each HRTF is assigned a similarity score that corresponds exactly to increasing values of the mismatch function calculated with Eq. (1) (for a single (k, n) pair).
- **Ranked position:** each HRTF is assigned a similarity score that is an integer corresponding to its ranked position taken from the previous mismatch values (for a single (k, n) pair).
- **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.

[4] https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/

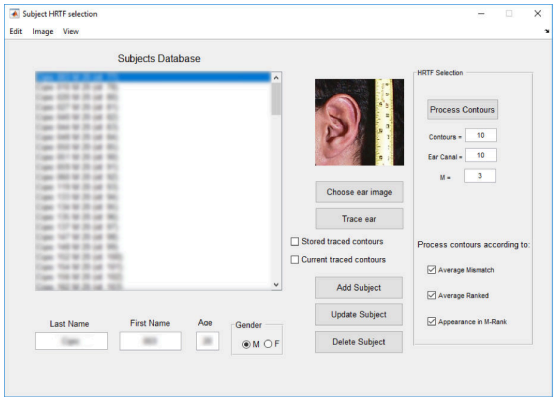

Figure 2. Tool for HRTF selection with pinna anthropometry: main graphical user interface.

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. 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 [18] 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 the HRTFs On-demand for Binaural Audio (HOBA) rendering framework for web browsers. 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. An overview of the technical description of the framework together with git repository information has been published in [19].

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 [20] 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 [21]. 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.

[5] http://sofacoustics.org/data/headphones/ari
[6] https://sourceforge.net/projects/equalizerapo/

4. EXPERIMENTS

4.1 Study 1 - Screening test

The aim of Study 1 was to conduct a screening of the subject pool's abilities of accurately locating spatialized sounds, presented in two audio rendering conditions, either a generic HRTF (a dummy head - CIPIC subject 165, *Generic* hereafter) or a customized HRTF selection with the method described in 3.2.1 (from now on referred to as *Custom*).

The main focus of this experimental design was to keep the execution quick and comfortable for participants (10 minutes maximum) in such a way to be used as screening test before any immersive virtual reality experience. A first attempt was conducted in a previous study [22] with non-visual virtual reality environments with the following experimental approach: a goal-reaching task provided navigation performances that were sensitives to elevation perception with customized HRTFs. In this work, we adopted a typical sound source localization task and the test was 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 1m radius. The sphere was also equipped with lines indicating the horizontal, median and traversal planes. The implementation is part of the HOBA-VR framework [19].

The auditory stimuli was a train of noise bursts, presented at 60 dBA level [8] when measured from the earphone cup; directional filtering through HRTFs rendered 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 led to a total of 6 (azimuths) $\times$ 4 (elevations) $+$ 1 (elevation 90°) spatial locations; at the start of each session, subject head was located at the origin of the coordinate system. The distance of the sound sources was fixed set to 1m, which corresponds coherently with CIPIC HRTF measurements in the far-field and to the dimensions of the sphere in the visual environment. The presentation order of these locations was randomized; test locations were presented once per audio-rendering condition.

A game controller with a virtual representation of a laser pointer was implemented allowing participants 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. After each condition, a break was issued and the participant was asked to fill in a short questionnaire with items related to their performance in the localization task. The questionnaire items were the following:

- Q1: Localizability - Estimating the location of the sound source was (More difficult - Easier)
- Q2: Did you perceive elevation? (Yes - No)
- Q3: Satisfaction - How satisfied were you with your own performance? (Not at all - Very much)
- Q4: Confidence - How confident were you that you pointed at the correct location of the sound source? (Not very confident - Very confident)

Q1 was adopted from previous literature [23]. Q2 was adopted similarly to the screening test conducted in an earlier study [22]; Q1, Q3 and Q4 were provided with seven-point rating scales.

Type	*Behavior*	*Level (dBA)*
Old transistor radio	Static - positioned at a table while playing a static radio noise	45.5
Fireplace	Static - placed at ground level, playing a looped fire recording	35.3
Bird	Static - placed in a tree at approximately head height, playing a loop of birdsong with twittering heard at regular intervals	50
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	36.6
Grasshopper	Static - positioned at ground level in a tuft of grass at the side of the path	32.7

Table 1. Sound samples and their reference loudness level at 1 m from the measurement point.

4.2 Study 2 - Virtual reality scene

The aim of Study 2 was to evaluate the effect of spatial audio when presented with a slightly more complex virtual scene when compared to previous studies, using fewer audio sources [5]. A night scene with a partially lit path in an area of sand dunes was designed to accommodate this experiment. 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 stimula. These audiovisual sources are described in table 1.

