Audio Engineering Society

Convention Express Paper 440

Presented at the AES 160th Convention
2026 May 28–30, Copenhagen, Denmark

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Can the individual winner HRTFs be determined in a shooting task during onboarding for an Audio Only VR?

Max Væhrens[1], Stefania Serafin[2], Flemming Christensen[1], and Dorte Hammershøi[1]

[1]*Acoustics and Hearing, AI and Sound, Department of Electronic Systems, Aalborg University*
[2]*Department of Engineering Technology and Didactics, Technical University of Denmark*

Correspondence should be addressed to Max Væhrens (maxv@es.aau.dk)

ABSTRACT

This study is inspired by the ambition to determine the individually best HRTFs during an onboarding task for an audio-only VR experience using a 'shooting down sound source' task. The study is motivated by the needs of blind and visually impaired users, who may rely more critically on correct rendering of auditory spatial cues. As a first step, we have implemented an experimental VR platform that renders sound source 'targets' in a virtual environment and logs head and handheld controller positional data while participants localize and 'shoot' the audible 'targets' one by one. In a pilot study, six game-relevant stimuli were tested across visible familiarization 'targets' followed by ten invisible 'targets' per session. Data were collected from seven sighted participants, which demonstrate individual precision-time trade-offs, with emerging prototype search patterns (e.g., hemisphere swaps and cone-of-confusion iterations). The prototype patterns are likely quantifiable via angular error, response time, path length, and search speed. Notably, for 'shooting' precision, head orientation slightly outperforms controller pointing for invisible 'targets'.

1 Introduction

Virtual Reality (VR) games generally make wide use of visuals for the game scene rendering as well as for the game controls. Blind VR gamers are challenged, and the overall purpose of the present efforts is to further develop VR game opportunities, which may serve the blind and visually impaired (BVI) gamers better in rendering convincing game effects with natural life relevance, as well as controlling and engaging actively in tailored VR games.

Literature suggests that BVI individuals could be a more critical user population, when considering the auditory rendering of the game scenes [1–8] (although also contested [9, 10]). BVI use their hearing for everyday navigation and identification of the sources in their surroundings, and may be attentive to subtle differences in the room acoustics, which sighted gamers

would not notice. Likewise, we may also speculate that they might rely more critically on having their own head-related transfer functions (HRTFs) for rendering the spatial aspects of the sources in the game scene than sighted gamers, who can see the sources.

Several studies have demonstrated the advantage of using individual HRTFs in tasks that require localization of the sound sources [11–14], and there are several suggestions for how to select the 'best' HRTFs [15–20]. Yet it can be difficult to know, if these advantages are merely a minor fine-tuning giving only a marginal advantage, or whether they are the critical parameters for a given game task and experience. It may be the first for sighted gamers and the latter for the BVI population.

One option is to facilitate the configuration of HRTFs as part of the onboarding process, where the gamer may actively engage with components of the game environment to reveal the better HRTFs for the gamer's performance. This can in principle be determined autonomously without the participant's subjective assessment.

In the present study, an experimental game platform utilizing a classical 'shooting down sound sources' task has been developed with the purpose of providing a game component, where the gamer has to utilize active sound localization to perform well. The present paper reports pilot data and shows a few typical data trends. More experiments are in progress, which among other will test the effect of different HRTFs. Whether the 'best' HRTFs are 'best' based on performance measures because they closely match the individual's own characteristics or because they provide spatial cues that are generally superior remains to be explored.

2 Methods

An experimental test system has been implemented in a standard VR platform, which renders sound sources visually and audibly - inspired by a 'classical shooting down sound sources' task [21]. Participants are instructed to find and 'shoot' down the sound sources, which are placed one at a time in different fixed positions in the virtual environment. Whenever a source has been shot, the sound will switch to a new position for the participant to find. The scenario enables evaluation of localization ability, and reveals search and find patterns that may among other characterize the participants personal time-accuracy trade-offs.

