DOI: 10.61782/fa.2023.0441

forum acusticum 2023

IMPORTANCE OF HRTF PERSONALISATION FOR AUDIO RENDERING IN MUSIC-RELATED VIRTUAL ENVIRONMENTS

Michael Oehler[1*] **Tray Minh Voong**[1] **Marlon Regener**[1]
Maurício do V. M. da Costa[1] **Christoph Reuter**[2]
[1] Music Technology and Digital Musicology Lab, Osnabrück University, Germany
[2] Musicological Institute, University of Vienna, Austria

ABSTRACT

This paper investigates the relevance of numerically simulated head-related transfer functions (HRTFs) for the perceived authenticity, plausibility, localization accuracy, and immersion in music-related virtual environments. For this purpose, personalized HRTFs were created for 46 individuals using a 3D scan of the head and torso. Numerical calculations were performed with the Mesh2HRTF library, using the fast-multipole BEM solver to generate the HRTFs. The subjects had to evaluate jazz and classical pieces played by musicians in a virtual concert hall when using the personalized HRTF, a generic HRTF (KEMAR) and a simplified personalized HRTF. Additionally, the influence of music preference and musical sophistication was measured. The results of mixed factorial repeated measures analyses of variance showed that, compared to the other two HRTFs, the personalized HRTF statistically significantly improves perceived immersion, plausibility, and localization accuracy for the participant group with a high musical sophistication score, contrary to the group with a low musical sophistication score. No effect of musical preference was found. Individuals with a high degree of musical sophistication thus seem to benefit particularly from the personalized HRTFs, regardless of the genre of music and individual music preferences.

Keywords: *Numerical HRTF Simulation, Virtual Concerts, Musical Sophistication, Music Preference.*

*__Corresponding author__: michael.oehler@uos.de.

1. INTRODUCTION

In recent years, the relevance of personalized Head-Related Transfer Functions (HRTFs) for virtual acoustic environments has been studied from many perspectives, and various methods for obtaining personalized HRTFs have been proposed (for reviews see [1, 2]). A special focus is currently on various hybrid approaches that combine machine learning methods with acoustic measurements, 3D scans of the head/torso, images of the ears, scaling through anthropometric parameters etc. However, a central question remains as to how exactly personalized fitting can best be realized without great effort and without special and expensive devices. In this context, numerical simulations of personalized HRTFs based on 3D scans of the head/torso in particular can facilitate the creation of HRTFs with mobile devices (e.g. with integrated LIDAR scanner) and make them more suitable for everyday use.

Moreover, different virtual scenarios, and therefore different applications, may rely differently on certain HRTF aspects, e.g., localization accuracy might be crucial for navigation solutions, whereas sound quality/authenticity may play a more important role for virtual concerts or similar music-related applications. However, most validations of personalized HRTFs focus on spatial localization (see [3, 4]) or specific perceptual dimensions [5–7]. Music perception is often not included in such a setting and even minor methodological details can be decisive for the results in this area (see e.g. [8]). Those studies that focus on music perception often either do not use personalized HRTFs [9] or have a very specific scope, such as audio-visual seating preference in virtual concert halls [10]. Furthermore, for music-related applications, certain user characteristics also seem to play a

role in the relevance of personalized HRTFs, e.g., musical preferences or the degree of musical sophistication [11].

The main objective of this paper is, therefore, to investigate whether personalized HRTFs lead to a better evaluation of the performance in a virtual concert situation compared to generic HRTFs and whether this depends on the degree of musical sophistication and the music preferences of the (virtual) concert audience. This is addressed by conducting perception experiments in a virtual environment to evaluate the performance of the different HRTFs.

Individual HRTFs were numerically simulated for all participants, with which they performed a music perception experiment in a virtual concert hall. In addition to the perceptual evaluation of the concert, music preferences, and musical sophistication were assessed. The latter was measured with the German version of the Gold-MSI (General Factor) [12], whereas music preferences were measured with the STOMP-R [13]. The basic methodology for the numerical simulation of HRTFs and creation of a realistic virtual concert environment was already used in [11].

