applied sciences

MDPI

Article

Head-Related Transfer Functions for Dynamic Listeners in Virtual Reality

Olli S. Rummukainen [1,2,*], Thomas Robotham [1,3] and Emanuël A. P. Habets [1,2,3]

[1] International Audio Laboratories Erlangen, 91058 Erlangen, Germany; thomas.robotham@audiolabs-erlangen.de (T.R.); emanuel.habets@audiolabs-erlangen.de (E.A.P.H.)
[2] Fraunhofer Institute for Integrated Circuits, 91058 Erlangen, Germany
[3] Faculty of Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
* Correspondence: olli.rummukainen@iis.fraunhofer.de

Abstract: In dynamic virtual reality, visual cues and motor actions aid auditory perception. With multimodal integration and auditory adaptation effects, generic head-related transfer functions (HRTFs) may yield no significant disadvantage to individual HRTFs regarding accurate auditory perception. This study compares two individual HRTF sets against a generic HRTF set by way of objective analysis and two subjective experiments. First, auditory-model-based predictions examine the objective deviations in localization cues between the sets. Next, the HRTFs are compared in a static subjective ($N = 8$) localization experiment. Finally, the localization accuracy, timbre, and overall quality of the HRTF sets are evaluated subjectively ($N = 12$) in a six-degrees-of-freedom audio-visual virtual environment. The results show statistically significant objective deviations between the sets, but no perceived localization or overall quality differences in the dynamic virtual reality.

Keywords: head-related transfer function (HRTF); spatial audio; virtual reality; binaural rendering; audio-visual integration; 6-degrees-of-freedom

Citation: Rummukainen, O.S.; Robotham, T.; Habets, E.A.P. Head-Related Transfer Functions for Dynamic Listeners in Virtual Reality. *Appl. Sci.* **2021**, *11*, 6646. https://doi.org/10.3390/app11146646

Academic Editor: Hyunkook Lee

Received: 14 June 2021
Accepted: 17 July 2021
Published: 20 July 2021

1. Introduction

Virtual sounds, rendered to headphones, often fail to localize in space or externalize from within the listener's head. Sound localization in a 3-dimensional space relies on acoustic cues described by the head-related transfer functions (HRTFs) [1]. Using the listener's own HRTFs results in the best localization performance in static listening, but measuring or modeling individual HRTFs remains a challenging process. With head tracking and multimodal environments, such as in virtual and augmented reality (VR, AR), generic, non-individual HRTFs may still enable accurate auditory perception comparable to individual sets. This paper presents an objective analysis and two subjective experiments to compare two different individual HRTF sets with a generic set in 6-degrees-of-freedom (6-DoF) VR, where rotational movement around the x, y, and z axes (pitch, yaw, and roll), and translational movement along the axes (surge, strafe, and elevation) are possible.

Static, non-head-tracked listening with generic HRTFs results, most evidently, in localization errors in elevation and in front-back or quadrant confusions when compared to free-field listening in an anechoic chamber or with individual HRTFs [2–4]. Similar findings of increased elevation and quadrant errors with generic HRTFs during static reproduction have also been recently reported [5]. Additionally, externalization may be degraded when listening with generic HRTFs compared to individual HRTFs, as noted in a recent review [6]. However, the results regarding modified spectral cues are mixed, and other cues, such as reverberation, movement, and vision, may be more critical factors for robust externalization.

Studies going beyond static scenarios have found a reduction in front-back confusions when head movements are allowed [7–9] both with and without headphones. Significant decreases in elevation errors have also been identified when moderate head motions are

permitted in free-field listening [10]. Few studies investigate auditory perception during full-body 6-DoF movement. Although the listener is in translational motion, the minimum audible angle has been found to increase compared to static listening [11]. This finding hints towards an increase in localization blur during full-body motion. Such phenomena may mask any potential localization benefits granted when using individualized over generic HRTF sets.

Directly comparing individual and generic HRTFs in dynamic localization tasks showed mixed results. In some papers, no effect on localization accuracy between individual and generic HRTFs is found when head-tracked rendering is used [5,8]. At the same time, some studies show significant improvements in localization accuracy or preference for individual over generic HRTFs, even in dynamic conditions [12,13]. However, localization is only one attribute of quality in a VR environment. One study compared the evaluation of overall audio quality either using generic or individual HRTFs in the context of 6-DoF VR critical listening via a multistimulus method. Apart from a small effect in the participants' sensitivity to elevation errors in favor of individual HRTFs, all HRTF sets yielded similar quality scores [14]. No perceptual differences were found between realistic binaural auralizations of a classroom scenario when rendered with either individual or generic HRTFs [15]. Finally, an experiment conducted using spatialized music content found a general preference of generic HRTFs over individual HRTFs [16].

Instead of arduously measuring individual HRTFs, different strategies have been developed to select the anthropometrically closest matching HRTF set from an existing HRTF database [17]. Based on an auditory model, the estimated localization accuracy has been shown to objectively increase when using anthropometric parameters to select the most suitable HRTF set [18]. However, anthropometry-based HRTF matching has been perceptually investigated in VR, where no effect on questionnaire results was found between the best match HRTFs and generic HRTFs for a free exploration task [19]. Comparing best and worst match HRTFs in a VR shooter game found a minimal impact of the HRTFs on participant performance for the most elevated target regions. Still, the benefit was limited to the early stages of the game, signaling a potential adaptation to the HRTFs as the game progressed [20].

Multiple sensory modalities help us to construct a mental representation of our surroundings [21]. In VR, audio-visual cross-modal training quickly improves source localization accuracy after short training periods [22–24]. Active tasks, such as head movements [25] or interaction with a sound source [26], are found to further increase the adaptation from our biological set to altered localization cues. A many-to-one mapping mechanism in sound localization has been suggested, where the plasticity in the auditory cortex allows us to learn multiple HRTF sets that lead to the same percept [27].

The emerging picture from prior literature indicates mixed results for the need of individual HRTFs in 6-DoF VR. On the one hand, static localization is more accurate with individual HRTFs than with generic HRTFs. On the other hand, VR experiences are head-tracked and multimodal, allowing the listener to use different modalities and motor actions to support auditory localization and ultimately adapt the auditory perception. From an objective perspective, the use of individual HRTFs provides more accurate localization cues tailored for each listener, but the perceptual evidence gathered so far in VR environments does not show uniform support for the need of individual audio delivery. This study compares two individual HRTF sets, attained using measurements or computational modeling, and a generic HRTF set, first objectively using auditory models, then followed by two subjective experiments addressing: (I) static localization; and (II) dynamic audio-visual localization, timbre, and overall quality in a 6-DoF VR. Additionally, another generic HRTF set, containing only horizontal plane measurement points, is included in the dynamic audio-visual localization experiment.

Hypothesis 1. *The two sets of individual HRTFs are objectively different from the generic HRTF on average.*

Hypothesis 2. *The individual HRTF sets are objectively similar to each other on average.*

Hypothesis 3. *Static localization accuracy is higher with individual HRTFs compared to generic HRTFs.*

Hypothesis 4. *The HRTF sets are equally preferred in localization, timbre, and overall quality in 6-DoF VR.*

2. HRTF Sets and Analysis

2.1. HRTF Sets

Three different HRTF sets were obtained for the experiment: a *measured*, *modeled*, and a *generic* set. The measured and modeled HRTFs were produced for the same group of people ($N = 12$), who also participated in the subjective experiments. A fourth HRTF set was derived from the generic HRTF set by keeping only the measurement points located in the horizontal plane and utilized in the second subjective listening experiment.