2.1 Test Platform

The test platform was developed in the Unity engine [22] for the Meta Quest Pro VR head-mounted display (HMD) and the two included controllers. Position and rotation of the HMD and controllers were tracked continuously during tests at 50-60 Hz.

The virtual environment (VE) is an empty sandbox scene with a distant horizon and a wide plane illustrating the floor. The VE has anechoic acoustic conditions. The pilot data were collected in an office using the HMD's built-in speakers, hence not entirely sound isolated and with risk of noise from the surroundings.

2.2 Shooting Task

Participants were instructed to search for the sound, and 'shoot' it down (by pulling a trigger with their index finger on any of the two hand-held controllers). As the main purpose of the pilot tests was exploratory, there were no success criteria other than the participants' decision to 'shoot', nor any instructions on how to balance time and accuracy (subjective certainty or confidence in having found the target).

The participants were seated in an office chair and provided the information that sounds could occur in any random direction in space, and that they should explore the space by turning the chair and moving their head. There were no further instructions regarding how to explore the space. It was observed that some would turn the chair frequently, whereas others would move very little. As for the use of the hand-held controller, no instructions were given to the 'correct' use. Hence, some participants used it as a pointer during exploration of space, whereas others would leave their arm hanging until they had located the source and were ready to 'shoot' it.

2.3 Participants

The seven participants all had normal vision and were three of the authors, two other colleagues, and two adult friends - all familiar with the project and its overall purpose.

2.4 Test duration

The sessions for the seven participants lasted 10 to 35 minutes, where the two longest sessions were affected by technical problems and required restart of the system. The total duration the seven participants were actively engaged in the 'shooting' task lasted 3-12 minutes.