The area in which the sound sources were placed was surrounded by a stone wall to remind the participants not to attempt to wander away from the scene. Invisible collider-walls were also added to prevent this.

Three 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;

Subj. ID	$Az_{generic}$	$El_{generic}$	$Slope_{generic}$	Az_{custom}	El_{custom}	$Slope_{custom}$
7	72.47, ±42.28	38.61, ±25.38	-0.2	26.71, ±26.15	18.72, ±14.86	0.77
8	10.15, ±8.65	31.24, ±23.12	0.54	30.77, ±44.12	33.75, ±16.17	0.64
9	34.89, ±50.8	39.31, ±28.95	0.38	26.29, ±35.42	33.1, ±28.22	0.59
10	7.93, ±7.58	22.39, ±17.01	0.71	10.58, ±19.08	28.22, ±19.87	0.28
11	15.77, ±31.34	24.92, ±19.85	0.33	12.86, ±16.69	32.97, ±18.67	0.54
13	31.07, ±50.83	27.97, ±25.72	0.43	35.98, ±45.32	25.71, ±23.76	0.32
14	10.58, ±12.08	20.87, ±21.93	0.27	20.68, ±38.88	27.76, ±19.2	0.07
15	9.12, ±8.73	10.72, ±9.48	0.84	3.67, ±2.97	15.58, ±12.47	0.63
16	30.29, ±37.49	31.16, ±22.87	0.004	22.85, ±37.43	32.4, ±22.5	0.005
17	19.54, ±21.22	38.67, ±31	0.13	6.18, ±4.57	31.15, ±18.75	0.2
18	27.32, ±32.91	33.79, ±23.1	0.43	6.26, ±7.61	17.6, ±14.77	0.73
19	42.12, ±35.62	34.33, ±20.6	-0.08	33, ±34.21	30.27, ±17.9	0.01

Table 2. The mean-values, standard deviations for azimuth and elevation errors in degrees and slope-values obtained during the screening test in Study 1, for the Generic HRTF condition (left side) and the Custom HRTF condition (right side).

- *Custom HRTF:* 3D audio with HOBA loading an individually selected HRTF set with the tool described in 3.2.1.

The distance attenuation in Unity (the Volume Rolloff setting) was set to "Logarithmic".

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 an walking-in-place locomotion technique, using an algorithm described in [24]. 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. Hence, real walking was not a possible solution.

The experiment involved three trials in randomized order. One for each audio condition. After each trail, a break was issued and the subjects were asked to fill in a questionnaire with questions regarding the level of experienced presence, spatial audio quality (adapted from [23]) and attention [6]. 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.

Item	*Global*
Q1 Localizability	4.17, ±1.52
Q3 Satisfaction	3.96, ±1.52
Q4 Confidence	3.83, ±1.58

Table 3. The mean and standard deviation values of the responses to the questionnaire items from Study 1 (seven-point ratings), for both audio conditions *(Global)*.

5. RESULTS

Ninteen subjects participated voluntarily in the study, but a number of participant data were removed due to technical issues. After that, twelve participants remained. 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. They also reported their previous amount of experience with immersive virtual reality. One subject had no experience, three had little, one was experienced, and seven were very experienced.

5.1 Study 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 it's perceived position i.e. the logged coordinates from the virtual laser pointer. From this information, a linear regression analysis was performed on the elevation errors only. It is known

Item	*Global*	χ^2	*df*	*p*
Q1 Externalization	5.5, $\pm$1.25	4.05	2	.13
Q2 Responsiveness	2.42, $\pm$1.48	2.64	2	.27
Q3 Naturalness	5.33, $\pm$1.01	2.26	2	.32
Q4 Presence	5.22, $\pm$1.12	.41	2	.81
Q5 Att. audio	5.31, $\pm$1.69	1.11	2	.57
Q6 Att. visuals	4.17, $\pm$1.48	4.84	2	.09
Q7 Realistic	4.5, $\pm$1.38	.26	2	.88

Table 4. 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)*.

from the literature that performances in vertical localization vary remarkably among individuals more than horizontal/azimuthal localization [9].

Along with the screening test, a questionnaire was administered after each pointing task evaluating the perceived localizability, satisfaction with performance, and confidence with performance. Due to non-normality of data distributions, Wilcoxon signed-rank tests were adopted in order to investigate if the responses were statistically different between the *Generic* and the *Custom* conditions.