Individual HRTFs were measured in a semi-anechoic chamber at 1.2 m distance with a procedure described in [28]. The measurement setup comprised an array of 64 loudspeakers mounted on a semi-circle in elevation direction. A multiple exponential sweep was used to allow overlapping signal playback from multiple loudspeakers in parallel. The spatial resolution was 5° in azimuth and 2.5° in elevation covering the full azimuth circle and elevations from +90° to −70° resulting in $Q = 4608$ measurement positions with a filter length of 386 samples. The measurements were made at the entrance to the blocked ear canal. The floor reflection was cut out from the measurements in post-processing, and a low-frequency boost was applied to compensate for the non-flat magnitude response of the small drivers in the measurement setup.

The modeled HRTFs were based on high-precision 3D scans of the participants' pinnae and a set of anthropometric measurements of the head size and ear position. The measurements were used to scale the shape and dimensions of a scanned generic head-and-torso model, onto which the individual pinna models were stitched. A boundary element method numerical simulation then produced the modeled HRTFs [29]. Each modeled HRTF set had the same distribution and distance of measurement points and the same filter length as the measured HRTF set.

The generic binaural head (Neumann KU100) HRTFs were obtained from the spatial audio for domestic interactive entertainment (SADIE) database (https://www.york.ac.uk/sadie-project/index.html (accessed on 29 May 2019)). These measurements comprise $Q = 1550$ measurement positions at 1.5 m distance with azimuthal increments of 5° and elevation increments of 10° with added measurement points to satisfy specific Ambisonics reproduction setups. The head-related impulse response length is 256 samples. Notably, the KU100 binaural head does not include a torso, removing the shoulder reflection effect from the generic HRTFs. The measured, modeled, and generic HRTF set signal processing chains were level aligned to 68 dB_A sound pressure level at 1 m distance using pink noise and a binaural head with a virtual sound source rendered directly in front.

Following [4,30], we employed a decoupled equalization technique using a reference sound field, in our case the diffuse field, to equalize the measured HRTFs and the headphones. This technique is opposed to a non-decoupled equalization, where a headphone transfer function (HPTF) is measured with the same measurement setup (including the head and microphones) as for the recording of the HRTFs. The decoupled equalization technique is found to offer the same degree of fidelity as using the HPTF approach, but with the added flexibility of equalizing the recording system and reproduction system in independent sessions using different equipment. The HRTFs were diffuse-field equalized for this experiment by calculating the average magnitude over all directions, inverting it, and creating a minimum-phase filter, which was then convolved with the HRTFs. The KU100 binaural head readily produces diffuse-field equalized measurements.

2.2. Objective Analysis

The objective analysis used two different auditory models, both implemented in the Auditory Modeling Toolbox [31], to gain objective metrics on the deviations in azimuth and elevation localization cues between the measured, modeled, and generic HRTF sets. We used a binaural azimuth estimation model [32] to obtain interaural time and level difference (ITD; ILD) estimates for each participant ($N = 12$) and the KU100 generic binaural head. The equivalent rectangular bandwidth (ERB) center frequency was set to 1000 Hz for ITD modeling and 4300 Hz for ILD modeling.

Figure 1 displays the auditory-model-based estimates for the ITD (upper panel) and ILD (lower panel). Focusing on the maximum ITDs, the $\phi = 90°$ azimuth horizontal plane mean ITD for the set of measured HRTFs is 723 µs (SD = 27 µs) and for the modeled HRTFs 669 µs (SD = 26 µs). For the KU100 generic head, the modeled ITD at 1 kHz is 786 µs. The difference in mean ITDs between the individual sets and compared to the generic set are above the just noticeable difference value of 20 µs found in previous literature [33]. Additionally, a paired sample Wilcoxon signed rank test showed that there is a statistically significant difference between the ITD populations ($p = 0.002$) and one-sample Wilcoxon signed rank tests showed significant median differences compared to the generic head ITD (measured $p < 0.001$, modeled $p < 0.001$).

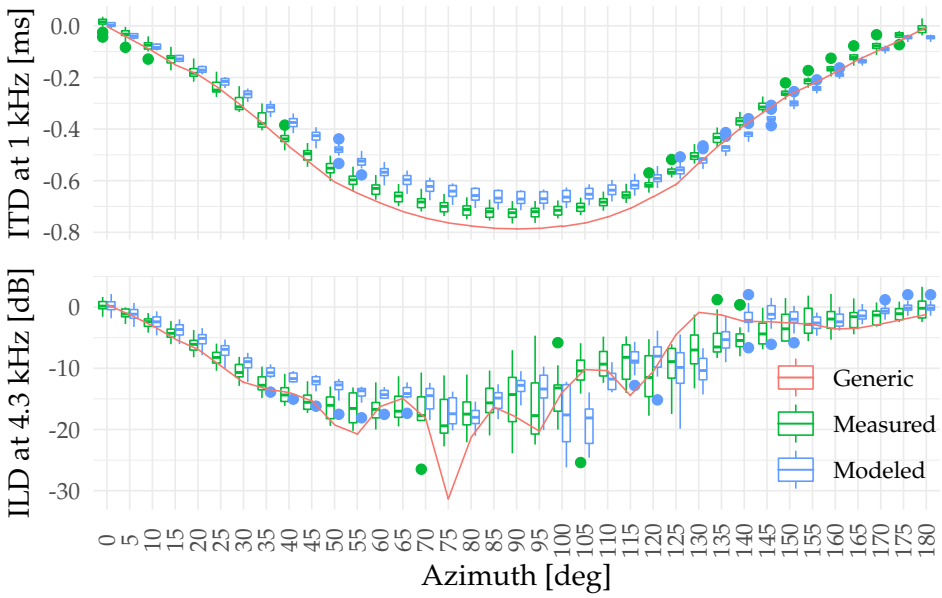

Figure 1. ITD in the ERB band centered at 1.0 kHz (bandwidth 133 Hz) and the ILD in the ERB band centered at 4.3 kHz (bandwidth 489 Hz) based on an auditory model by Dietz et al. [32]. The boxplots show the sample median and the first and third quartiles, any data within 1.5 times the interquartile range, and any outliers outside the range. Azimuth $\phi = 0°$ corresponds to the direction in front of the listener and positive angles are clockwise towards behind the head.

The deviations in ILD are more difficult to summarize. Still, a consistent ILD discrepancy can be observed in the azimuth range $\phi = \{30, ..., 70\}°$, where the measured and generic HRTFs yield larger ILD values compared to the modeled HRTFs. Further to the side the ILD predictions become more varied between the HRTF sets, before converging to similar values beyond $\phi = 160°$ behind the head.

For localization estimates in elevation, we used a sagittal-plane sound localization model [34]. This model assumes the human brain holds an internal *template* set of learned HRTFs, which can be compared with an incoming sound filtered with a *target* HRTF that can either be the person's own HRTF or a generic HRTF with some degree of deviation from the template. The model predicts the percentage of quadrant errors ($\theta > 90°$) and the root-mean-square (RMS) local elevation errors in degrees ($\theta \leq 90°$) based on a comparison

process between the target representation and template HRTFs. Here, the prediction was conducted for the elevation angles from −70° to +250° in 2.5° increments in the 0° azimuth sagittal plane.

Table 1 displays the prediction results averaged over all elevation angles for the cases: (I) template and target are the modeled HRTFs; (II) template and target are the measured HRTFs; (III) template are the measured HRTFs, target the modeled HRTFs; (IV) template are the modeled HRTFs, target generic HRTFs; and (V) template are the measured HRTFs, target generic HRTFs. The results are further depicted graphically in Figure 2. The RMS local elevation errors for matching template and target HRTFs are significantly lower than the RMS elevation error averages when combining measured templates with modeled target HRTFs. Similar large elevation error values are observed when combining generic target HRTFs with the modeled or measured HRTF templates.