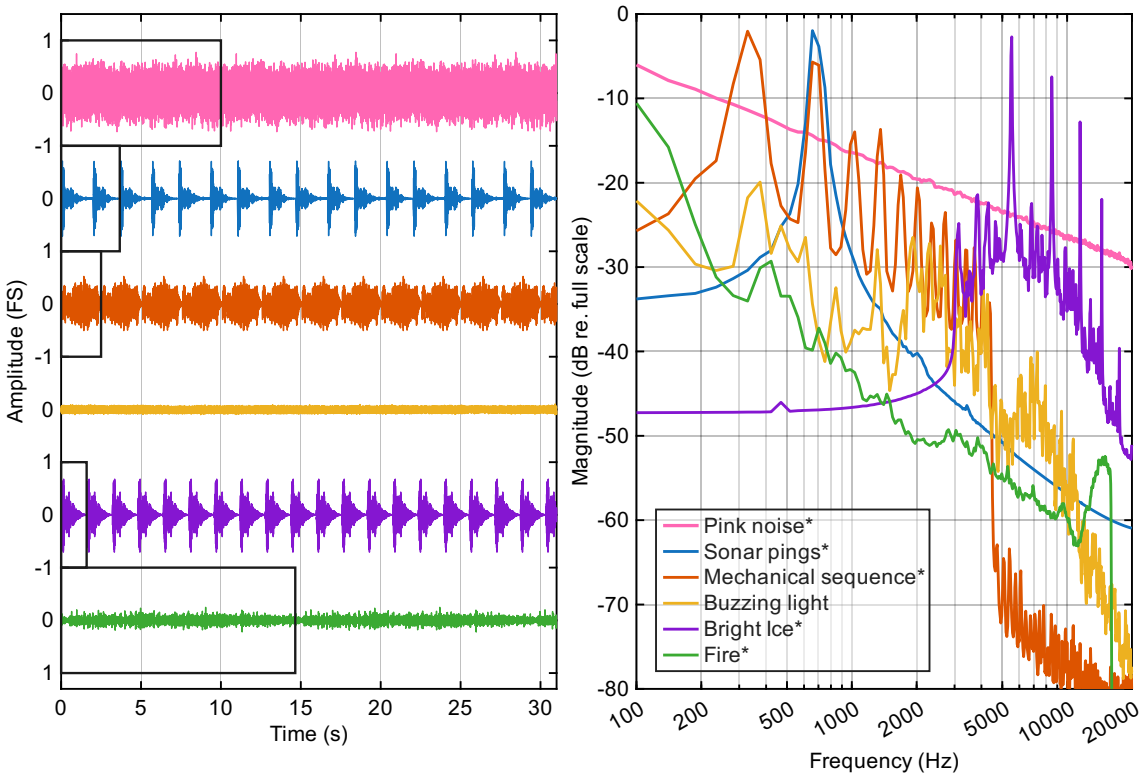


Fig. 1: Waveforms (left panel) and magnitude spectrum (right panel) of the six sound stimuli (resampled to 48 kHz). Stimulus marked with * has been looped a number of times to illustrate a common time axis, boxes around waveforms indicate length of each stimulus. Full scale level is represented by ± 1 and 0 dB, respectively.

2.5 Binaural rendering

The sound was spatialized using the Meta XR Audio SDK for Unity (v77.0). The built-in HRTFs and built-in speakers were used in the pilot tests reported here.

2.6 Target positions

In the pilot experiment the source positions used the same spherical band distribution of 19 positions (radius between 1.5 and 2.6 meters) used in [21] with the sphere centred at position (x,y,z)=(0,0,1.7) in the VR scene. The same fixed protocol for source positions was applied for the first five trials for familiarization with the task and sound stimuli. These five positions were also represented visually as pink balls (radius=0.5m) in the VE. Then followed playback from ten invisible source positions randomly selected from the 19 available positions with the only limitation that a given position would not occur consecutively.

2.7 Target stimuli

Six different types of sound stimuli that by nature could be game relevant were used. Waveforms and spectrum of the sound stimuli are shown in Fig. 1, and included are: Pink noise, 'sonar pings', 'mechanical sequence', 'buzzing light', 'bright ice', and 'fire'. Two transient sound stimuli were included: The 'sonar pings' contain two mid-frequency bursts and the 'bright ice' contains a single high frequency burst. All stimuli were retrieved online (from FreeSound [23]), and had information about natural sound level. All sounds had a relatively broad spectral content. The 'sonar pings' and the 'mechanical sequence' both had audible tonal components, as also reflected in the spectra.

2.8 Stimuli order and playback level

The six different stimuli were used in six separate sessions, and played back in the same order for all participants (as listed above). The sound looped continuously

