Impact of HRTF individualization on player performance in a VR shooter game II

David Poirier-Quinot, Brian F. G. Katz

Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio for Virtual and Augmented Reality
2018 August 20 – 22, Redmond, WA, USA

Impact of HRTF individualization on player performance in a VR shooter game II

David Poirier-Quinot and Brian F.G. Katz

Sorbonne Université, CNRS, Institut Jean Le Rond d'Alembert, UMR 7190, F-75005 Paris, France

Correspondence should be addressed to David Poirier-Quinot
(david.poirier-quinot@sorbonne-universite.fr)

ABSTRACT

We present the extended results of a previous experiment [1] to assess the impact of individualized binaural rendering on player performance in the context of a VR "shooter game", as part of a larger project to characterize the impact of binaural rendering quality in various VR applications. Participants played a game in which they were faced with successive enemy targets approaching from random directions on a sphere. Audio-visual cues allowed for target localization. Participants were equipped with an Oculus CV1-HMD, headphones, and two Oculus Touch hand tracked devices as targeting mechanisms. Participants performed six sessions alternatively using their best and worst-match HRTFs from a "perceptually orthogonal" optimized set of 7 HRTFs [2]. Results suggest that the impact of the HRTF on participant performance (speed and movement efficiency) depends both on participant sensitivity and HRTF presentation order.

1 Introduction

"Binaural hearing" refers to the capability of integrating information from the two ears to perceive a sound in three-dimensional space (azimuth, elevation, and distance). Psychophysical studies have shown that various mechanisms are involved in the human auditory system for sound localization [3]. To infer the angular direction of a sound source, these mechanisms rely on direction-dependent audio cues, resulting from the propagation of an acoustic wave from the source to both ears. Using digital signal processing, these cues can be applied to any audio input to simulate a sound object at a virtual position in a listener's 3D auditory space (experienced over headphones). The set of these direction-dependent cues for a given person is typically referred to as a Head Related Transfer Function (HRTF).

HRTFs are individual, directly resulting from the interactions between a person's morphology and an impinging acoustic wave during its propagation around the head [3]. Individuals listening to a binaural audio scene rendered using their own HRTF will perceive each of its components with more spatial precision than those presented with a random HRTF set [4]. Various methods have been proposed to select a "best-match" HRTF from an existing database [5, 6], as measuring an in-

dividual's actual HRTF is a demanding operation [7]. This process of selection and use of a best-match HRTF for binaural synthesis is here referred to as HRTF individualization.

This study is part of a larger research project aiming to characterize the impact of binaural rendering quality in the context of different Virtual Reality (VR) application contexts. This study focuses on the impact of HRTF individualization on performance in the context of a VR shooter game. While the core of the gameplay is built around an audio-visual localization task, it extends the existing literature [8, 9, 10, 11] in that the final experience is truly a game, where participants are placed under increasingly difficult time and performance constraints.

The underlying hypothesis of this study is that using individualized HRTFs will result in an increase in participant performance (reaction time, movement efficiency), more so as the overall game dynamic (enemy spawn interval, flight speed, etc.) increases.

2 Experimental design

The experiment consisted of two sequential parts. The objective of Part 1 was to identify the *best* and *worst* match HRTFs from a subset of 7 for each participant. Part 2 was the VR shooter game. A total of 20 participants undertook the experiment (8 female, mean age 31.8 ± 11.5 years).

2.1 HRTF classification

The HRTF subset database was assembled from the LISTEN database, defined from a "perceptually orthogonal" optimized HRTF collection [2]. Per-participant best and worst match HRTF sets were selected based on the method elaborated in [12], establishing a classification based on perceptual-space distance between a spatialized audio trajectory and a described reference. Two trajectories were presented: horizontal plane (12 angles [0°:30°:330°]) and median plane (19 angles [–45°:15°:225°]), as illustrated in Figure 1.