Table 1. Auditory-model-based predicted mean unsigned elevation errors and quadrant errors with the bootstrapped 95% confidence intervals.

Template	Target	Unsigned Elevation Error	Quadrant Errors
Modeled	Modeled	31.0° (CI 29.8° to 32.2°)	16.0% (CI 13.5% to 18.6%)
Measured	Measured	37.5° (CI 34.9° to 40.3°)	20.3% (CI 15.9% to 25.1%)
Measured	Modeled	50.3° (CI 49.0° to 51.7°)	41.7% (CI 38.4% to 45.7%)
Modeled	Generic	48.7° (CI 46.8° to 50.7°)	41.9% (CI 38.7% to 44.8%)
Measured	Generic	46.6° (CI 45.0° to 48.2°)	39.6% (CI 37.3% to 41.9%)

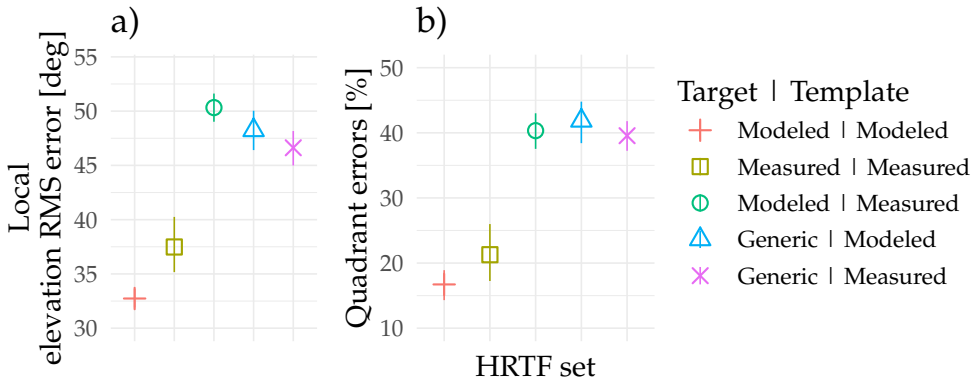

Figure 2. Auditory-model-based predictions of the (**a**) local ($\theta \leq 90°$) elevation RMS error and (**b**) quadrant errors ($\theta > 90°$). The error bars denote the bootstrapped 95% confidence intervals of the mean. The applied model is the Baumgartner et al. [34].

Comparable trends are observed with the predicted quadrant errors percentages, where the matching modeled HRTFs show the least quadrant errors followed by the matching measured HRTFs. Mixing measured template HRTFs with modeled target HRTFs yields a significantly higher errors percentage. A similar range of quadrant errors is observed for the generic target HRTFs combined with modeled or measured template HRTFs.

3. Subjective Experiments Overview

There were two sessions of subjective experiments: static localization (Section 4) and dynamic 6-DoF virtual reality (Section 5). The experiments were conducted starting with the dynamic VR experiment followed by the static localization experiment, but reported here in reversed order for readability. There was a gap of more than 12 months between the respective experiments. The dynamic VR experiment was further divided into a main experiment and a follow-up, where two additional scenes focusing on localization in elevation were evaluated a month after the main experiment. The following paragraph provides an overview of the participants in each of the subjective sessions.

The sample size for the subjective experiments was 12 participants, limited only due to the difficulties in ascertaining high fidelity measured and modeled individual HRTF sets. Subgroups of the subject pool with slight variations in numbers participated in the two experiment due to availability. For the *static localization experiment*, 8 people took part (1 female). Their average age is 37.4 years (SD = 11.7). In total, 12 people (2 female) took part in the *main dynamic 6-DoF VR experiment*. Their average age is 38.6 years (SD = 9.8). In the *follow-up dynamic 6-DoF VR experiment*, the two scenes focusing on localization in elevation, *Above* and *Below*, were evaluated on a separate day, where a subset of 8 participants (1 female) from the full pool did the test. This group had an average age of 37.5 years (SD = 9.3).

The participants were employees of Fraunhofer IIS or the International Audio Laboratories Erlangen at the time of the experiments, and can be considered expert listeners in audio quality evaluations [35]. However, they were not trained in VR-specific listening tests, and they were naive about the specific conditions under examination here. All of the participants reported no hearing impairments and had a normal or corrected-to-normal vision by self-report. All participants provided written informed consent to participate in each session of the experiment.

4. Experiment I: Static Localization

4.1. Stimuli

The experiment was constructed using an audio-visual virtual environment created in Max 8. Head-tracking, visual display, and user interaction system were provided by the HTC Vive Pro head-mounted display (HMD), using the SteamVR Version 2.0 tracking system. A visual interface in VR showed an aiming cross-hair and three concentric guide circles around the head, denoting the median, frontal, and horizontal planes. The cross-hair was coupled to the head direction. A starting location marker for each trial was drawn in VR as a circle with 2° radius as seen from the central viewpoint at (0, 0) degrees. The stimulus sound was pulsed pink noise with a total duration of 3 s, a pulse length of 400 ms with pauses of 100 ms. The pulsed noise was used to provide onset and offset cues for auditory localization [36]. For each trial, the pulse train was convolved with HRTFs from either the measured or modeled HRTF database, matching the participant's own HRTFs, or using the generic HRTFs. Room acoustic cues were not modeled in this localization experiment, to focus only on the localization cues provided by the HRTFs. Target locations were selected as follows for the azimuth angles: $\phi = \{0, 15, 75, 90, 105, 165, 180\}^\circ$; and elevation angles: $\theta = \{-40, -20, 0, 20, 40\}^\circ$. All combinations of the target azimuths and elevations were presented to the participants once in random order with random assignment to either the left or right hemisphere, assuming a left–right symmetry in auditory localization. The HRTFs were randomly interleaved from the three databases, i.e., the participant was unaware of the HRTF condition in each trial. The audio was played back at 68 dB_A sound pressure level through an RME Fireface UCX audio interface and Beyerdynamic DT770 Pro headphones, diffuse-field equalized by the manufacturer.

4.2. Procedure

The starting location marker turned from red to green when the participant's head orientation was within the circle to indicate the next stimulus could be started. The participant was required to press a button on a hand-held controller to start the stimulus playback and keep their head static during the 3 s stimulus. Once the sound stopped, they should rotate their head orientation, thus the aiming cross-hair towards the perceived location of the sound event. Dynamic audio rendering was disabled during the stimulus presentation to ensure only static cues were available. Another button press by the participant was used to signal the perceived location of the sound event as indicated by head orientation. At the end of the trial, the participant was required to resume facing the starting marker. Four familiarization trials were included where a visual cue was spatially matching the sound event location. The practice targets were located in the horizontal plane at $\pm 20^\circ$ and $\pm 90^\circ$

azimuth. After the familiarization session, there were 105 trials (7 azimuths × 5 elevations × 3 HRTF sets).

4.3. Results

Figure 3 collects the local azimuth and elevation unsigned errors (panels a to d) and quadrant errors percentage (panel e). The quadrant errors percentages display a tendency for lower errors for the modeled and measured HRTFs compared to the generic HRTFs, but the effect is not significant (Friedman test: $\chi^2_{(2,N=8)} = 4.96$, $p = 0.084$, Kendall's $W = 0.31$ (moderate effect size)). The local azimuth and elevation error angles are calculated after the quadrant errors (>90°) are filtered from the data. The azimuth errors are collected across all elevation angles and elevation errors across all azimuth angles. In the front, for azimuth errors, no large differences can be observed apart from the modeled HRTFs yielding lower errors at the $\phi = 15°$ angle compared to the two other HRTFs. On the side, the measured HRTFs result in lower azimuth errors than the modeled or generic HRTFs, while at the back orientations, no differences are observed. The average azimuth angular errors collated across all azimuth angles are observed to be 17° (Measured), 21° (Modeled), and 20° (Generic), with a non-significant main effect of the HRTF set (Friedman test: $\chi^2_{(2,N=8)} = 4.00$, $p = 0.135$, Kendall's $W = 0.25$ (small effect size)).