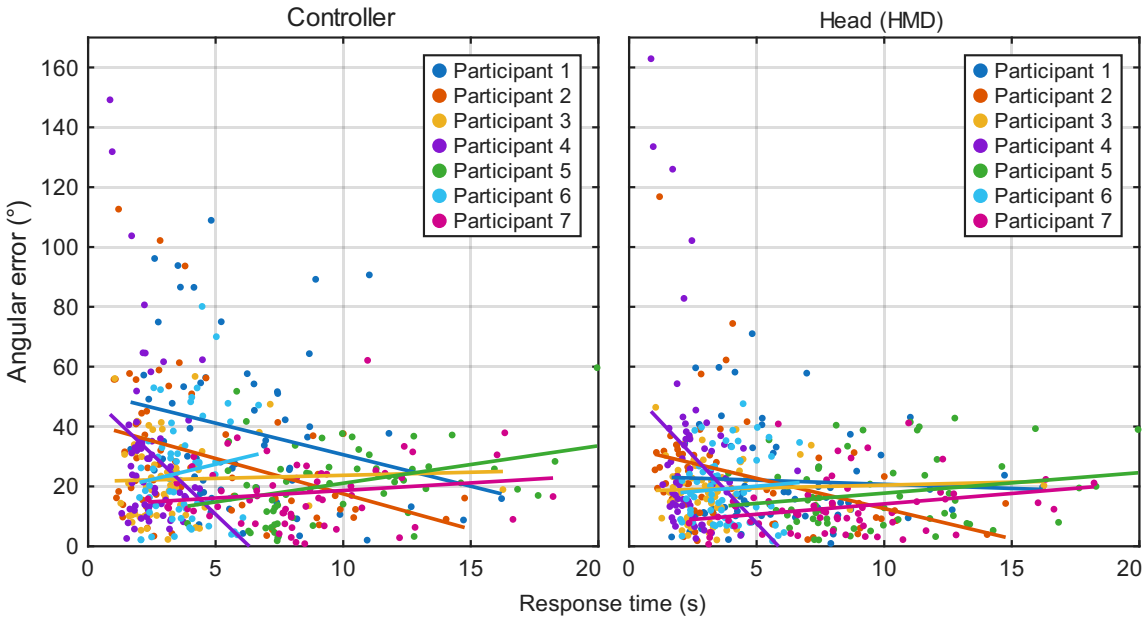


Fig. 2: Angular error versus response time for the hand-held controller (left panel) and the head-mounted display (right panel) for the ten invisible 'targets' in the pilot experiment. Each data point represents a unique response. Data for each of the seven participants are plotted with a unique colour, including corresponding regression lines, for all six sound stimuli.

during the session until all 15 'targets' had been 'shot'. The participant would listened only to one stimulus type at a time at their preferred level. There was a small break between each sound type. The playback level was sometimes adjusted during the first trials when the sound type changed.

2.9 Data processing

Continuous position and orientation data for HMD and controller, as well as registration of discrete trial events (e.g. trigger pulls, trial outcomes, angular error, and response time) were logged through a custom script utilizing Unity Experimental Framework [24]. The start and end positions for each search and find process are determined and the length of the actual spherical search path computed as well as the speed of the search (computed from frame-to-frame displacement of direction vectors divided by the frame time interval). The precision for the 'hit' (computed as angular error, i.e. the angular difference between 'target' direction and the direction indicated) and the elapsed time before the subject gave their response (response time) and corresponding average values and standard-deviations have been computed.

Trends in the dependence of subject, sound, and source position will be explored both qualitatively and quantitatively. If possible, prototype search patterns will be identified through visual inspections and tailored statistics that characterize the different prototype patterns (monotonous direct path, front-back or other cone-of-confusion iterations, source independent behaviours, etc.)

3 Results

The pilot test results are presented in the following.

3.1 Precision versus elapsed time

The angular error versus the response time is plotted in Fig. 2 for data from the hand-held controller and the head (HMD) for the ten invisible 'targets' in the pilot experiment. A notable observation is when comparing all data points and average values, the data from the