No statistically significant differences were found between the audio rendering conditions in any of the approaches to grouping the subjects. The mean-values and standard deviations for each questionnaire item are presented in table 3. Apart from this, two subjects reported that they had not noticed any elevation in *Custom* condition, and one reported that they did not notice any elevation in the *Generic* condition.

5.2 Study 2 - Virtual reality scene

For the second 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. Data was analyzed in order to investigate if there were statistically significant differences after experiencing the environment among audio conditions. Due to non-normality of data distribution, non-parametric tests were performed: Friedman's test and repeated Wilcoxon signed-rank tests with Bonferroni correction. However, no statistically significant differences were found between the audio rendering conditions on the questionnaire items. The mean and standard deviations from all questionnaire items, grouped by condition, are presented in Fig. 3, and all the audio conditions combined are presented in Table 4 together with χ^2 statistics. 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 14, was one of those who could be considered bad localizer due to a low slope-value from the screening tests. 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, also here there were no significant differences found among the questionnaire responses.

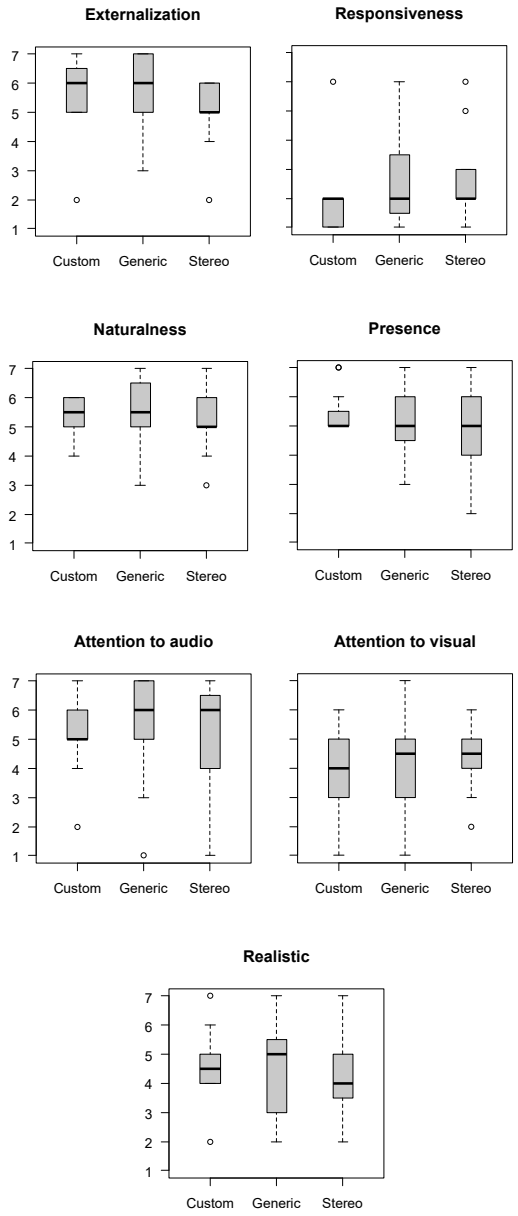

Figure 3. Responses to questionnaire items grouped by condition in Study 2.

6. DISCUSSION

The two HRTF conditions in the screening test were not rated differently on the performance related questionnaire items, between the two 3D audio condition and stereo condition. The two HRTF conditions in the screening test were not rated differently on the performance related questionnaire items, between the two 3D audio condition and stereo condition. A longer screening procedure including

repeated evaluations of each position could possibly yield different results with higher reliability. However, keeping the screening test short was one of the main motivations in our study thus we were not surprised at this result.

The results from Study 2 show that there were no perceived difference among the audio rendering conditions, when evaluated on questionnaire items that were used in the study. The additional analysis comparing the two HRTF conditions with the stereo condition also reinforces this. Reasons for 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) [25, 26], and the division of attention associated with interactive tasks and audio quality evaluations. Previous research on the influence of interactive tasks on audio quality evaluations have 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) [27, 28]. Generally, the outcomes of these studies have found that the users in the active conditions were more tolerant to degradations. Some of these results goes partly against previous work, for example Barfield and Hendrix's study on spatialized audio [5], 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 [6]. 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.