For elevation errors, negative elevation angles below the horizontal plane tend to yield greater unsigned localization errors than positive elevations. A tendency for the modeled HRTFs to yield larger errors than the other HRTF sets in the horizontal plane and below can be observed. The average elevation angular errors collated across all elevation angles are observed to be 21° (Measured), 26° (Modeled), and 21° (Generic), with a non-significant main effect of the HRTF set (Friedman test: $\chi^2_{(2,N=8)} = 4.00$, $p = 0.135$, Kendall's $W = 0.25$ (small effect size)). Average duration to indicate a target was found after the stimulus playback was 5.9 s (SD = 2.0 s).

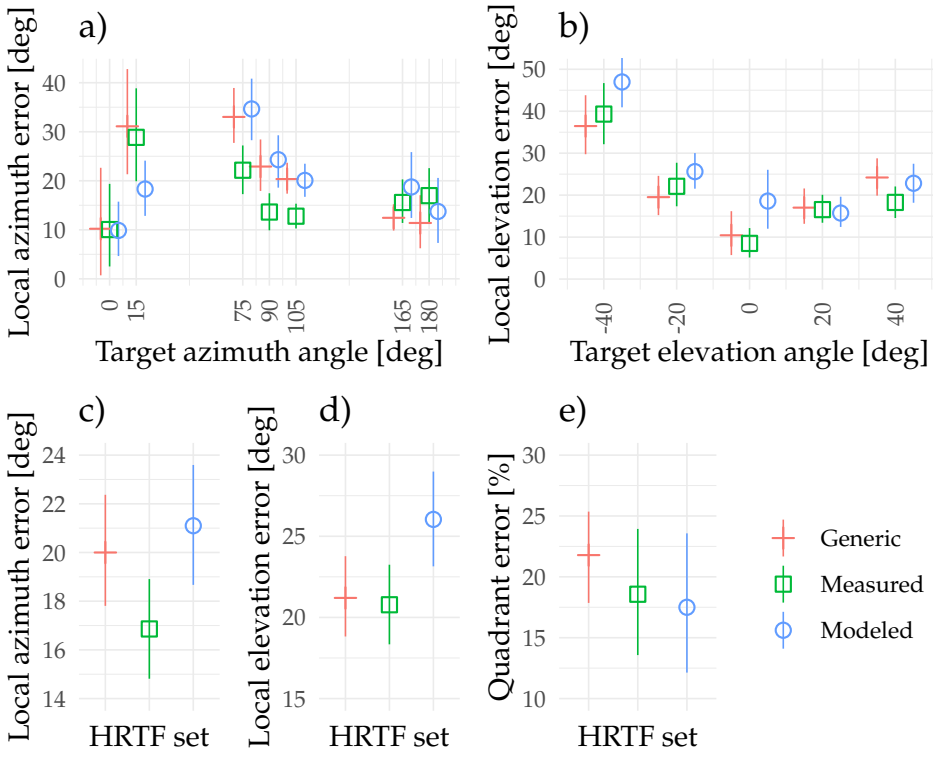

Figure 3. Subjective scores from the static localization experiment for (**a**) local unsigned azimuth error, (**b**) local elevation unsigned error, (**c**) average local azimuth unsigned error, (**d**) average local elevation unsigned error, and (**e**) quadrant errors percentage. The error bars denote the bootstrapped 95% confidence intervals of the mean.

5. Experiment II: Dynamic 6-DoF Virtual Reality

5.1. Virtual Reality Environment

A platform for real-time evaluation of binaural renderers has been developed, allowing participants to switch between conditions with no interruption to audio-visual sensory input. The basic structure is presented in this section; for a thorough walk-through, please see [37]. The platform may be broken into three components: (1) VR device, (2) graphical rendering engine, and (3) audio rendering engine. For Component 1, the HTC Vive Pro HMD and SteamVR Version 2.0 tracking system were used for positional tracking, visual presentation, and control interface. For Component 2, the game engine Unity was used for the graphical rendering, and hosted positional information for all audio objects and participant's position and orientation using the SteamVR asset. All relevant positional and rotational data were then sent (via an Open Sound Control (OSC) data package at a 10 ms interval) to Max 8. In Component 3, binaural renderers, each equipped with a unique HRTF set, were hosted in Max 8 and were fed, in parallel, the positional and rotational information received from Unity.

The interface was designed such that it could be instantiated anywhere in the VR scene. A semi-transparent panel appeared at eye level in the participants' field of view by pressing a button on a hand-held controller. Pressing the button again hid the panel, allowing the user to explore the environment fully. A virtual laser pointer was used to switch between renderers, assign ratings, and move on to the next scene.

5.2. Stimuli and Conditions

Five conditions were created for each of the 6-DoF VR scenes based on the HRTF sets: *Measured*, *Modeled*, *Generic*, *Generic horizontal*, and *No HRTF*. The first three conditions were the same as in Experiment II. The *Generic horizontal* condition was created by retaining only the measurement points on the horizontal plane from the generic HRTF set. Lastly, condition *No HRTF* was created by disabling the binaural rendering and playing back a mono-downmix of the scene audio that did not react to listener movements.

There were four scenes focusing on different aspects of localization accuracy (*In scene, Handheld, Above*, and *Below*), two scenes focusing on timbre (*Cello* and *Castanets*), and two scenes focusing on overall audio quality (*Park* and *Fountain*). The scenes, task in the scene, and audio content are summarized in Table 2. Visually, all of the scenes presented the participants with a large empty space where a white sphere ($r = 10$ cm) denoted the location of the audio object. The environment was purposefully minimalistic in appearance to provide only the necessary cues to support audio-visual sensory integration, and not distract the participants with too much visual information.

The audio object was either placed in the scene (*In scene*) at 1.5 m height from the floor level or attached to the hand-held controller (*Handheld*), allowing the participant to move it around their head freely. In the scene *Above*, the audio object was placed 3 m above the floor level, as opposed to being on the floor in the scene *Below*. In these four scenes the audio content was pink noise. In the two scenes focusing on timbre, the audio object was placed at 1.5 m height, and the participant was instructed to stand on a marker 1 m in front of the audio object. These scenes featured anechoic recordings of a cello or castanets.

The overall quality scenes were composed of two spheres denoting two audio objects placed 3 m from each other at 1.5 m height. In *Park*, the two audio objects were a field recording of ducks and another field recording of shrub pruning with scissor clipping sounds and the rustling of leaves. In the *Fountain* scene, the audio objects were a dry studio recording of piano performance and a field recording of a water fountain.

In all of the scenes, the participant's movement was tracked in 6-DoF, and the audio rendering was updated dynamically to reflect the audio object's position in relation to the listener. The update was performed by choosing the nearest HRTF pair without interpolation. The source signals were convolved with each of the HRTF sets under study in parallel in real-time. To enable accurate comparison of the HRTFs, the acoustics of the virtual space were not modeled. The omitted early reflections and reverberation remove

any supporting localization cues apart from the directional information in the actual HRTFs under test. Due to the lack of virtual acoustics, the direct-to-reverberant ratio cue was not modeled. Thus, distance rendering relied only on the intensity cue realized by applying the inverse square law with a maximum level reached at 0.1 m from the sound object. Auditory near-field effects were not modeled. Similar to the static localization experiment, the audio was played back through an RME Fireface UCX audio interface and Beyerdynamic DT770 Pro headphones.

Table 2. Scenes, task in the scene, and audio content.