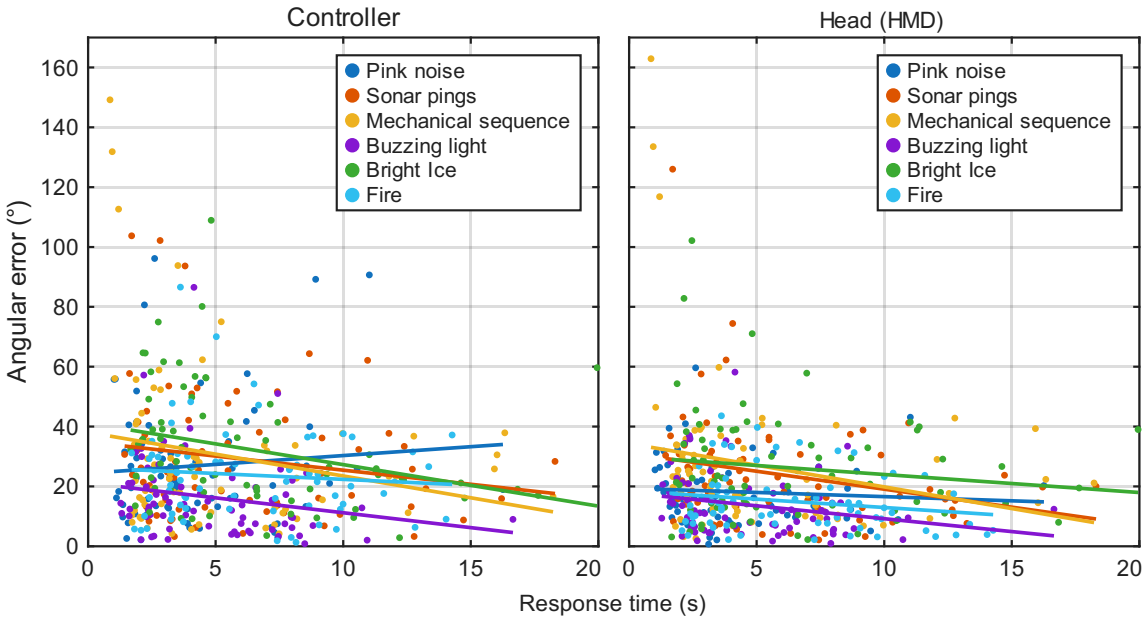


Fig. 3: Angular error versus response time for the hand-held controller (left panel) and the head-mounted display (right panel) for the ten invisible 'targets' in the pilot experiment. Each data point represents a unique response. Data corresponding to each of the six stimuli are plotted with a unique colour, including corresponding regression lines, for all seven participants.

HMD appear to give slightly higher precision. This is a bit surprising, given that the instructions were to 'shoot' the sound sources with the hand-held controller, and the subjects were not instructed to use their head in any particular way, nor to look at the invisible source when 'shooting' it. These data suggests that even when 'targets' were invisible, sighted participants explored space as if there was something to look for.

In Fig. 2, it can also be seen that the seven individuals employ different search and find strategies with respect to precision versus time. Participants 4 (purple) and 6 (teal) systematically use less than five seconds before 'shooting', whereas most other participants vary the time they spent searching. Analyses of the path length between two consecutive source directions may reveal that under a fixed time paradigm (as apparently chosen by participant 4 and 6), lower precision may occur with sources further apart or from targets with weaker localization cues. The individual regression lines in Fig. 2 reveal that some (e.g. participant 1, 2 and 4 appear to have the expected precision-time trade-off, with higher precision when more time is spent searching.

Fig. 3 shows that when comparing across stimuli, it seems to be the 'mechanical sequence' and the 'bright ice' that are hardest to localize and pink noise, 'buzzing light' and 'fire' that are easiest to localize. The three latter stimuli are likely the easiest to localize due to having a constant non-transient response and the broadest frequency content, whereas the 'mechanical sequence' and 'bright ice' both have a narrower frequency content and contain transient sounds. The data in Fig. 3 also suggests that when the personal traits are averaged out (showing only dependence of stimuli), the regression lines are also closer to the expected trend in the precision-time trade-off. However, one exception is the regression line for controller and the pink noise stimulus (blue, left panel), which may be due to it being the first stimuli presented and participants were not as well trained in using the controller to point towards invisible sources.

3.2 Search and find trajectories

Figure 4 shows the difference between the 'actual' source direction (in the virtual environment) and the

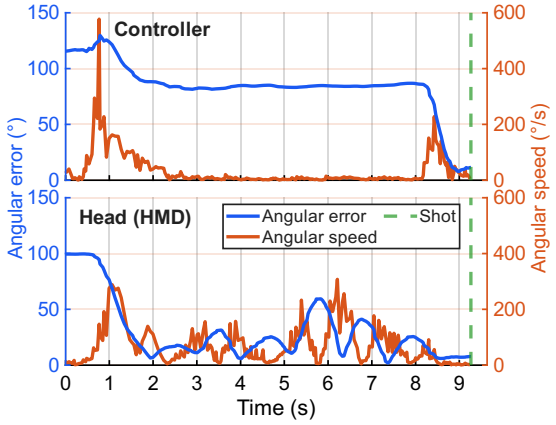


Fig. 4: Angular error and angular speed versus elapsed time for the trajectory mapped on a unit sphere of the hand-held controller (top panel) and the head-mounted display (bottom panel) for a single 'shooting' trial.