1.1 Numerical HRTF Simulation

In order to produce the individualized HRTFs, the first step was the acquisition of the 3D scans of the participants' head and torso. This was performed using the POP 3D scanner from Revopoint [1]. The subjects used a nylon hair net, which provided as much detail of their head's shape as possible.

The second step consisted in processing the acquired meshes in Blender. [2] First, a smoothing procedure was performed in the parts of the mesh were unwanted details related to hair and clothes or artifacts related to the scanning were present. Different resolutions were adopted for different regions of the mesh, as to reduce the computational burden needed for simulating the HRTFs. In this procedure, for a given region, the flatter and the greater its distance to the ear canals, the relatively larger its faces could get. As a consequence, the average final resolution for the subjects' ears was about 1.3 mm; the resolution for the rest of the head averaged around 6 mm; and, finally, the upper section of the torso had a 33 mm resolution, on average. As a consequence, this resulted in meshes that typically had around 24 k to 30 k triangular faces.

The numerical calculations were then performed with the open-source library Mesh2HRTF [14], using the fast-multipole BEM solver to generate the final HRTFs. The simulations were run for both ears independently using faces placed at the ear canal as vibrating elements, with the frequency spectrum being uniformly sampled in intervals of 100 Hz and ranging from 0 Hz to 16 kHz. In the end, the results were sampled in 1550 collocated spatial directions distributed at the surface of a sphere with 1.2 m of radius and centered at the participant's head. Finally, the HRIRs obtained were then resampled up to 44.1 kHz.

A simplified version of the participants' heads was also simulated by means of spherical meshes whose diameter matched the distance between the ear canals. These HRTFs were also computed using the Mesh2HRTF library and the exact same procedure to finalize the files as described above. This 'spherical head model' variant [15] was added to investigate the performance of a simple model based on an individual feature (ear distance) that can be easily measured and simulated fast.

The third model chosen for the perceptual experiments was the KEMAR artificial head [16], which did not need to be simulated. We used the diffuse-field equalized version implemented in Spat [17], which runs in Max, to make it as comparable as possible to the simulated HRTFs.

1.2 Apparatus

The virtual concert hall was implemented using Unity, Max and Spat. Unity and Max were synchronized via OSC, as to send the information regarding the orientation of the participant's head in real time. The audio signal was played back via Sennheiser IE900 in-ear headphones, which were plugged into a Focusrite Scarlett 2i2 audio interface. The headphones level was previously calibrated with the miniDSP EARS headphone test fixture [3] as to provide an equivalent of 85 dB_{SPL} loudness. Considering the smoothness of the frequency response of the headphones used and the fact that it produces the sounds directly into the ear canals, the headphone-ear transfer functions (HpTFs) were not accounted for. An HTC VIVE Pro Eye and two VIVE Controllers were used.

[1] Available from `http://www.revopoint3d.com` (last viewed: Jun. 24, 2023)

[2] Available from `http://www.blender.org` (last viewed: Jun. 24, 2023)

[3] For more information, see `http://www.minidsp.com` (last viewed: Jun. 24, 2023).

1.3 Virtual Concert Hall

The virtual concert hall designed for this study is a recreation of the small broadcast studio of the WDR Broadcast Studios, in Cologne, Germany. The virtual scenario was created in Unity resembling the same dimensions, materials, and spatial configuration of the studio.

Acoustic measurements of the room (available in [18]) were used to faithfully recreate the acoustic environment. To this end, it was used the set of directional room impulse responses (DRIR) measured using a rigid sphere ('VariSphear microphone array'), with a diameter of 17.5 cm and 110 points in the Lebedev grid. In practice, only a selected subset of the responses recorded using the SonicBall loudspeaker placed at the center of the stage [18] was used, comprising 21 different directions distributed in the upper part of the sphere, namely: 12 equally spaced directions in the horizontal plane, eight different directions with elevation $\phi = 45°$, and the DRIR measured for position $\phi = 90°$.