Scene	Task	Content
In Scene	Localization accuracy	One audio object: Pink noise
Handheld	Localization accuracy	One audio object: Pink noise
Above	Localization accuracy	One audio object: Pink noise
Below	Localization accuracy	One audio object: Pink noise
Cello	Timbral fidelity	One audio object: Anechoic cello
Castanets	Timbral fidelity	One audio object: Anechoic castanets
Park	Overall audio quality	Two audio objects: Ducks and tree clipping
Fountain	Overall audio quality	Two audio objects: Piano and fountain

5.3. Procedure

The scenes *Above* and *Below* were evaluated on a separate day in a follow-up experiment by a subset of 8 participants. The rest of the scenes were evaluated in one main experiment, one after the other by all of the 12 participants. The presentation order was fixed to start with localization, followed by timbre, and finishing with overall quality. The participant was presented with a multiple stimuli quality scaling task in VR, where the localization accuracy, timbral fidelity, and overall audio quality were to be rated on a 100-point continuous scale, depending on the task. There were verbal labels marking the regions of the scale as bad, poor, fair, good, and excellent in 20-point intervals. The five HRTF conditions were presented side-by-side in the VR user interface, and the participant was allowed to switch between them. The HRTF conditions and interface buttons were randomized for each scene. The interface is displayed together with the sphere denoting audio object location in Figure 4.

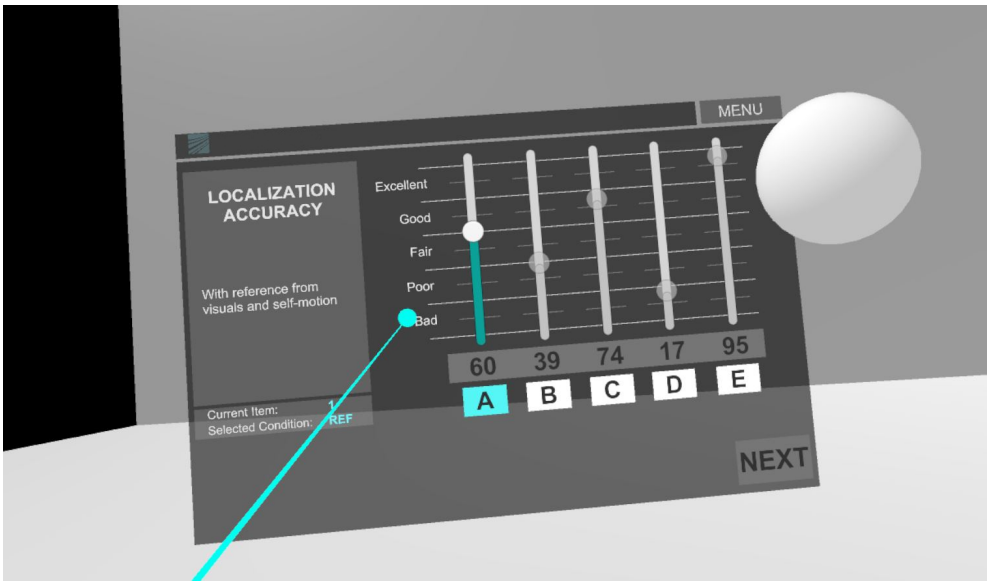

Figure 4. User interface in VR for the localization task. The white sphere denotes visually the location of the audio object.

For the localization and overall quality tasks, the participants were instructed to freely move and teleport around in the 6-DoF scene and focus on either the localizability or general audio aspects of the experience. In these scenes, there was no audio reference to compare against; instead, the participants were assumed to possess an internal reference obtained from the day-to-day experience of listening in the real world. The internal reference should inform the participant how self-movement in relation to visual cues should be reflected in the audio reproduction. In the timbre scenes, the participant was instructed to stand on a marker 1 m away from the audio object and face towards the sound. In these scenes, the *No HRTF* condition was defined as the timbre reference, and all the ratings are relative to this condition. The *No HRTF* was included in the set of conditions as a hidden reference.

5.4. Results

The data from the subjective experiments are analyzed by non-parametric Friedman tests, with the participant defined as a random effect repeated across measurements. The effect sizes are calculated using the Kendall's W metric, which describes the level of agreement among the participants. To interpret the effect size, we refer to the guidelines laid out by Cohen [38]: $0.1 \leq \mathrm{W} < 0.3$ (small effect), $0.3 \leq \mathrm{W} < 0.5$ (moderate effect), and $\mathrm{W} \geq 0.5$ (large effect).

The localization accuracy was found significantly different for the different HRTF sets, when the audio object was placed in the scene, $\chi^2_{(4,N=12)} = 29.6, p < 0.001$, or the audio object was hand-held $\chi^2_{(4,N=12)} = 37.6, p < 0.001$. The corresponding effect sizes are $W = 0.62$ and $W = 0.78$, respectively. Similarly, the localization accuracy was found significantly different for the different HRTF sets, when the audio object was above the listener, $\chi^2_{(4,N=8)} = 24.8, p < 0.001$, or when the audio object was below the listener $\chi^2_{(4,N=8)} = 22.5, p < 0.001$. The corresponding effect sizes are $W = 0.77$ and $W = 0.70$, respectively. Post-hoc pairwise permutation tests revealed no significant differences between the measured, modeled, and generic HRTFs in any of these scenes. These results are displayed graphically in the first four panels a, b, c, and d, in Figure 5.

The timbral fidelity was found significantly different for the different HRTF sets, when listening to an anechoic cello sample, $\chi^2_{(4,N=12)} = 29.8, p < 0.001$, or when listening to an anechoic castanets sample $\chi^2_{(4,N=12)} = 35.4, p < 0.001$. The corresponding effect sizes are $W = 0.62$ and $W = 0.74$, respectively. These results are shown in Figure 5 panels e and f. Post-hoc tests revealed all the other conditions to be rated significantly lower than the hidden reference *No HRTF* in both scenes. In the *Cello* scene, the generic HRTFs were rated significantly higher than the measured or modeled HRTFs. In *Castanets*, the measured HRTFs were rated significantly lower than all the other conditions.

The overall audio quality was found significantly different for the different HRTF sets, when experiencing the park scene, $\chi^2_{(4,N=12)} = 27.3, p < 0.001$, or the fountain music scene $\chi^2_{(4,N=12)} = 29.6, p < 0.001$. The corresponding effect sizes are $W = 0.57$ and $W = 0.62$, respectively. The last panels g and h, in Figure 5 display these results. In *Park*, all of the HRTF sets, even the *Generic horizontal*, were rated equally and above the *No HRTF* condition according to post-hoc tests. In *Fountain*, no difference was found between measured, modeled, and generic HRTFs, but the *Generic horizontal* was significantly lower than *Generic*.

