Audio Engineering Society

Convention Paper 10638

Presented at the 154th Convention
2023 May 13–15, Espoo, Helsinki, Finland

Navigation of virtual mazes using acoustic cues

Sebastià V. Amengual Garí, Paul Calamia, and Philip Robinson

Reality Labs Research, Meta, Redmond, USA

Correspondence should be addressed to Sebastià V. Amengual Garí (samengual@meta.com)

ABSTRACT

We present an acoustic navigation experiment in virtual reality (VR), where participants were asked to locate and navigate towards an acoustic source within an environment of complex geometry using only acoustic cues. We implemented a procedural generator of complex scenes, capable of creating environments of arbitrary dimensions, multiple rooms, and custom frequency dependent acoustic properties of the surface materials. For the generation of the audio we used a real-time dynamic sound propagation engine which produces spatialized audio with reverberation by means of bi-directional path tracing (BDPT) and is capable of modeling acoustic absorption, transmission, scattering, and diffraction. This framework enables the investigation of the impact of various simulation properties on the ability of navigating a virtual environment. To validate the framework we conducted a pilot experiment with 10 subject in 30 environments and studied the influence of diffraction modeling on navigation by comparing their navigation performance in conditions with and without diffraction. The results suggest that listeners are successfully able to navigate VR environments using only acoustic cues. In the studied cases we did not observe a significant effect of diffraction on navigation performance. A significant amount of participants reported strong motion sickness effects, which highlights the ongoing issues of locomotion in VR.

1 Introduction

Environmental acoustic cues are critical for spatial awareness. Humans are capable of perceiving the presence of objects using echolocation by exploiting level, spectral envelope, echo delay, and binaural differences [1]. Although blind people generally perform better, both blind and sighted people are capable of navigating spaces utilizing echo-location [2], as well as producing spatial maps of the environment by leveraging the presence of active sources in the environment, producing comparable results in real and virtual environments [3]. As such, spatial audio has been extensively used for the training and investigation of spatial awareness, both in audio only applications [4, 5, 6, 7, 8, 9] and in multi-modal applications e.g. a virtual haptic-acoustic cane [10] or interactive audio-tactile maps [11]. However, these environments generally show some limitation in the form of low geometrical complexity of the environments (often being reduced to single rooms with non-rectangular geometries), small variety of investigated scenes, and/or lack of head tracking, rendering the interaction unnatural.

In this paper we present a navigation experiment conducted in Virtual Reality (VR), in which users are asked to navigate inside procedurally generated complex environments with the goal of locating an acoustic source. The main advantage of our framework lies in the flexibility of the procedural generation of environments coupled with a real-time dynamic sound propagation engine capable of handling complex geometries and

Parameter	Type	Description
Size X	Integer	Size of the maze in the X dimension, in number of cells.
Size Y	Integer	Size of the maze in the Y dimension, in number of cells.
Room types	Integer	Number of different room types contained in the scene. There are 4 predefined room types, each with different acoustic and visual materials, which can be customized.
Large rooms	Boolean	Determines whether rooms are created by combining multiple cells in both X and Y dimensions, or whether rooms are formed like corridors in a maze style.
Portal probability	Float	Probability of two adjacent cells from a different room type being connected by a door frame. The higher the number, the more rooms are connected.
Seed	Integer	Initializes a specific instance of geometry, source position, and initial listener position.

Table 1: Control parameters of the procedurally generated environment.

frequency dependent material properties, allowing us to conduct the navigation task in any arbitrary acoustical environment. In this work we leverage the introduced framework in the investigation on the relevance of diffraction modeling in real-time simulations and we evaluate the impact of diffraction on the acoustical navigation performance of humans in complex VR environments. Additionally, besides the evaluation and validation of real-time audio simulations, these kinds of navigation tasks could be leveraged for a variety of purposes, such as training first responders in acoustically guided navigation in emergencies, or the investigation of relevant acoustical phenomena in spatial awareness and perceptual spatial mapping, among others.

2 Framework

2.1 Procedural creation of environments

In order to efficiently generate a large number of environments, we implemented a procedural maze generator in Unity, which allows the experimenter to specify the properties of the environment to obtain a randomly generated environment. The external boundaries of the environment always correspond to a rectangular space, whose size is determined by a configurable number of cells. An arbitrary number of room types can be generated, each with different acoustic and visual properties. A flag controls whether rooms resemble those found in daily environments or they are generated as corridors, similar to traditional mazes. The presence of openings between rooms can be controlled using a probability value. Finally, the use of a seed value permits the pseudo-random generation of a virtually infinite number of environments sharing the same statistical properties. The positions of both listener and sound source(s) is randomized and depends on the generation seed. The number of sources and their content can be arbitrarily defined.

Details about the control parameters are provided in Tab. 1. Examples of a few illustrative environments, together with their control parameters are provided in Fig. 1. The size and complexity of environments can be arbitrarily manipulated, generating from simple small environments, to very complex and large spaces.