In order to have a lightweight simulation, the binaural reverberation was simulated independently from the dry signal of the instruments [11]. The initial transient portion of the DRIRs, which is related to the direct acoustic path, was suppressed, and then all DRIRs were shifted back in time to align the average time location of such initial transients in $t = 0$. This resulted in DRIRs where only the actual room response is present and that could be combined (summed) with the direct path processed in parallel, presenting no time delay, hence being considerably time aligned with the original direct path, now absent in the DRIRs. A mixed version of the dry (anechoic) signals $\sum_j s_j(t)$ is convolved with each DRIR(θ_i, ϕ_i), producing the reverberated sound $r_i(t)$. The Spat [17] plugin thus provides the spatialization of the dry signals s_j coming from the specific direction of each musician on the stage, while the reverberated sound signals $r_1, r_2, ..., r_{21}$ come from a set of virtual speakers in fixed positions around the listener, with the speakers placed in the same azimuth θ_i and elevation ϕ_i in which the DRIRs were recorded.

The reverberation of the instruments was simulated altogether, since the virtual musicians will all be on the stage and, thus, their sounds come roughly from the same direction. As a result, only the differences in the direct path are used to convey the sense of directionality of the instruments. As for the RIR, the directionality attained using the rigid sphere is used to convey the directional aspects of the incoming reflections in the room.

The relatively coarse spatial resolution attained using the low number of sources for the room reverberation only affected the reverberant part of the incoming signal, for the simulation of the dry signal $s(t)$ of each instrument will be performed in parallel. This is desirable since the diffuse nature of the reverberation helps the listener have a sense of envelopment without the need for many sound sources.

The mono mixes of the dry recordings were made as to mimic the relative intensities of each instrument in real life and the convolutions with each DRIR were performed offline, further saving CPU power for the real-time computations. During the experiment, a dedicated computer running Spat in Max received the relative position of the subject's head in real-time and rendered the auditory spatialization for the set of signals just described. Despite the relatively small number of reverberated signals $r(t)$, this procedure results in a convincing scenario with a precise sense of localization combined with a realistic sense of the space of the room.

1.4 Participants

A total of 46 participants took part in the test trial (average age, in years, $M = 23.85$, $SD = 3.286$; 71.7% female, 28.3% male). Each person got a compensation of 20 € for participating and the whole experiment took about 60 min per participant.

1.5 Procedure

During the experiment, participants sat in a virtual concert hall and were asked to rate a two-minute piece each of jazz and classical music from the dataset published in [19]. Those pieces were chosen to ensure better comparability with other studies. Both pieces were performed by 4 virtual musicians on stage, represented by point clouds, as illustrated in Figure 1.

The *Aria* in string quartet instrumentation by Johann Sebastian Bach (BWV 1068 No. 3) was used as the classical music piece, and *Don't Mean a Thing*, written by Duke Ellington, was used as the jazz music piece. The order of the musical pieces was randomized for each participant.

In order to familiarize the participants with the environment, another piece of music of [19] was played before the actual experiment (*Minor Swing 2* by Django Reinhardt). At the same time, the experimental instructions were displayed in the VE. The participants were seated facing toward the stage, but were encouraged to explore the environment by looking around. The listener was positioned in the center of the concert hall, about 4 m from

the stage.

Participants performed the tasks using three different HRTFs: their individual numerically simulated HRTF, the simplified spherical head model, and the HRTF of the KEMAR artificial head. The loudness for the three HRTFs was set to the same maximum short-term loudness according to EBU R 128 [20]. The order of the HRTFs was randomized for each participant.

After each performance, the participants had to complete the Immersive Music Experience Inventory (IMEI) [9] and additionally rate the following three items on a scale from 1 (does not apply at all) to 4 (fully applies): (1) The listening experience is as authentic as a real concert situation; (2) The listening experience is plausible within the virtually represented concert situation; (3) The sound of the musical instruments can be perceived from the places in the concert hall in the same way as they are visually represented. This item selection covers some categories frequently mentioned in the context of perceptual quality in (music-related) VR environments: plausibility, authenticity, immersion and localization accuracy. In [6,7,9,21], for example, corresponding evaluation methods and measurement scales are described.

Figure 1. Virtual concert scene for the jazz piece.