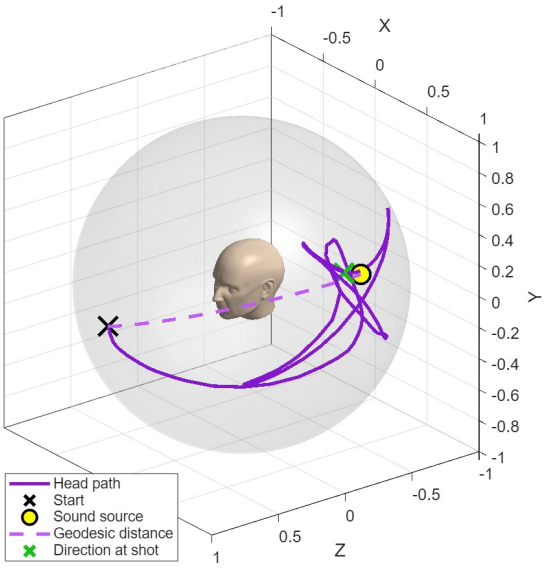


Fig. 5: Trajectory for head-mounted display pointing onto a unit sphere from start position (X) to shot position (circle) for a single 'shooting' trial. The head model points towards start position at trial start. Full circle indicates the 'target' (sound source direction) and dashed line indicates shortest distance on sphere from start direction to the 'target'.

pointing direction computed from the hand-held controller and the HMD as a function of time. Initially, the angular error is high for both trackers. For the hand-held controller, the speed is high in the beginning, where the hand is likely taken down from the preceding 'shot', then the angular speed is close to zero, in the period where the participant is searching, and eventually there is a small movement before the 'shot', where the angular error is small for both trackers.

The data from the HMD (bottom panel in Fig. 4) show the head-based search pattern. The head orientation gets really close to the 'target' direction already within the first two seconds, and yet the participant iterates away from the right direction several times before giving the 'shot' after approximately eight seconds. This would suggest that although the final response is close to the 'target', the subject seemed very much in doubt all the way and may have needed assurance through head movement before confirming the 'shot'. Fig. 5 reveals that the head search path starts at a hemisphere opposite to the 'target', and immediately moves to the 'target' area, where the head direction is zigzagged around the 'target' position. This pattern might be a prototype for poor spatial cues.

Another possible prototype pattern can be seen from Figs. 6 and 7, where the participant starts in a spatial region close to the 'target', but after less than two seconds turn to the spatial region in the opposite hemisphere, which is then explored for 6-7 seconds, before returning to the correct hemisphere. The exploration continues for another 10 seconds before eventually concluding the search with a 'shot' quite close to the 'target'. Having to swap from one hemisphere to another during search suggest that the localization cues are weak, even though the 'target' is eventually found with good precision.

4 Discussion

During game on-boarding, many instructions are delivered that refer to the game experience, how to navigate and advance, and how to use the controls. This information would logically be conveyed in a spoken dialogue in an audio-only game for blind users. The present pilot study employed artificial sounds that may be relevant for the game elements, but which the participant may be unfamiliar with at the beginning. This might not be the best-case scenario for successful localization, and