Each audio trajectory was generated with the 7 HRTFs from the subset. Participants were instructed to rate the 7 resulting versions of each trajectory on a fixed 9-point scale. They were encouraged to distribute their notations on that scale, and required to indicate at least one best (9) and one worst (1) match. Both median and horizontal rating sessions were repeated 3 times, to take into account HRTF rating consistency [13]. Participants completed the study in a listening booth, ambient noise level < 30 dBA using Sennheiser HD600 headphones and RME Fireface UC audio interface.

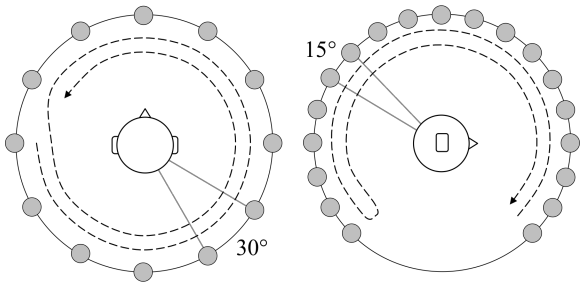

Fig. 1: Trajectory descriptions for HRTF quality ratings: horizontal (left) and median (right) plane trajectories indicating the start/stop position and trajectory direction ($- \rightarrow$).

2.2 VR Shooter Game

During the VR shooter game, participants were equipped with an Oculus CV1 Head Mounted Display (HMD), a pair of headphones (those of the CV1), and a pair of hand tracked devices (Oculus Touch controllers). The game started with participants immersed in a virtual scene, standing on a 0.5 m radius platform mounted on a pole at the center of a 20 m radius spherical structure. Enemy targets could "spawn" from any of the 29 evenly distributed holes in the structure, flying in straight lines towards the participant until collision, either with a bullet or the participant. Participants were instructed to shoot at the incoming targets using a pair of hand-held blasters, the avatar representations of the hand tracked devices in the virtual scene, destroying as many as possible, as fast as possible, in the given time limit.

Enemy targets emitted specific event-based sounds for: spawning, launching, flight, and collision. All sounds were spatialized using the Anaglyph binaural audio engine v0.9 [14]. Anechoic conditions were employed, no room effect was included to keep the study's focus on HRTF effects.

The game was designed using the Unity v2017.3.0 game engine with modelled assets designed in Blender v2.79. The OSC (Open Sound Control) protocol [15] was used for communications between Unity and the

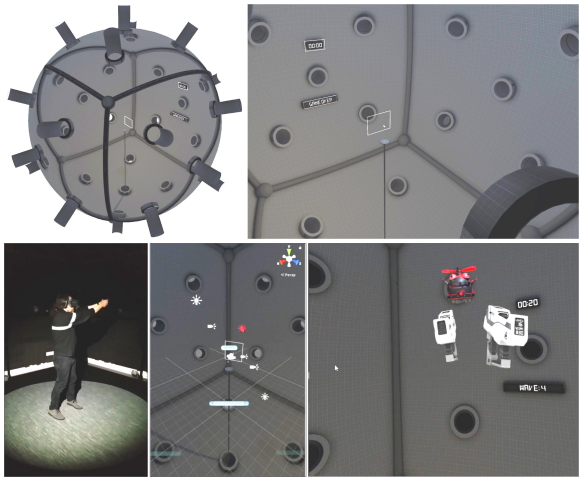

Fig. 2: Game scene overview: (upper-left) overall view of the virtual environment, (upper-right) focus on the platform atop which participants stand during the game, (lower-left) participant in the VR room, (lower-middle) virtual environment during gameplay, and (lower-right) in-game screenshot.

Anaglyph engine running as a VST in Cycling'74 Max v7.3. Figure 2 depicts the game setup and the VR scene. A video extract showing a game session is available[1,2].

A short training session introduced the controls and the difficulty level mechanism, implemented so that the overall game dynamic (enemy spawn interval, flight speed, etc.) increased as the game progressed and participant's skill (in-game "level" in Table 1) improved. The 5 min game was played in six sessions, alternating between participant's best and worst match HRTF. Best/worst match presentation order was evenly balanced among participants, resulting in two groups. To avoid fatigue on this rather demanding task (see video extract), participants played sessions S1–2, directly following the HRTF classification task, followed by a week pause interval, then sessions S3–6.