2. RESULTS

While music preference does not appear to have a statistically significant effect on the subjects' perceptual ratings, there are some significant rating differences when considering musical sophistication. Therefore, the four rated categories of plausibility, authenticity, immersion, and localization accuracy were analyzed individually using mixed-factor repeated-measures analyses of variance (ANOVA) with HRTF and genre as within-subjects variables, the dichotomous factor musical sophistication (median-split of Gold-MSI scores) as a between-subjects variable, and the four different perceptual ratings as the dependent variable in each ANOVA. The lower half of the Gold-MSI scores ranged from 39 to 96 ($M = 74.74$, $SD = 16.899$), and the upper half ranged from 97 to 119 ($M = 105.35$, $SD = 5.441$). For all categories, the Mauchly sphericity test for the interaction effect of HRTF and genre was not significant, so sphericity was assumed.

2.1 Immersive Music Experience

No interaction effect could be found between HRTF, genre, and musical sophistication ($F(2,88) = 0.144$, $p = .866$, $\mu_p^2 = .003$), between HRTF and genre ($F(2,88) = 0.006$, $p = .994$, $\mu_p^2 < .001$) and between genre and musical sophistication ($F(1,44) = 1.062$, $p = .308$, $\mu_p^2 = .024$), but there is an interaction effect between HRTF and musical sophistication ($F(2,88) = 5.786$, $p = .004$, $\mu_p^2 = .116$). Therefore, the main effect of genre and the simple effects of HRTF and musical sophistication were considered in the following analysis.

A Wilks-Lambda ANOVA showed no significant effect for the factor genre ($F(1,44) = 2.497$, $p = .121$, $\mu_p^2 = .054$). For the within-subjects variable HRTF, a significant effect was found for the participants with high musical sophistication scores (Wilks-Lambda $F(2,43) = 4.234$, $p = .021$, $\mu_p^2 = .165$), but no significant effect was found for the participants with low musical sophistication scores (Wilks-Lambda $F(2,43) = 1.345$, $p = .271$, $\mu_p^2 = .059$).

A pairwise comparison (LSD) of the judgments of participants with high musical sophistication showed that there was a significant difference ($p = .020$) in the immersion ratings between the individually simulated HRTF and the spherical head model (2.217, 95%-CI[0.364, 4.071]) and a significant difference ($p = .008$) between the individually simulated HRTF and the KEMAR artificial head (2.652, 95%-CI[0.724, 4.581]). There is no statistically significant difference ($p = .617$) between the spherical head model and the KEMAR (0.435, 95%-CI[-1.304, 2.174]). Figure 2 shows the estimated marginal means of the two groups with high and low scores for musical sophistication for the three different HRTFs.

The perceptual ratings of the two groups with high and low musical sophistication scores differed significantly ($p = .003$) only when the individually simulated HRTF was used (4.522, 95%-CI[1.595, 7.449]), but not for the spherical head model ($p = .387$) and not for the KEMAR ($p = .873$). The estimates are shown in Tab. 1.

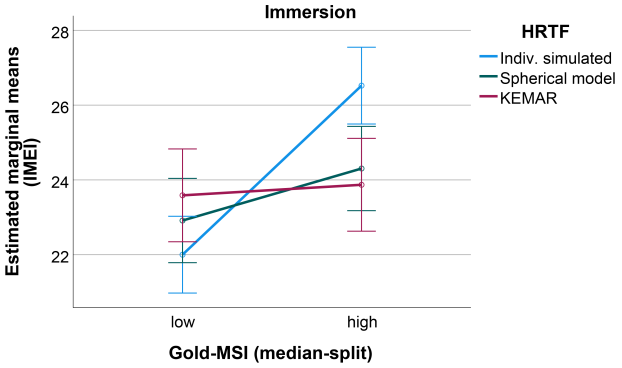

Figure 2. Perceptual ratings for immersion. The error bars indicate 1 SE.

Table 1. Perceptual rating estimates for immersion.