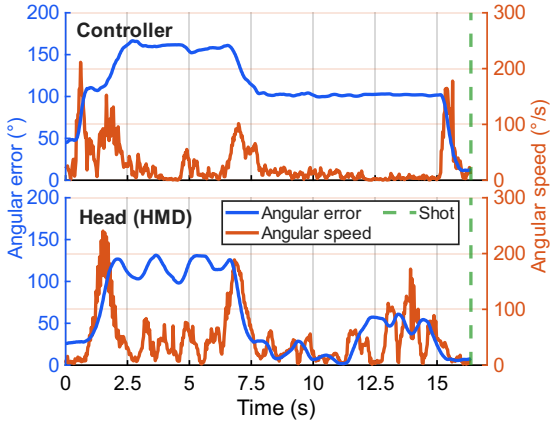


Fig. 6: Angular errors and angular speed versus time for the trajectory mapped on a unit sphere of the hand-held controller (top panel) and the head-mounted display (bottom panel) for a single 'shooting' trial.

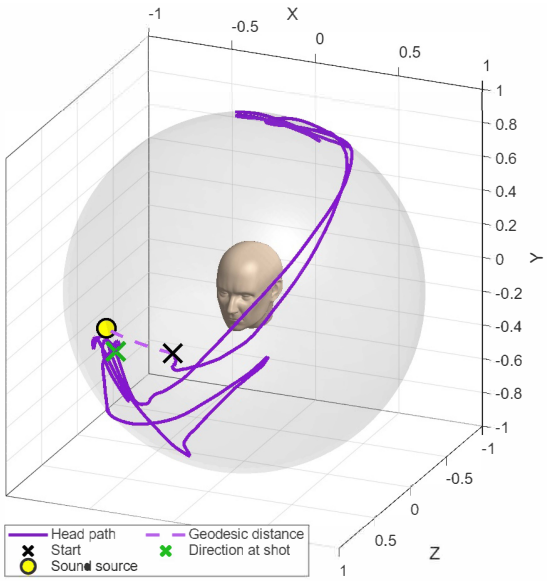


Fig. 7: Trajectory for head-mounted display pointing onto a unit sphere from start position (X) to shot position (circle) for a single 'shooting' trial. The head model points towards start position at trial start. Full circle indicates the 'target' (sound source direction) and dashed line indicates shortest distance on sphere from start direction to the 'target'.

determining the best HRTFs. On the other hand, they all had broad spectral content, allowing the participant to use the salient features of the HRTFs for localization, and would provide the added benefit of familiarization before game start.

5 Conclusion

An experimental platform for testing the user's localization abilities in a simplistic virtual environment has been implemented. Sources can be configured arbitrarily in space, with positions either pre-configured or in the end solution computed on the fly during on-boarding to the game. A test paradigm that introduces a 'shooting' task and monitors the successes and failures is suggested as a game on-boarding component. The motivation is to enable the determination of the most favourable HRTFs for the individual gamer, a feature yet to be implemented and further explored. Further studies may exploit whether it is possible to find the objectively 'best' HRTF, and may also assess the user satisfaction for this and alternative paradigms, which are driven by user preferences rather than objective advantages.

6 Acknowledgement

Funding by the Independent Research Fund Denmark (grant 4264-00072B) is sincerely acknowledged. The collaboration with Stine Froulund Astorp, Nicoleta Diana Radu, Antionia Line Krummheuer, and Thomas Bjørner as well as technical support from Claus Vestergaard Skipper is sincerely acknowledged. We would also like to acknowledge the mutual inspiration from two master students working in a parallel project that through the use of AI would determine the better sound quality control.

References

[1] Morgan, M., "Sensory perception: Supernormal hearing in the blind?" *Curr. Biol.*, 9(2), pp. R53–R54, 1999.

[2] Voss, P., Lassonde, M., Gougoux, F., Fortin, M., Guillemot, J.-P., and Lepore, F., "Early- and Late-Onset Blind Individuals Show Supra-Normal Auditory Abilities in Far-Space," *Curr. Biol.*, 14(19), pp. 1734–1738, 2004.

[3] Schepers, I. M., Hipp, J. F., Schneider, T. R., Röder, B., and Engel, A. K., "Functionally specific oscillatory activity correlates between visual and auditory cortex in the blind," *Brain*, 135(3), pp. 922–934, 2012.