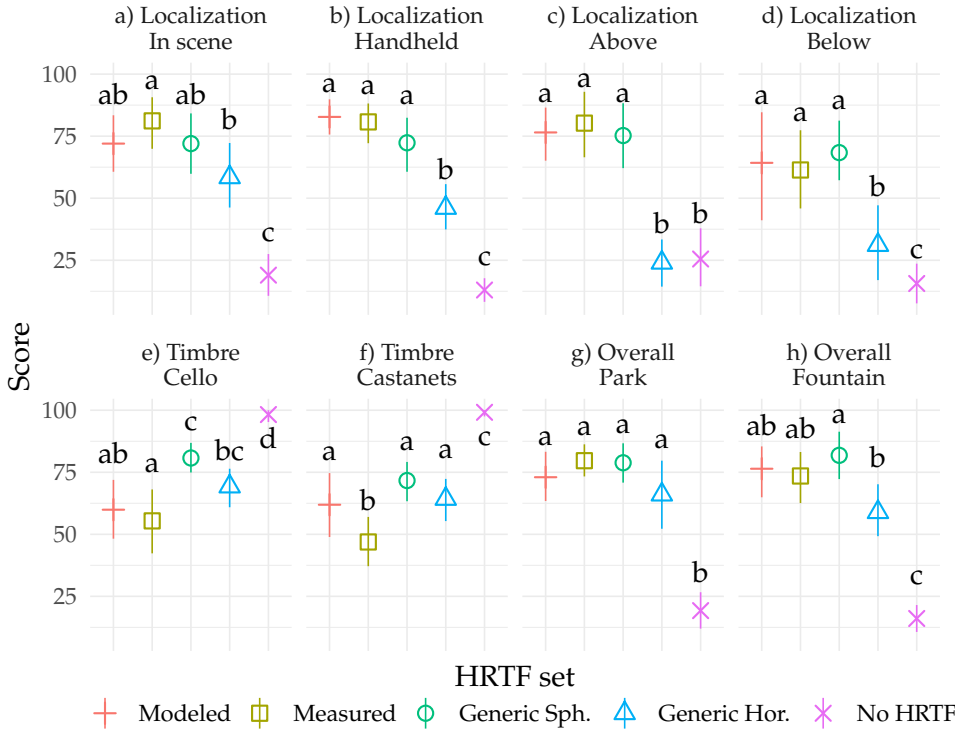

Figure 5. Subjective scores from the 6-DoF VR experiment. Panels (**a**–**d**) display results from the localization scenes, panels (**e**,**f**) timbre scenes, and panels (**g**,**h**) overall quality scenes. The error bars denote the bootstrapped 95% confidence intervals of the mean and the letters denote the significant differences ($p \leq 0.05$) between conditions calculated by pairwise permutation tests. Conditions with the same letter are not significantly different from each other.

6. Discussion

6.1. Objective Deviations

From the analysis in Section 2, the generic HRTFs were found statistically different from either of the individual sets based on the auditory model predictions, which supports our first hypothesis. However, the measured and modeled HRTF sets were found to be objectively significantly different also from each other. If the spectral cues in measured and modeled HRTF sets were similar, we would assume the elevation error predictions, in Figure 2, for the measured template and modeled target HRTF case to be close to the matching template and target errors, as shown in [34] for the own versus other HRTF cases. This, however, is not the case; the local elevation error and quadrant errors for this mixed prediction are closer to the cases with generic HRTFs. Similarly, the ITD and ILD analysis in Figure 1 showed significant differences between all three sets. Therefore, we reject our second hypothesis about the similarity of these sets. Recent work has found corresponding results regarding deviations between simulated and measured HRTFs, where significant differences were found in perceived coloration and localization between the sets [39]. There, Brinkmann et al. [39] suspected the absorbing effect of hair and clothing, missing from the modeled HRTFs, to at least partially explain the differences. However, the absorbing effect does not explain the relatively large deviation found in ITD values at lateral positions in our analysis. Here, the anthropometric measurement process may be a source of error.

6.2. Static Localization

In Experiment I, no clear improvement was found in static localization accuracy when using individual HRTFs. Our third hypothesis receives partial support from the non-significant ($p = 0.083$), but moderate effect size, of reduced quadrant errors percentage

with individual HRTFs (measured 19%, modeled 18%) compared to the generic HRTFs (22%). In previous literature, using static localization tasks, Middlebrooks [4] showed a reduction in quadrant errors from 21% to 8% when switching from generic to individual HRTFs, Wightman and Kistler [40] a reduction from 11% to 6% when switching from individual HRTFs to free field, and finally Wenzel et al. [2], moving from generic HRTFs to free field conditions, a change from 31% to 19% of front-back confusion percentage. However, some studies show no effect in quadrant errors rate between individual and generic HRTFs, for example, Begault et al. [8] 59% and Bronkhorst [3] 28% for both HRTF types. There, especially the level of participant training in localization tasks has been suspected to mask the differences in HRTF cues. Considering a commercial scenario, most users are likely to be untrained, non-expert listeners, which might further minimize the localization benefits of individual HRTFs. Although not showing a significant reduction in quadrant errors percentage, our results appear to be in line with what may be expected to be the range of quadrant errors in static localization tasks using individual or generic HRTFs.

After cleaning the data from quadrant errors, the azimuth error angles should be approximately the same between HRTFs and free-field listening based on previous research [2,3,40]. For headphone localization tasks using either generic or individual HRTFs, the values found in previous research range from 15° to 22° [4,8]. These values may be contrasted with free field localization results, which yielded azimuth errors of 11° on average in [4]. The results for average local azimuth errors in our static localization experiment varied between 17° and 21° depending on the HRTF set, which is in line with previous results. Similar to previous work, we found no statistically significant difference between the HRTF sets in the average azimuth error.

However, previous studies on static localization comparing individual and generic HRTFs have found significant differences in elevation localization accuracy in favor of the individual HRTFs [3,4], but also a lack of an effect is reported [8], where an average elevation error value of 18° was found for both individual and generic HRTFs. In our static localization experiment, no significant reduction in unsigned elevation errors was observed when switching from generic HRTFs to either of the individual sets, which is in contrast with the majority of previous results. The values we found for average unsigned elevation errors ranged from 21° to 26° for the different HRTF sets. Some exemplary elevation error values found in the literature for static localization are 10° or 21° in the free field [4,9], 11°, 29°, or 37° using individual HRTFs [4,9,41], and 42° using generic HRTFs [4]. Comparing absolute localization accuracy results between studies is challenging due to different stimulus signals, target locations, pointing methods, visual settings, and participant training. Despite the lack of reduction in relative unsigned elevation errors between the HRTF sets, the values found in our study are in line with the previously found error values and more closely match the results found in free field or individual HRTF listening. One potential reason for the lack of significant results here could be an adaptation to the generic HRTFs since the static localization experiment was conducted after the dynamic 6-DoF VR experiment, although there was a lengthy gap between the experiments (12 months).

6.3. Dynamic Results

The lack of significant differences between the individual and generic HRTFs in dynamic localization accuracy and overall audio quality in the 6-DoF VR experiment supports our fourth hypothesis regarding the similarity of the HRTFs in this setting. This finding is corroborated by previous research in dynamic localization, where there are often no differences found between individual and generic HRTFs [5,8].

However, a few recent studies also show the contrary. Ben-Hur et al. [12] utilized a constricted movement range where head rotations were allowed, but the sound playback would stop outside of the range. The audio targets were placed beyond the allowed movement range so that dynamic cues could be incorporated in the localization process, but the localization accuracy could be tested for source directions other than only in front.

Ben-Hur et al. [12] found a significant localization improvement with individual HRTFs over generic HRTFs, which is in contrast to the findings of our study. Their experimental setup did not provide visual cues or motor cues beyond the limited movement range, possibly explaining part of the differences.

Another study using audio-visual VR showed that individual HRTFs were preferred in localizability and realism over a generic HRTF set when a target object was a drone flying overhead along the sagittal plane in a pre-programmed animation [13]. This finding is in contrast to our dynamic experiment results. One reason for the contrasting results could be the dynamic audio object and seated participant employed by Jenny and Reuter [13] in contrast to emphasizing the movement of the listener. A dynamically moving audio-visual object has been found to reduce listener movement in 6-DoF VR audio quality evaluations [42], which might lessen the adaptation to generic HRTF cues, but this assertion will require further studies.