Gold-MSI	HRTF	Mean	SE	95% CI
low	indiv. simulated	22.000	1.027	19.930 - 24.070
low	spherical model	22.913	1.127	20.641 - 25.185
low	KEMAR	23.587	1.241	21.085 - 26.089
high	indiv. simulated	26.522	1.027	24.452 - 28.591
high	spherical model	24.304	1.127	22.033 - 26.576
high	KEMAR	23.870	1.241	21.368 - 26.371

2.2 Plausibility

No interaction effect could be found between HRTF, genre, and musical sophistication ($F(2, 88) = 0.510$, $p = .603$, $\mu_p^2 = .011$), between HRTF and genre ($F(2, 88) = 1.529$, $p = .223$, $\mu_p^2 = .034$) and between genre and musical sophistication ($F(1, 44) = 0.441$, $p = .510$, $\mu_p^2 = .010$), but there is an interaction effect between HRTF and musical sophistication ($F(2, 88) = 3.148$, $p = .048$, $\mu_p^2 = .067$). As a result, the main effect of genre and the simple effects of HRTF and musical sophistication were considered in the following analysis.

A Wilks-Lambda ANOVA showed no significant effect for the factor genre ($F(1, 44) = 0.159$, $p = .692$, $\mu_p^2 = .004$). For the within-subjects variable HRTF, a significant effect was found for the participants with high musical sophistication scores (Wilks-Lambda $F(2, 43) = 4.207$, $p = .021$, $\mu_p^2 = .164$), but no significant effect was found for the participants with low musical sophistication scores (Wilks-Lambda $F(2, 43) = 0.458$, $p = .636$, $\mu_p^2 = .021$).

A pairwise comparison (LSD) of the judgments of participants with high musical sophistication showed that there was a significant difference ($p = .008$) in the plausibility ratings between the individually simulated HRTF and the spherical head model (0.283, 95%-CI[0.077, 0.489]) and a significant difference ($p = .047$) between the individually simulated HRTF and the KEMAR artificial head (0.239, 95%-CI[0.003, 0.475]). There is no statistically significant difference ($p = .717$) between the spherical head model and the KEMAR (0.043, 95%-CI[-0.197, 0.284]). Figure 3 shows the estimated marginal means of the two groups with high and low scores for musical sophistication for the three different HRTFs.

Table 2. Perceptual rating estimates for plausibility.

Gold-MSI	HRTF	Mean	SE	95% CI
low	indiv. simulated	1.457	0.143	1.168 - 1.745
low	spherical model	1.522	0.136	1.247 - 1.797
low	KEMAR	1.565	0.140	1.282 - 1.848
high	indiv. simulated	1.913	0.143	1.624 - 2.202
high	spherical model	1.630	0.136	1.355 - 1.905
high	KEMAR	1.674	0.140	1.391 - 1.957

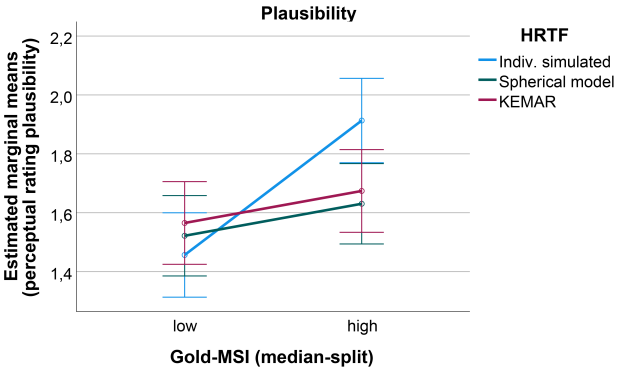

Figure 3. Perceptual ratings for plausibility. The error bars indicate 1 SE.

The perceptual ratings of the two groups with high and low musical sophistication scores differed significantly ($p = .029$) only when the individually simulated HRTF was used (0.457, 95%-CI[0.048, 0.865]), but not for the spherical head model ($p = .576$) and not for the KEMAR ($p = .587$). The estimates are shown in Tab. 2.

2.3 Localization Accuracy

No interaction effect could be found between HRTF, genre, and musical sophistication ($F(2, 88) = 0.102$, $p = .903$, $\mu_p^2 = .002$), between HRTF and genre ($F(2, 88) = 0.190$, $p = .828$, $\mu_p^2 = .004$) and between genre and musical sophistication ($F(1, 44) = 0.157$, $p = .694$,

$\mu_p^2 = .004$), but there is an interaction effect between HRTF and musical sophistication ($F(2, 88) = 4.593$, $p = .013$, $\mu_p^2 = .095$). Thus, the main effect of genre and the simple effects of HRTF and musical sophistication were considered in the following analysis.