[4] Boroujeni, F. M., Heidari, F., Rouzbahani, M., and Kamali, M., "Comparison of auditory stream segregation in sighted and early blind individuals," *Neurosci. Lett.*, 638, pp. 218–221, 2017.

[5] Shim, H. J., Go, G., Lee, H., Choi, S. W., and Won, J. H., "Influence of Visual Deprivation on Auditory Spectral Resolution, Temporal Resolution, and Speech Perception," *Front. Neurosci.*, 13, 2019.

[6] Battal, C., Occelli, V., Bertonati, G., Falagiarda, F., and Collignon, O., "General Enhancement of Spatial Hearing in Congenitally Blind People," *Psychol. Sci.*, 31(9), pp. 1129–1139, 2020.

[7] Sabourin, C. J., Merrikhi, Y., and Lomber, S. G., "Do blind people hear better?" *Trends Cogn. Sci.*, 26(11), pp. 999–1012, 2022.

[8] Fine, I. and Park, W. J., "Do you hear what I see? How do early blind individuals experience object motion?" *Philos. Trans. R. Soc. B Biol. Sci.*, 378(1869), 2023.

[9] Zwiers, M. P., Opstal, A. J. V., and Cruysberg, J. R. M., "A Spatial Hearing Deficit in Early-Blind Humans," *J. Neurosci.*, 21(9), pp. RC142–RC142, 2001.

[10] King, A. J., "What happens to your hearing if you are born blind?" *Brain*, 137(1), pp. 6–8, 2014.

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

[12] Moller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D., "Binaural Technique: Do We Need Individual Recordings?" *J. Audio Eng. Soc.*, 44(6), pp. 451–469, 1996.

[13] Møller, H., Jensen, C. B., Hammershøi, D., and Sørensen, M. F., "Using a typical human subject for binaural recording," in *In Proc. of 100th Audio Engineering Society Convention*, 4157, pp. 1–18, 1996.

[14] Fels, J., Braren, H., and Bomhardt, R., "Comparison of the front-back confusion rate of individual and individualized head-related transfer functions," in *DAGA 2018*, 2018.

[15] Poirier-Quinot, D. and Katz, B. F., "On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3D room response," *Acta Acustica*, 5, p. 25, 2021.

[16] Seeber, B. U. and Fastl, H., "Subjective Selection of Non-individual Head-related Transfer Functions," in *Proc. of the 2003 Int. Conf. on Auditory Display*, pp. 259–262, 2003.

[17] Katz, B. F. G. and Parseihian, G., "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Am.*, 131(2), pp. EL99–EL105, 2012.

[18] Daugintis, R., Barumerli, R., Picinali, L., and Geronazzo, M., "Classifying Non-Individual Head-Related Transfer Functions with A Computational Auditory Model: Calibration And Metrics," in *ICASSP 2023 - 2023 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[19] Bahu, H., "Auditory localization in the context of non-individual binaural synthesis," in *Summary of the PhD*, pp. 1–9, Institute of Research and Coordination in Acoustics and Music (IRCAM), 2016.

[20] Zagala, F., Noisternig, M., and Katz, B. F. G., "Comparison of direct and indirect perceptual head-related transfer function selection methods," *J. Acoust. Soc. Am.*, 147(5), pp. 3376–3389, 2020.

[21] Sandvad, J., "Dynamic Aspects of Auditory Virtual Environments," in *Proc. of 100th Audio Engineering Society Convention, Copenhagen, May 11-14, 1996*, pp. 1–14, 1996.

[22] Unity Technologies, "Unity Engine," https://unity.com/, 2025.

[23] Freesound, "Online sound database," https://freesound.org/, 2025.

[24] Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., and Mushtaq, F., "Studying human behavior with virtual reality: The Unity Experiment Framework," *Behav. Res. Methods*, 52(2), pp. 455–463, 2020.