The lack of perceptual differences in dynamic localization accuracy is surprising, keeping in mind the different number of measurement points in the individual sets ($Q = 4608$) compared to the generic set ($Q = 1550$). The difference stems from the elevation resolution since the azimuth resolution was the same for all sets (5°). In the localization scene with the audio object in the scene at 1.5 m height, even the *Generic Horizontal* condition with $Q = 72$ measurement points could not be statistically differentiated from the *Generic* set or the *Modeled* set, further highlighting the strength of visual influence on auditory localization. On another note, in dynamic localization without visual cues, a sparse HRTF set with just $Q = 36$ measurement points covering the whole sphere combined with linear interpolation for intermediate locations was shown to yield comparable localization accuracy with a high-resolution HRTF set with $Q = 612$ [12]. Our experiments did not employ an interpolation strategy, which might reduce possible differences between the HRTF sets with different numbers of measurement points. Using the 100-point localization accuracy scale with intermediate quality labels, the individual and spherical generic HRTF sets were evaluated in the *good* (60–80 points) and *excellent* (80–100 points) localizability range. Although the HRTF sets are to some extent affecting each others' relative scores in the absence of an auditory reference, it may be assumed that the visual and motor cues provided sufficient reference conditions for an absolute judgment of the localization accuracy for the experienced listeners in our experiments. This result adds confidence that the sound delivery chain comprising the diffuse-field equalization of the HRTFs and headphones, can convey a convincing spatial impression.

The lowest static and dynamic localization errors for virtual sounds in recent studies have resulted from experiments conducted with floating headphones, where in-ear headphones are mounted in front of the open ear canal entrance and individual headphone calibration is performed at the beginning of each experiment [9,12], using a procedure originally described in [43]. This procedure is different from the diffuse-field calibration performed in the present experiments and suggested in the previous literature [4,30]. Although yielding accurate localization results in the absence of visuals, the individual calibration procedure is laborious and unfeasible with current headphone technology for the general public. The lack of differences between generic and individual HRTFs in this study in a 6-DoF VR context add evidence to the cross-modal integration and adaptation effects on sound localization, possibly circumventing the need for high fidelity individual spatial cues delivery.

The preference in timbre in favor of the generic set is in agreement with previous research [16], where the effect of higher quality microphones in the generic binaural head was suspected to be one reason for the higher perceived timbre fidelity. In the study at hand, defining the *No HRTF* condition as the timbre reference may cause potential confusion in the participants since any of the HRTF conditions is spatially different from the reference. Moreover, the HRTFs cause timbre coloration by definition, making the comparison to a non-filtered stimulus unfair. It might be argued that the stronger localization cues the HRTF can convey, the more different the timbre will be from a non-spatial stimulus. In informal

discussions with the participants, all of them had understood the goal of comparing the non-spatial timbre with a spatialized version, requiring a certain degree of mental simulation of how a real sound source should sound like if placed in front of the listener. The timbre results should only be viewed together with the localization and overall quality results to attain an accurate picture of the quality of the HRTF sets.

The overall quality results support our findings in localization and timbre: the individual and generic HRTF sets are perceived to be highly similar in quality. Here, the participants were allowed to focus on any aspect of the auditory experience while keeping in mind the boundary conditions set by the visuals and motor actions. The scenes were also more complicated than the previously presented scenes in this experiment, both in the number of audio objects and their content. Still, it may be argued that the employed scenes were far from a scene that might be experienced in the real world. In a real scene or a more realistic virtual scene, there would most likely be multiple competing audio objects active simultaneously and blending into a holistic perception of a soundscape. Then, given a visual scene to match the soundscape, the contribution of any individual audio object's auditory localizability to the overall experience would most likely be small. The experiments in this study were conducted in anechoic conditions to highlight the impact of HRTF fidelity on localizability and overall audio quality, representing the worst-case scenario for virtual audio. Based on previous literature, we may assume that adding early reflections and reverberation would further reduce the prominence of differences between individual and generic HRTF sets since these cues affect the spatial impression and externalization [8,44]. Furthermore, more realistic acoustic conditions would ease the integration of visual and auditory environments by reducing the room divergence effect [45], potentially reinforcing the adaptation to generic localization cues.

In summary, our results showed significant objective differences between three HRTF sets, but minimal perceptual differences in dynamic 6-DoF VR scenes. In future, instead of focusing on individualizing the HRTFs, we see potential in systematically adapting the user to listen with generic HRTFs and to listen to spatial audio in general. Recent evidence shows using audio-visual-motor cross-modal training to further improve the adaptation effectiveness compared to audio-visual training [26]. Such a scenario is applicable in virtual reality, where the hands are tracked in addition to the head.

7. Conclusions

This study compared measured and modeled individual HRTF sets with a generic HRTF set and found significant objective differences based on auditory model predictions between all three sets. A static localization experiment showed minor differences in favor of the individual HRTFs over the generic HRTFs in quadrant errors percentage, but a localization and overall quality experiment in 6-DoF audio-visual VR revealed no difference between the HRTF sets. The experimental results suggest there is no need to employ individual HRTFs when applied in a multimodal and dynamic environment.

Author Contributions: Conceptualization, all authors; methodology, O.S.R.; software, O.S.R. and T.R.; validation, all authors; formal analysis, O.S.R.; investigation, O.S.R. and T.R.; resources, all authors; data curation, O.S.R.; writing—original draft preparation, O.S.R.; writing—review and editing, all authors; visualization, O.S.R.; supervision, E.A.P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review was not sought for the present study due to the common research ethics standards of the relevant scientific societies as well as survey research. The experiments do not pose any particular risks, i.e., participation in the study does not produce harm or discomfort beyond everyday experience.