3 Results

Preliminary elements of the results were previously presented [1]. That previous study was based on the current experimental design and protocol, involving 30 participants for just the first two game sessions. The present study extends these results to a six session game so as to investigate the impact of HRTF individualization on the "long" term, to dissociate it from the game learning effect that may have affected the results of the first part of this study. All of the 20 participants involved in this study took part in the first experiment, each undertaking four more sessions approximately a week after they took the first two.

Performance assessment was based on three metrics, calculated from sessions logs: in-game level, spawn-spot reaction time and spawn-spot travelled angular distance. The in-game level was related to the number of enemies destroyed versus those that hit participants: increasing one unit for every three consecutive kills, decreasing one unit for every two consecutive fails.

Participant spawn-spot reaction time corresponded to the time interval between the spawn of a target and its entering the visual field of view, defined as a 50° cone centered around the current forward view axis. The event of seeing the target, rather than destroying it, was chosen so as to remove the impact of skill at aiming and destroying targets from the analysis (the task of *targeting* being independent of the acoustic rendering quality). Targets visible upon spawn (spawned in the current field of view) were discarded from spawn-spot reaction time analysis. For targets that never entered participants field of view, colliding with or shot by the participants, the spawn-to-collision time was used as the spawn-spot reaction time. Such targets, shot without ever being visible, represented less than 3% of the total number of target spawned.

The associated spawn-spot travelled angular distance metric corresponds to the angular distance traversed by the participant's head from target spawn to target spot sub-events. This last metric represented movement efficiency, and served to differentiate between participants using binaural cues to localize targets and those randomly looking around [16].

Table 1 summarizes the independent and dependent variables of the experimental protocol. Result significance was assessed using a Wilcoxon signed rank test (p-value threshold of 0.05) as all compared paired-sample distributions proved to follow a non-normal (skewed) distribution.

[1] https://youtu.be/q6muds1qW-w

[2] https://www.youtube.com/c/LAMSorbonneUniversité

Table 1: Independent and dependent variables of the experimental protocol.

Independent variables		
Participant ID	20	random variable
HRTF ID	7	best, worst
Session ID	6	S1, S2, ..., S6
Spawn region	6	R1, R2, ..., R6
Dependent variables		
angular distance	event-wise	raw and norm
time	event-wise	raw and norm
mean level	session-wise	raw and norm

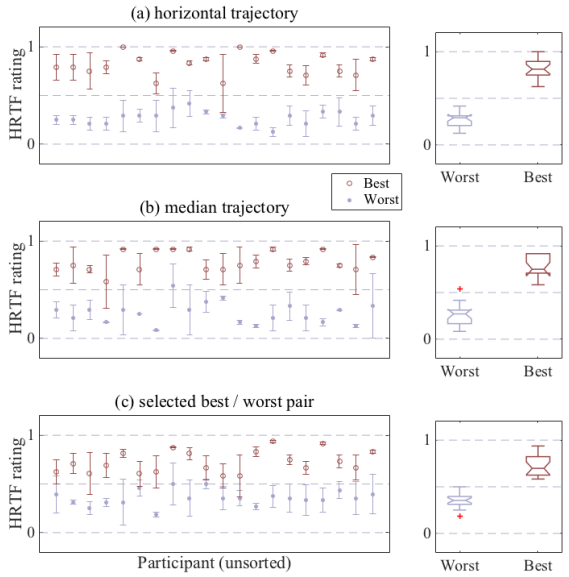

Fig. 3: Results of the HRTF classification task for all participants. The reported rating values correspond to the average normalized rank given by participants to their *determined* best/worst HRTF. A value of 0 (resp. 1) indicates that the HRTF was always rated as the least (resp. most) representative of the described trajectory across the 3 rating repetitions. Scores for the best and worst HRTF matches for (a) horizontal and (b) median trajectories. (c) Combined trajectory mean score results for the *determined* best and worst match HRTF. Error bars indicate the variance of participant ratings for best and worst HRTF.