2.2 Audio simulation engine

The real-time acoustic simulations were performed using the RLR-Audio-Propagation custom engine running as a plug-in for Unity. The engine conducts a bi-directional path tracing (BDPT) [12] simulation in N arbitrary logarithmically spaced frequency bands and generates the final audio by means of artificial reverberators spatialized in the spherical harmonics domain (SHD). In this study we used simulations in 4 frequency bands. The BDPT process handles the modeling of frequency dependent acoustic absorption, transmission, scattering, and high order diffraction. In particular, scattering is modelled using Bidirectional Reflectance Distribution Functions (BRDF). A detailed description of the diffraction implementation is provided in [13]. The final binaural signals are rendered utilizing generic Head Related Transfer Functions (HRTF) encoded into the SHD using Magnitude Least Squares as in [14].

The audio engine has been succesfully used in a variety of multimodal Deep Learning tasks, such as the navigation of virtual agents in complex acoustic environments [15], the audiovisual reconstruction of floorplans [16], or the learning of spatial cues via echolocation [17]. Recently, the acoustic simulator has been deployed with SoundSpaces 2.0, a simulation platform for audiovisual Deep Learning tasks [18] and is publicly available for research purposes.

An instrumental validation of the acoustic engine is provided in [18], comparing the Reverberation Time (RT60) and Direct-to-Reverberant Ratio (DRR) of measurements against the results of the simulations. In order to minimize the error introduced by the manual annotation of materials, a first annotation using acoustic material data from a dataset was followed by an

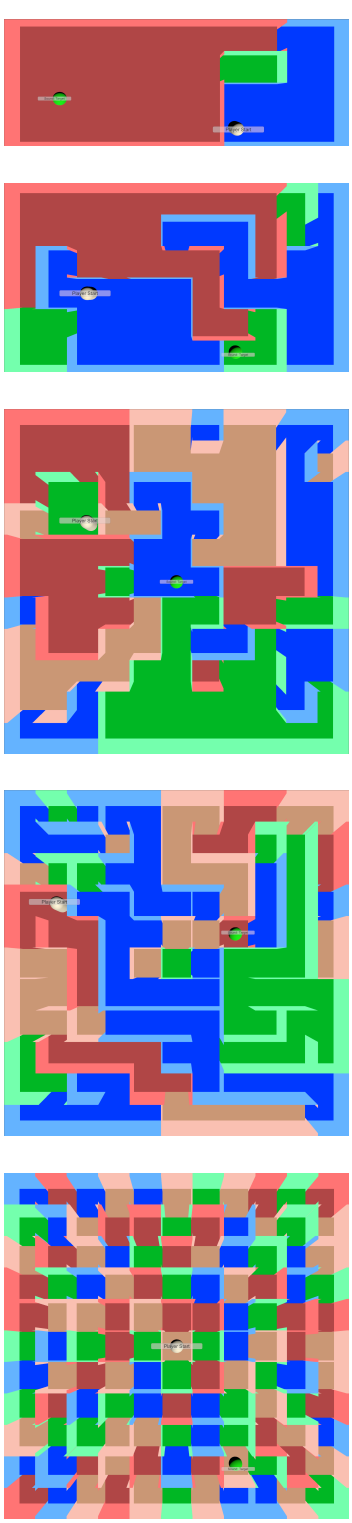

Fig. 1: Example environments of varying complexity.

acoustic optimization process that iteratively modifies the material properties to minimize the discrepancy between measurements and a series of reference simulations [19]. The engine adaptively adjusts the quality of the simulation to fulfill computational constrains of the scene. In the highest fidelity mode, the RT60 error is 0%, while in the fast mode it increases to 9.5% in a benchmark dataset (highly reverberant apartment scene). Overall, the mean RT60 relative error on a larger dataset is 12.5%, suggesting that the errors are generally comparable to audible RT60 discrimination thresholds [20]. The errors on DRR are 0.98 dB on average, with maximum errors of 2 dB in the benchmark scene, which are below the discrimination thresholds [21]. We refer the reader to [18] for further details on the acoustic analysis.

3 Experiment

3.1 Task

The goal of the experiment is twofold: First, to determine whether subjects are consistently able to navigate to and locate sources in complex multi-room virtual environments, and second, to investigate whether the rendering of sound diffraction in a complex environment significantly affects their navigation performance. To this end, an experiment was conducted in VR, and subjects were asked to navigate as fast as possible to an invisible target sound source (a periodic broadband noise burst). No distractor sources were present in this task.

The test platform was the Meta Quest 2 connected via USB-C to a host PC running the Unity test application. The scene was visible to the participants in the Head Mounted Display (HMD). The locomotion was performed using the HMD controllers via a combination of the left joystick for continuous translation and the right joystick for rotation in 90° increments. Rotation of the head could also be used for continuous rotation.

In each trial, subjects spawned in a random location and after navigating to the sound source they had to locate it with one hand and hold the trigger on that controller for a short period of time. Before the actual test, 4 trials were provided with experimenter guidance to become acquainted with the task, procedure, and apparatus. A total of 30 scenes were evaluated. Half of the scenes were presented with diffraction effects disabled and the other half with diffraction enabled.