Informed Consent Statement: Informed consent was obtained from all participants prior to the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Møller, H.; Sørensen, M.F.; Hammershøi, D.; Jensen, C.B. Head related transfer functions of human subjects. *J. Audio Eng. Soc.* **1995**, *43*, 300–321.
2. Wenzel, E.M.; Arruda, M.; Kistler, D.J.; Wightman, F.L. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.* **1993**, *94*, 111–123. [CrossRef]
3. Bronkhorst, A.W. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.* **1995**, *98*, 2542–2553. [CrossRef]
4. Middlebrooks, J.C. Virtual localization improved by scaling non-individualized external-ear transfer functions in frequency. *J. Acoust. Soc. Am.* **1999**, *106*, 1493–1510. [CrossRef] [PubMed]
5. Oberem, J.; Richter, J.G.; Setzer, D.; Seibold, J.; Koch, I.; Fels, J. Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods. *bioRxiv* **2020**, 1–11. [CrossRef]
6. Best, V.; Baumgartner, R.; Lavandier, M.; Majdak, P.; Kopčo, N. Sound externalization: A review of recent research. *Trends Hear.* **2020**, *24*. [CrossRef]
7. Wallach, H. On sound localization. *J. Acoust. Soc. Am.* **1939**, *10*, 270–274. [CrossRef]
8. Begault, D.R.; Wenzel, E.M.; Anderson, M.R. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.* **2001**, *49*, 904–916. [PubMed]
9. Romigh, G.D.; Brungart, D.S.; Simpson, B.D. Free-field localization performance with a head-tracked virtual auditory display. *IEEE J. Sel. Top. Signal Process.* **2015**, *9*, 943–954. [CrossRef]
10. McAnally, K.I.; Martin, R.L. Sound localization with head movement: Implications for 3-d audio displays. *Front. Neurosci.* **2014**, *8*, 1–6. [CrossRef]
11. Rummukainen, O.S.; Schlecht, S.J.; Habets, E.A.P. Self-translation induced minimum audible angle. *J. Acoust. Soc. Am.* **2018**, *144*, EL340–EL345. [CrossRef]
12. Ben-Hur, Z.; Alon, D.L.; Robinson, P.W.; Mehra, R. Localization of virtual sounds in dynamic listening using sparse HRTFs. In Proceedings of the Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality, Online, 13 August 2020; pp. 1–9.
13. Jenny, C.; Reuter, C. Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization. *JMIR Serious Games* **2020**, *8*, e17576. [CrossRef]
14. Rummukainen, O.S.; Robotham, T.; Plinge, A.; Wefers, F.; Herre, J.; Habets, E.A.P. Listening tests with individual versus generic head-related transfer functions in six-degrees-of-freedom virtual reality. In Proceedings of the 5th International Conference on Spatial Audio (ICSA), Ilmenau, Germany, 26–28 September 2019; pp. 55–62. [CrossRef]
15. Blau, M.; Budnik, A.; Fallahi, M.; Steffens, H.; Ewert, S.D.; van de Par, S. Toward realistic binaural auralizations—Perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario. *Acta Acust.* **2021**, *5*. [CrossRef]
16. Armstrong, C.; Thresh, L.; Murphy, D.; Kearney, G. A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database. *Appl. Sci.* **2018**, *8*, 2029. [CrossRef]
17. Pelzer, R.; Dinakaran, M.; Brinkmann, F.; Lepa, S.; Grosche, P.; Weinzierl, S. Head-related transfer function recommendation based on perceptual similarities and anthropometric features. *J. Acoust. Soc. Am.* **2020**, *148*, 3809–3817. [CrossRef]
18. Spagnol, S. HRTF selection by anthropometric regression for improving horizontal localization accuracy. *IEEE Signal Process. Lett.* **2020**, 1–5. [CrossRef]
19. Sikström, E.; Geronazzo, M.; Kleimola, J.; Avanzini, F.; de Götzen, A.; Serafin, S. Virtual reality exploration with different head-related transfer functions. In Proceedings of the 15th Sound and Music Computing Conference, Limassol, Cyprus, 4–7 July 2018; pp. 85–92. [CrossRef]
20. Poirier-Quinot, D.; Katz, B.F. Assessing the impact of head-related transfer function individualization on task performance: Case of a virtual reality shooter game. *J. Audio Eng. Soc.* **2020**, *68*, 248–260. [CrossRef]
21. Ernst, M.O.; Bülthoff, H.H. Merging the senses into a robust percept. *Trends Cogn. Sci.* **2004**, *8*, 162–169. [CrossRef]
22. Parseihian, G.; Katz, B.F.G. Rapid head-related transfer function adaptation using a virtual auditory environment. *J. Acoust. Soc. Am.* **2012**, *131*, 2948–2957. [CrossRef] [PubMed]
23. Berger, C.C.; Gonzalez-Franco, M.; Tajadura-Jiménez, A.; Florencio, D.; Zhang, Z. Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. *Front. Neurosci.* **2018**, *12*, 1–9. [CrossRef]
24. Stitt, P.; Picinali, L.; Katz, B.F. Auditory accommodation to poorly matched non-individual spectral localization cues through active learning. *Sci. Rep.* **2019**, *9*, 1063. [CrossRef] [PubMed]
25. Steadman, M.A.; Kim, C.; Lestang, J.H.; Goodman, D.F.; Picinali, L. Short-term effects of sound localization training in virtual reality. *Sci. Rep.* **2019**, *9*, 18284. [CrossRef] [PubMed]
26. Valzolgher, C.; Campus, C.; Rabini, G.; Gori, M.; Pavani, F. Updating spatial hearing abilities through multisensory and motor cues. *Cognition* **2020**, *204*, 104409. [CrossRef] [PubMed]
27. Trapeau, R.; Aubrais, V.; Schönwiesner, M. Fast and persistent adaptation to new spectral cues for sound localization suggests a many-to-one mapping mechanism. *J. Acoust. Soc. Am.* **2016**, *140*, 879–890. [CrossRef] [PubMed]

28. Richter, J.G.; Behler, G.; Fels, J. Evaluation of a fast HRTF measurement system. In Proceedings of the Audio Engineering Society 140th Convention, Paris, France, 4–7 June 2016; pp. 1–7.
29. Gumerov, N.A.; O'Donovan, A.E.; Duraiswami, R.; Zotkin, D.N. Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation. *J. Acoust. Soc. Am.* **2010**, *127*, 370–386. [CrossRef]
30. Larcher, V.; Jot, J.M.; Vandernoot, G. Equalization methods in binaural technology. In Proceedings of the Audio Engineering Society 105th Convention, San Francisco, CA, USA, 26–29 September 1998; pp. 1–28.
31. Søndergaard, P.; Majdak, P. The auditory modeling toolbox. In *The Technology of Binaural Listening*; Blauert, J., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 33–56. [CrossRef]
32. Dietz, M.; Ewert, S.D.; Hohmann, V. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Commun.* **2011**, *53*, 592–605. [CrossRef]
33. Klockgether, S.; van de Par, S. Just noticeable differences of spatial cues in echoic and anechoic acoustical environments. *J. Acoust. Soc. Am.* **2016**, *140*, EL352–EL357. [CrossRef]
34. Baumgartner, R.; Majdak, P.; Laback, B. Modeling sound-source localization in sagittal planes for human listeners. *J. Acoust. Soc. Am.* **2014**, *136*, 791–802. [CrossRef]
35. Schinkel-Bielefeld, N.; Lotze, N.; Nagel, F. Audio quality evaluation by experienced and inexperienced listeners. In Proceedings of the Meetings on Acoustics, Montreal, QC, Canada, 2–7 June 2013; Volume 19, pp. 1–8. [CrossRef]
36. Stecker, G.C. Exploiting envelope fluctuations to enhance binaural perception. In Proceedings of the Audio Engineering Society 140th Convention, Paris, France, 4–7 June 2016; pp. 1–7.
37. Robotham, T.; Rummukainen, O.; Habets, E.A.P. Evaluation of binaural renderers in virtual reality environments: Platform and examples. In Proceedings of the Audio Engineering Society 145th Convention, New York, NY, USA, 17–20 October 2018; pp. 1–5.
38. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Routledge: New York, NY, USA, 1988; p. 283. [CrossRef]
39. Brinkmann, F.; Dinakaran, M.; Pelzer, R.; Grosche, P.; Voss, D.; Weinzierl, S. A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses. *J. Audio Eng. Soc.* **2019**, *67*, 705–718. [CrossRef]
40. Wightman, F.L.; Kistler, D.J. Headphone simulation of free-field listening. II: Psychophysical validation. *J. Acoust. Soc. Am.* **1989**, *85*, 868–878. [CrossRef]
41. Majdak, P.; Goupell, M.J.; Laback, B. 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Atten. Percept. Psychophys.* **2010**, *72*, 454–469. [CrossRef] [PubMed]
42. Rummukainen, O.; Wang, J.; Li, Z.; Robotham, T.; Yan, Z.; Li, Z.; Xie, X.; Nagel, F.; Habets, E.A.P. Influence of visual content on the perceived audio quality in virtual reality. In Proceedings of the 145th Audio Engineering Society International Convention, New York, NY, USA, 17–20 October 2018; pp. 1–10.
43. Langendijk, E.H.A.; Bronkhorst, A.W. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *J. Acoust. Soc. Am.* **2000**, *107*, 528–537. [CrossRef] [PubMed]
44. Catic, J.; Santurette, S.; Dau, T. The role of reverberation-related binaural cues in the externalization of speech. *J. Acoust. Soc. Am.* **2015**, *138*, 1154–1167. [CrossRef] [PubMed]
45. Werner, S.; Klein, F.; Mayenfels, T.; Brandenburg, K. A summary on acoustic room divergence and its effect on externalization of auditory events. In Proceedings of the 8th International Conference on Quality of Multimedia Experience, Lisbon, Portugal, 6–8 June 2016; pp. 1–6. [CrossRef]