3.1 HRTF Classification Results

Results of the HRTF ratings of Part 1 are summarized in Figure 3, focusing on the scores obtained by each participant's best and worst HRTF match for both trajectories. Participants were consistent in their classification with regards to these extrema. As audio sources in Part 2 of the experiment were to arrive from all directions, an average best- and worst-HRTF match across trajectories was established for each participant. The rating statistics for selected best- and worst-HRTF matches are shown in Fig. 3c. Most participants proved consistent in their ratings to clearly distinguish between best and worst match for both horizontal and median trajectories. As also shown in [12], participant HRTF ratings for the horizontal and the median trajectories were not correlated. This explains the observed decrease in rating values for the trajectory mean results (Fig. 3c) as compared to the individual horizontal and median plane trajectory ratings (resp. Fig. 3a and Fig. 3b). For almost all participants, average-best and average-worst HRTF scores remained sufficiently distinct to distinguish both populations.

3.2 VR Shooter Game Results

Result analysis was subdivided into session wise and event wise analysis. Session wise analysis concerns participant mean results across sessions. Event wise analysis decomposes each session in events, an event being defined as an enemy target {spawn, launch, flight, collision} sequence. Events related to targets spawned in participant's field of view (see Section 3) have been discarded from analysis. Events where the target was not destroyed and ended up colliding with the participant are included in the analysis, adding both total event time and angular distance to participant's results.

Out of 20 participants, 10 started S1 with their worst-match HRTF (group 1), 10 with their best-match HRTF (group 2). The analysis below concerns $20 \times 6 = 120$ sessions for a total of 14430 events (average of 120.3 ± 18.3 events per session, since game dynamics vary with participant in-game level).

3.2.1 Statistical analysis across participants

Figure 4 illustrates the evolution of participants in-game level across sessions. Mean in-game level significantly increased across sessions, until it reached a

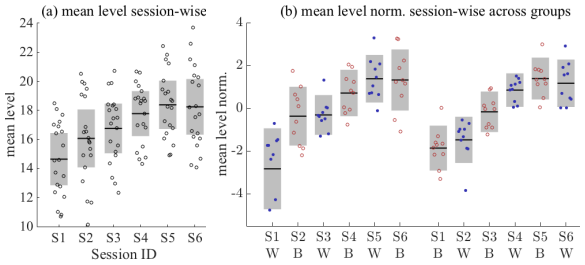

Fig. 4: Mean value, confidence interval (95%), and scattered representation of (a) participant mean game level across sessions, and (b) participant mean game level normalized – per-participant normalization – across sessions for both groups of participants. *W* and *B* indicates worst and best HRTF respectively.

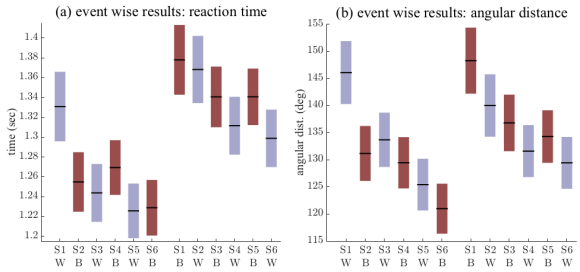

Fig. 5: Agregated (a) event-wise reaction time and (b) normalized angular distance (mean and 95% CI) across sessions and HRTF for both groups of participants.

plateau from S4 on (from 14.6 for S1 to 18.1 for S4–6). Mean spawn-spot angular distance traversed and response time (both averaged per participant session) likewise significantly diminished from S1 to S4 (from 147.7° and 1.37 s for S1, to 125.0° and 1.27 s for S4–5–6 resp.). These results highlight the game task learning effect independent of the HRTF condition.

Repeating the same analysis as a function of HRTF across both participant groups (Figure 4b) showed no significant impact of the HRTF on these metrics. These results would suggest at first glance that there was no benefit to using individual HRTFs in the VR shooter game.