A Wilks-Lambda ANOVA showed no significant effect for the factor genre ($F(1, 44) = 1.925$, $p = .172$, $\mu_p^2 = .042$). For the within-subjects variable HRTF, a significant effect was found for the participants with high musical sophistication scores (Wilks-Lambda $F(2, 43) = 7.052$, $p = .002$, $\mu_p^2 = .247$), but no significant effect was found for the participants with low musical sophistication scores (Wilks-Lambda $F(2, 43) = 0.017$, $p = .983$, $\mu_p^2 = .001$).

A pairwise comparison (LSD) of the judgments of participants with high musical sophistication showed that there was a significant difference ($p = .001$) in the localization accuracy ratings between the individually simulated HRTF and the spherical head model (0.522, 95%-CI[0.212, 0.832]) and a significant difference ($p = .001$) between the individually simulated HRTF and the KEMAR artificial head (0.587, 95%-CI[0.250, 0.924]). There is no statistically significant difference ($p = .632$) between the spherical head model and the KEMAR (0.065, 95%-CI[-0.206, 0.337]). Figure 4 shows the estimated marginal means of the two groups with high and low scores for musical sophistication for the three different HRTFs.

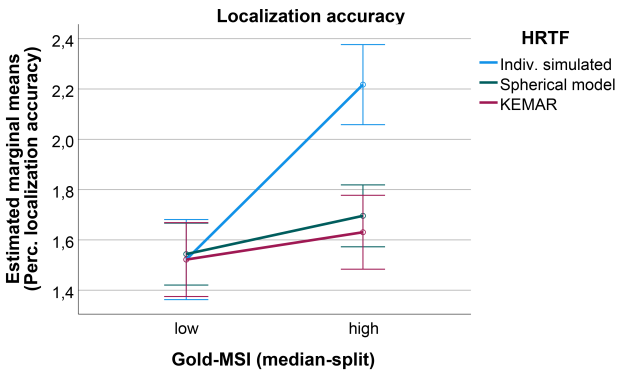

Figure 4. Perceptual ratings for localization accuracy. The error bars indicate 1 SE.

The perceptual ratings of the two groups with high and low musical sophistication scores differed significantly ($p = .003$) only when the individually simulated HRTF was used (0.696, 95%-CI[0.241, 1.149), but not for the spherical head model ($p = .387$) and not for the KEMAR ($p = .604$). The estimates are shown in Table 3.

Table 3. Perceptual rating estimates for localization accuracy.

Gold-MSI	HRTF	Mean	SE	95% CI
low	indiv. simulated	1.522	0.159	1.201 - 1.843
low	spherical model	1.543	0.123	1.295 - 1.791
low	KEMAR	1.522	0.147	1.226 - 1.818
high	indiv. simulated	2.217	0.159	1.897 - 2.538
high	spherical model	1.696	0.123	1.448 - 1.944
high	KEMAR	1.630	0.147	1.334 - 1.927

2.4 Authenticity

No interaction effect could be found between HRTF, genre, and musical sophistication ($F(2, 88) = 0.827$, $p = .441$, $\mu_p^2 = .018$), between HRTF and genre ($F(2, 88) = 0.064$, $p = .938$, $\mu_p^2 = .001$), between genre and musical sophistication ($F(1, 44) = 0.154$, $p = .697$, $\mu_p^2 = .003$), and between HRTF and musical sophistication ($F(2, 88) = 0.453$, $p = .637$, $\mu_p^2 = .010$). As a result, the main effects of the genre, HRTF, and musical sophistication were considered in the following analysis.

A Wilks-Lambda ANOVA showed no significant effect for the within-subjects variable genre ($F(1, 44) = 3.849$, $p = .056$, $\mu_p^2 = .080$) and for the factor HRTF ($F(2, 43) = 0.449$, $p = .641$, $\mu_p^2 = .020$). No significant effect was found for the between-subjects variable musical sophistication either ($F(1, 44) = 0.654$, $p = .423$, $\mu_p^2 = .015$).