6.1 Limitations of the study

The screening test in Study 1 included no repetitions of each position in the localization task. A longer screening procedure would provide a more detailed listener characterization at the expense of a lightweight procedure. As for general limitations of study 2, the short exposure time (2 minutes of VR experience) and the small number of participants are factors that possibly narrowed the applicability of our results to wider VR contexts. Furthermore, in order to limit acoustical factors, there were no simulation of room acoustics enabled in this VR experience. A similar experiment conducted in a reverberant virtual space might yield different results due to additional dynamic localization cues, i.e. early/late reverberations, and direct-to-reverberant energy ratio [29].

7. CONCLUSIONS

The studies conducted in this paper aimed to investigate differences in the experience of HRTF-based spatial audio rendering with headphones. The first experiment used a screening procedure for assessing user localization performances using either a generic HRTF or a customized HRTF selection based on the shape of each participant's pinnae. The second experiment attempted to study differences in the experience of a virtual reality environment. A questionnaire was used for this purpose, however there were no statistically significant differences found between the audio conditions for any of the questionnaire items probably due to visual dominance.

Future research should further investigate how the auditory side of user characterization influence their experience of audio in virtual reality contexts; experimental validation with massive participation of human subjects will be highly relevant for the applicability of our findings to different VR scenarios and HRTF selection procedures. It is worthwhile to notice 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, such as Google Cardboard, Samsung Gear VR, or Oculus Go.

Acknowledgments

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

8. REFERENCES

[1] 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*, vol. 38, no. 2, pp. 31–43, 2018.

[2] R. Gunther, R. Kazman, and C. MacGregor, "Using 3d sound as a navigational aid in virtual environments," *Behaviour & Information Technology*, vol. 23, no. 6, pp. 435–446, 2004.

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

[4] 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*, vol. 15, no. 6, pp. 613–626, 2006.

[5] 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.* IEEE, 1995, pp. 74–82.

[6] K. Bormann, "Presence and the utility of audio spatialization," *Presence: Teleoperators and Virtual Environments*, vol. 14, no. 3, pp. 278–297, 2005.

[7] 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*, vol. 108, no. 6, pp. 3088–3091, Dec. 2000.

[8] 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)*, Florence, Italy, May 2014, pp. 4496–4500.

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

[10] H. Møller, M. Sørensen, J. Friis, B. Clemen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?" *J. Audio Eng. Soc*, vol. 44, no. 6, pp. 451–469, 1996.

[11] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1480–1492, 1999.

[12] 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*, vol. 53, no. 6, pp. 767–781, 2010.

[13] 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*, vol. 94, no. 6, pp. 965–980, 2008.

[14] 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1243–1256, Jul. 2018.

[15] 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*, vol. E91-A, no. 1, pp. 345–356, Jan. 2008.

[16] C. Mendonca, G. Campos, P. Dias, and J. A. Santos, "Learning Auditory Space: Generalization and Long-Term Effects," *PLoS ONE*, vol. 8, no. 10, p. e77900, Oct. 2013.

[17] 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)*, Edinburgh, UK, Sep. 2017, pp. 397–404.

[18] M. Wright, A. Freed *et al.*, "Open soundcontrol: A new protocol for communicating with sound synthesizers." in *ICMC*, 1997.

[19] M. Geronazzo, J. Kleimola, E. Sikstrm, A. De Gtzen, 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.

[20] 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*. Audio Engineering Society, Oct. 2014.

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

[22] M. Geronazzo, F. Avanzini, and F. Fontana, "Auditory navigation with a tubular acoustic model for interactive distance cues and personalized head-related transfer functions," *J Multimodal User Interfaces*, vol. 10, no. 3, pp. 273–284, Sep. 2016.

[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*, vol. 100, no. 5, pp. 984–994, 2014.

[24] 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*. Audio Engineering Society, 2014.

[25] C. Jackson, "Visual factors in auditory localization," *Quarterly Journal of Experimental Psychology*, vol. 5, no. 2, pp. 52–65, 1953.

[26] H. Witkin, S. Wapner, and T. Leventhal, "Sound localization with conflicting visual and auditory cues." *Journal of experimental psychology*, vol. 43, no. 1, p. 58, 1952.

[27] F. Rumsey, P. Ward, and S. K. Zielinski, "Can playing a computer game affect perception of audio-visual synchrony?" in *Audio Engineering Society Convention 117*. Audio Engineering Society, 2004.

[28] U. Reiter and M. Weitzel, "Influence of interaction on perceived quality in audiovisual applications: evaluation of cross-modal influence." Georgia Institute of Technology, 2007.

[29] 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*, pp. 1–23, Nov. 2015.