3.2.2 Statistical analysis by event

Combining the results of each event for all participants, the event-wise analysis also reflects the significant impact of game learning on performance, for spawn-spot angular distance traversed (from 147.2° for S1 to 125.1° for S6) and reaction time (from 1.35 s for S1 to 1.26 s for S6). As for mean session results in Section 3.2.1, event-wise statistics show no significant impact of the HRTF quality on either reaction time or angular distance traversed. This result differs from those reported in [1], limited to the first two sessions of the game. HRTF quality still has a significant impact on both metrics between the first two sessions for these 20 participants, yet none for all sections combined, nor for any S3 to S6 combination.

3.2.3 HRTF×Session interaction analysis

Subsequent analysis examines the interaction effect of HRTF presentation order across sessions. The event-wise mean and 95% confidence intervals for reaction time and angular distance traversed are shown in Figure 5. As seen in [1, Figure 5], a clear interaction effect can be observed on event-wise reaction time between S1 and S2, only for the group of participants going from worst to best HRTF. This suggested at the time that "a worst match HRTF negated the benefit that should result from training". The extended results of sessions S3–S6 show no further impact of HRTF quality, regardless of the presentation order.

Participants undertook S1–2 immediately after the HRTF classification task, S3–6 took place at least a week after that. The difference in performance improvement from S1 to S2 between both groups not being repeated in S3–6 could suggest either a difference in game related learning capability between the two groups, or a desensitisation to HRTF quality as acquired after the classification task. The week gap between S2 and S3 did not result in any clear performance drop, suggesting that participants did not "unlearn" the game between both sessions.

As seen in [1], no significant difference was observed between the performance of both groups in S1. In the long run, participants from the first group (started with worst) significantly out-performed those from the second group (average results from S3 to S6: 127.3° and 1.24 s versus 133.0° and 1.32 s), overall showing a steeper learning curve regardless of the HRTF quality. This difference in learning ability between both groups weakens the hypothesis made in [1] on the worst match

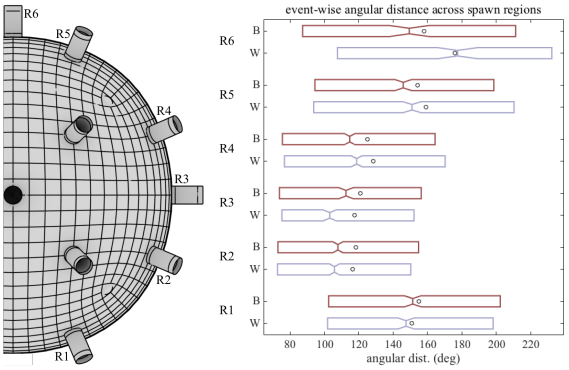

Fig. 6: Event-wise angular distance across target absolute spawn region, separating worst and best match HRTF. R_i represent the spawn region. All spawn positions of a given region share a common elevation.

HRTF negating the benefit that should result from training, which indicates rather an overall performance bias for group 1 participants with respect to group 2.

3.2.4 HRTF×Spawn elevation interaction analysis

Figure 6 presents event-wise angular distance traversed for target spawn elevation and HRTF. Spawn elevation regions R2–4 resulted in significantly lower distributions than R1 and R5–6 (121.7° versus 156.4°) due to participants resting head orientation (facing R3 or R2 most of the time), typical search pattern (azimuth search followed by elevation search), and up/down confusions induced by binaural rendering. As could be expected, participants had difficulties locating targets when spawned from the R6 region, significantly more so when using their worst HRTF (mean angular distance of 158.1° versus 176.0° for best and worst resp.). On average, participants only managed to locate and destroy half of the targets spawned from R6 (54% out of 281 spawns with best versus 44% out of 282 spawns with worst). No significant impact of HRTF quality on angular distance traversed was observed for any other spawn elevation region. No significant region×HRTF interaction was observed for participant reaction time.

3.2.5 Post hoc analysis of participant sensitivity to individualized HRTF

Correlation analysis of per-participant mean angular dist. versus reaction time shows poor results as a function of best and worst match HRTF sessions ($r = 0.30$ and $r = 0.26$ for best and worst resp., for combined spawn elevation regions R1 and R5–6). In contrast, examination of the per-participant mean *differences* between best and worst match HRTF angular dist. versus reaction time, again for spawn elevation regions R1 and R5–6, shows a strong correlation of $r = 0.89$, see Fig. 7.