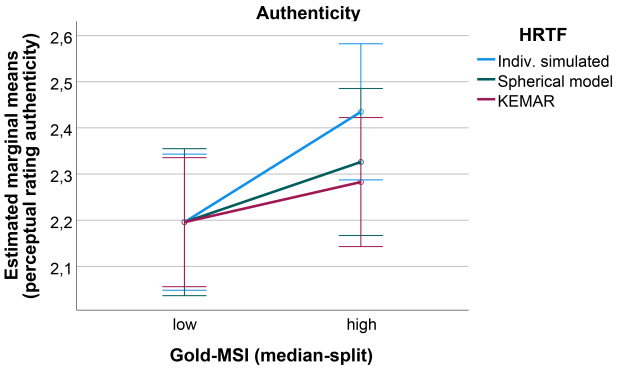

Figure 5. Perceptual ratings for authenticity. The error bars indicate 1 SE.

Although the statistical analysis did not yield significant results for the perceived authenticity, Figure 5 shows

a similar pattern as for plausibility, localization accuracy, and immersive music experience.

3. CONCLUSIONS

We investigated whether personalized HRTFs lead to a better performance evaluation in a virtual concert situation compared to simplified spherical head models and compared to the generic KEMAR HRTF. We also analyzed whether this depends on the degree of musical sophistication and the music preferences of the (virtual) concert audience.

It was shown that the genre of the musical stimuli (jazz or classical) and the individual music preferences did not affect participants' ratings of authenticity, plausibility, localization accuracy, and immersion within the virtual acoustic environment. However, the HRTF used significantly influenced the perceptual ratings for all parameters except authenticity for the group with a high degree of musical sophistication, but not for the group with a low degree of musical sophistication. For the group with a high degree of musical sophistication, the numerically simulated personalized HRTF resulted in the best ratings. However, only the difference between personalized HRTF and spherical head model and between personalized HRTF and KEMAR HRTF were statistically significant.

Individuals with a high degree of musical sophistication thus seem to benefit particularly from personalized HRTFs, regardless of the genre of music and individual music preferences. It is possible that this group perceives subtle details of the musical scene more explicitly, which can be better represented by the individually simulated HRTF. This finding could perhaps be helpful in the design of future music-related VE applications. Especially since the concept of musical sophistication underlying the study [12] takes into account the multi-faceted nature of musical expertise, such as engaging with music in several ways other than playing an instrument. For example, the decision for a specific target group of an application could be decisive for whether personalized HRTFs should be integrated, or whether a generic HRTF or a simplified model might be sufficient.

There are still several limitations that should be addressed in future studies. The relatively high median value of the Gold MSI compared to the norm sample [12] indicates a relatively high overall musical sophistication level in the pool of subjects, which might bias the results. In follow-up studies, there will be an effort to recruit people that provide a more balanced distribution in this respect. Moreover, in addition to the general factor of the Gold-MSI, more attention could be paid specifically to the perceptual abilities of the participants, since these might be more significant than the other subscales in an accordingly complex acoustic environment. The selection of evaluation scales is another influencing factor. While localization accuracy is a relatively commonly used parameter in comparable studies, both authenticity and plausibility are also frequently mentioned parameters, but there are a number of other suitable evaluation criteria, as discussed in [6, 7, 21]. Another aspect to be considered is the computational power required for the simulations of the individual HRTFs, which is still prohibitively high for lightweight devices, such as smartphones. Besides, the 3D scanning procedure is prone to produce artifacts. Although the procedure works without expensive and immobile scanning technology, post-processing and simulation still take a relatively long time, thus, devising automatic procedures to prepare the meshes would make this step more suitable for everyday use.

In addition, the current version of the virtual concert hall experiment seems to produce only relatively small effect sizes. This could be related to the relatively low evaluations overall and the non-significant results for the authenticity parameter. One goal here could be to improve the audiovisual representation of the musical performance, i.e., to design and implement further realistic, possibly 6DoF interactive, performances in VR/MR. At the same time, it would also be useful to improve the lightweight room simulation used in the study and to adapt it to a 6DoF variant.