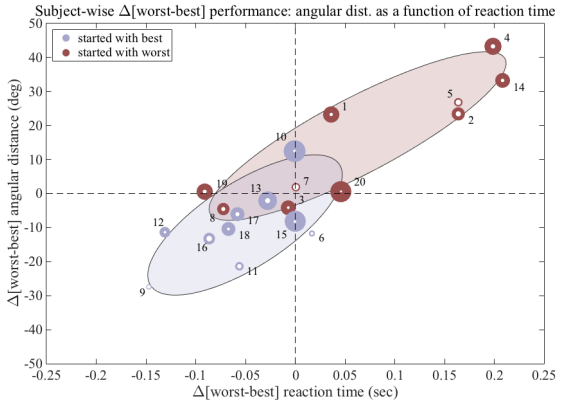

Fig. 7: Participants clustering based on the difference in event-wise performance using best and worst match HRTF for spawn elevation regions R1 and R5–6 combined. Each point represent the performance of a given participant, abscissa and ordinate values are differences in mean response time and angular dist. between best and worst match HRTF sessions respectively. Points in the top-right section (resp. bottom-left) represent participants who best performed with their best (resp. worst) match HRTF. Blue circles are participants who started the game with worst match HRTF, red for those who started with their best. The radius of each circle is proportional to the difference between best and worst match HRTF scores obtained during rating task. Radius of concentric white circles are proportional to mean variance of best and worst match HRTF ratings (see Fig. 3). Numbers are participants game performance rank, based on the best average level across all sessions. Blue and red patches are ellipse fits around blue and red clusters resp. using a least squares criterion.

Clustering of participant performance based on HRTF presentation order is highlighted by enveloping ellipses. This clustering indicates that most participants who performed best with their best match HRTF (resp. best

with their worst match) started S1 with their worst match (resp. with their best match). This result suggests that HRTF presentation order had an impact on HRTF-wise performance, regardless of participant specific HRTF ratings. The HRTF presented in session S2 seems more likely to be "efficient" than the one presented during the first session, where learning the game takes precedence over focus on audio localization. This result, and its interpretation, would deserve a more thorough study, as it could simply be related to a non-uniform distribution of participants ability among both groups as observed in Section 3.2.2. Worth noticing, 4 of the top 5 participants (mean level based ranking) started with their worst HRTF and are located in the upper-right quadrant, *i.e.* performed best with their best HRTF.

The clustering of Figure 7 strongly suggests that some participants are more sensitive to HRTF subjective quality than others, or that at least the classification resulting from the HRTF rating was more appropriate for these participants for the task at hand. Out of 20 participants, HRTF quality proved to have a significant impact on event-wise angular distance traversed for 4 of them, 3 who performed best with their best match HRTF, 1 with the worst match ({1, 4, 14} vs{6} in Fig. 7). For none of the participants did HRTF quality show a significant impact on spawn-spot reaction time.

No clear correlation could be observed between point coordinates and radius, *i.e.* between best-worst HRTF performance difference and best-worst HRTF ratings differences. Likewise for HRTF rating variance (points inner circle size, Fig. 7). This result does not support the "creation of an overall metric to rate participants affinity with binaural hearing [based on participant results to the HRTF classification task]" as suggested in [1].

4 Conclusion

We have presented the results of an experiment designed to assess the impact of individualized binaural rendering on player performance in the context of a VR "shooter game". Participants performed six game sessions, alternatively using their best- and worst-match HRTF, extending the two-session game published in [1] to further dissociate the impact of the learning effect and that of the HRTF on participants performance. During the game, participants had to locate and shoot at successive enemy targets approaching from random directions within a sphere.

Results indicate that the use of a best-match HRTF did not improve overall participants performance regarding the time they needed to localize the targets and the angular distance they travelled before doing so. An analysis focused on spawn region however revealed that angular distance traversed significantly decreased when participants used their best match HRTF for the top-most R6 region of Fig. 6 (mean of 158.1° versus 176.0° for best and worst resp.). Targets spawned in this region were also more often spotted before collision by subjects using their best-match HRTF (54% versus 44%).

Participant clustering in Fig. 7 indicated that 3 out of 20 participants were significantly more efficient in term of angular distance traversed to locate targets with their best HRTF. This result strengthen the hypothesis formulated in [1] that "the benefits of HRTF individualization for a sub-group of "aware listeners" would exceed those of the average participants". While these aware listeners accounted for less than a fifth of the tested participants, it may be that another HRTF selection method, or a different game design, could increase this ratio. A potential impact of the HRTF presentation order on participants sensitivity to best and worst match HRTF was inferred from the results of this clustering (see Section 3.2.5). This last observation would require further study before any assertion can be made.

A last extension of this study is currently being conducted, adding a control group for reference to characterize the learning effect for a fixed set of best versus worst match HRTF during the early stages of the game. A final and more thorough statistical analysis on participant results will then be conducted, to conclude on the impact of HRTF quality on participant performance for the task at hand.

Acknowledgements

This work was funded in part through a fundamental research collaboration partnership between Sorbonne Université, CNRS, Institut ∂'Alembert and Oculus VR, LLC.

References

[1] Poirier-Quinot, D. and Katz, B. F., "Impact of HRTF individualization on player performance in a VR shooter game I," in *2018 AES International Conference on Spatial Reproduction – Aesthetics and Science*, Audio Engineering Society, 2018.

[2] Katz, B. F. G. and Parseihian, G., "Perceptually based head-related transfer function database optimization," *J Acous Soc of Am*, 131(2), pp. EL99–EL105, 2012, doi:10.1121/1.3672641.

[3] Blauert, J., *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT press, 1997.

[4] Begault, D. R., Wenzel, E. M., and 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 Aud Eng Soc*, 49(10), pp. 904–916, 2001.

[5] Seeber, B. U. and Fastl, H., "Subjective selection of non-individual head-related transfer functions," in *Intl Conf on Auditory Display*, pp. 259–262, 2003.

[6] Zotkin, D., Hwang, J., Duraiswaini, R., and Davis, L. S., "HRTF personalization using anthropometric measurements," in *IEEE Workshop on Applications of Sig Proc to Audio and Acoustics*, pp. 157–160, 2003.

[7] Carpentier, T., Bahu, H., Noisternig, M., and Warusfel, O., "Measurement of a head-related transfer function database with high spatial resolution," in *Forum Acusticum (EAA)*, 2014.

[8] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L., "Localization using nonindividualized head-related transfer functions," *J Acous Soc of Am*, 94(1), pp. 111–123, 1993.

[9] Xu, S., Li, Z., and Salvendy, G., "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," in *Intl Conf on Virtual Reality*, pp. 397–407, Springer, 2007.

[10] Härmä, A., van Dinther, R., Svedström, T., Park, M., and Koppens, J., "Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface," in *Aud Eng Soc Conv 132*, 2012.

[11] Mehra, R., Nicholls, A., Begault, D., and Zannoli, M., "Comparison of localization performance with individualized and non-individualized head-related transfer functions for dynamic listeners," *J Acous Soc of Am*, 140(4), pp. 2956–2957, 2016.

[12] Andreopoulou, A. and Katz, B. F., "Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees," *JMUI*, (SI: Auditory Display), pp. 1–13, 2016.

[13] Andreopoulou, A. and Katz, B., "Investigation on Subjective HRTF Rating Repeatability," in *Aud Eng Soc Conv 140*, pp. 9597:1–10, 2016.

[14] Poirier-Quinot, D. and Katz, B. F., "The Anaglyph binaural audio engine," in *Aud Eng Soc Conv 144*, 2018.

[15] Wright, M., "Open Sound Control: an enabling technology for musical networking," *Organised Sound*, 10(3), pp. 193–200, 2005.

[16] Katz, B., Tarault, A., Bourdot, P., and Vézien, J.-M., "The use of 3D-audio in a multi-modal teleoperation platform for remote driving/supervision," in *Aud Eng Soc Conf: Intelligent Audio Environments*, pp. 1–9, Saariselkä, 2007.