4. ACKNOWLEDGMENTS

This work was partly funded by Volkswagen Foundation (VolkswagenStiftung) Germany (grant no. 96 881).

5. REFERENCES

[1] C. Guezenoc and R. Seguier, "Hrtf individualization: A survey," in *Proceedings of the 145th Audio Engineering Society International Convention*, (New York, USA), October 2018.

[2] L. Picinali and B. F. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic Interactions in Virtual Environments*, pp. 115–143, Springer International Publishing Cham, 2022.

[3] P. Paukner, M. Rothbucher, and K. Diepold, "Sound localization performance comparison of different hrtf-individualization methods," Technical report 620, Technical University of Munich, Munich, Germany, April 2014.

[4] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[5] R. Nicol, L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, B. F. G. Katz, and L. S. R. Simon, "A roadmap for assessing the quality of experience of 3d audio binaural rendering," in *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, (Berlin, Germany), March 2014.

[6] 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, pp. 984–994, September 2014.

[7] C. Jenny and C. Reuter, "Can i trust my ears in vr? literature review of head-related transfer functions and valuation methods with descriptive attributes in virtual reality," *International Journal of Virtual Reality*, vol. 21, pp. 29–43, October 2021.

[8] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Applied Sciences*, vol. 8, p. 2029, October 2018.

[9] Y. Wycisk, K. Sander, R. Kopiez, F. Platz, S. Preihs, and J. Peissig, "Wrapped into sound: Development of the immersive music experience inventory (imei)," *Frontiers in psychology*, p. 4894, 2022.

[10] Y. Chen, D. Cabrera, and D. Alais, "Modelling audio-visual seat preference in virtual concert halls," *Available at SSRN 4195786*.

[11] M. Oehler, da Costa, M. do V. M., M. Regener, and T. M. Voong, "Relevance of individual numerically simulated head-related transfer functions for different scenarios in virtual environments," in *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*, Audio Engineering Society, 2022.

[12] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The musicality of non-musicians: An index for assessing musical sophistication in the general population," *PloS One*, vol. 9, p. e89642, February 2014.

[13] P. J. Rentfrow and S. D. Gosling, "The do-re-mi's of everyday life: The structure and personality correlates of music preferences," *Journal of Personality and Social Psychology*, vol. 84, no. 6, pp. 1236–1256, 2003.

[14] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2hrtf: Open-source software package for the numerical calculation of head-related transfer functions," in *Proceedings of the 22nd International Congress on Sound and Vibration*, (Florence, Italy), July 2015.

[15] S. Bögelein, F. Brinkmann, D. Ackermann, and S. Weinzierl, "Localization cues of a spherical head model," in *In Fortschritte der Akustik - DAGA 2018 : 44. Jahrestagung für Akustik*, (Munich), March 2018.

[16] W. G. Gardner and K. D. Martin, "Hrtf measurements of a kemar," *The Journal of the Acoustical Society of America*, vol. 97, pp. 3907–3908, June 1995.

[17] T. Carpentier, "A new implementation of spat in max," in *Proc. of the 15th Sound Music Computing Conference (SMC)*, (Limassol, Cyprus), July 2018.

[18] P. Stade, B. Bernschütz, and M. Rühl, "A spatial audio impulse response compilation captured at the wdr broadcast studios," in *Proceedings of the 27th Tonmeistertagung - VDT International Convention*, (Cologne, Germany), November 2012.

[19] D. Thery and B. F. G. Katz, "Anechoic audio and 3d-video content database of small ensemble performances for virtual concerts," in *Proceedings of the 23rd International Congress on Acoustics*, (Aachen, Germany), September 2019.

[20] EBU, "Technical recommendation r128: Loudness normalisation and permitted maximum level of audio signals," 2020.

[21] F. Brinkmann and S. Weinzierl, "Audio quality assessment for virtual reality," in *Sonic Interactions in Virtual Environments*, pp. 145–178, Springer International Publishing Cham, 2022.