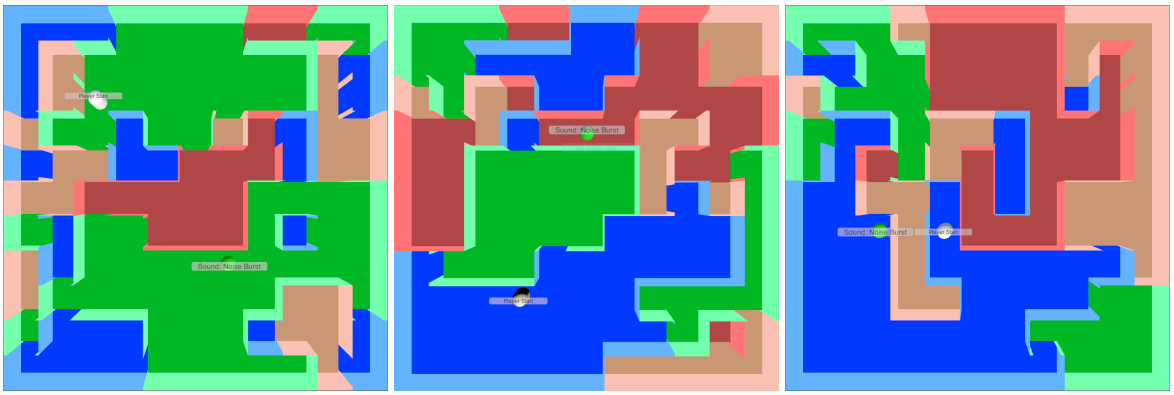

Fig. 2: Example of three environments used in the test. The parameters of the environments are *size X* = 10, *size Y* = 10, *room types* = 4, *large rooms* = *true*, and *portal probability* = 0.18.

The test was conducted remotely and 10 subjects participated using their own hardware. The study required users to have a VR-ready computer running Windows, a Meta Quest 2 HMD, and external headphones. The majority of the users completed the test using Beyerdynamic DT990 Pro headphones, except a few exceptions that used headphones of comparable quality. The users were asked to adjust the reproduction level to a comfortable level at the beginning of the session during the training trials and leave it constant for the remainder of the experiment. In order to minimize the risks of running remote tests in uncontrolled environments, the experimenter provided guidance via videoconference at the beginning of the experiment and during the training trials and was available during the experiment for troubleshooting and monitoring of the experiment. Subjects were asked to complete a post study questionnaire, which was completed by 9 out of 10 subjects.

3.2 Test environments

The 30 evaluated scenes were all of comparable geometrical complexity and were generated using the same parameters, differing only in the generation seed. A few examples of the test environments and the generation parameters are provided in Fig. 2.

Four room types with different acoustic materials were used. Blue rooms used concrete acoustic properties, green rooms used wood, red rooms used carpet, and tan rooms used steel and glass. See Tab. 2 for a description of the materials in each room surface and Tab. 3 for the absorption, scattering, and transmission coefficients of each material.

Room	Floor	Walls	Ceiling
Blue	Concrete rough	Concrete block	Concrete rough
Green	Wood floor	Wood thick	Wood thin
Red	Carpet heavy padded	Carpet heavy	Carpet
Tan	Steel	Glass heavy	Steel

Table 2: Acoustic materials of the rooms.

3.3 Metrics

In order to evaluate the navigation performance we propose the use of multiple metrics:

- *Source visibility*: A trial is considered successful if there is a direct line of sight (LOS) between the subject's head and the source at the end of the trial. Trials without a direct LOS are discarded in further analysis.
- *Distance to source*: The subject's final distance to the source, computed as the minimum distance between the source and the position of the controller used to input the location.
- *Total Navigation Time*: The total duration of a trial, since the listener spawns in the scene until they introduce their answer.
- *Cropped Navigation Time*: The time elapsed between the start of the user movement and the moment in which their distance to the source is less than an arbitrary distance.
- *Total Path Length*: The total distance travelled by the subject, since they spawn until they introduce their answer.

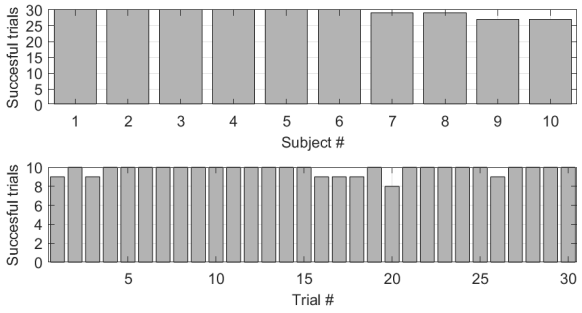

Fig. 3: Successful trials per subject (top) and per scene (bottom).

4 Results

4.1 Source visibility

The vast majority of trials were completed successfully (292/300 trials, 97.33% success rate), with all subjects completing at least 90% of the trials successfully. The unsuccessful trials were concentrated within 4 subjects, with 2 subjects failing to complete 3 trials, and 2 subjects failing in 1 trial. With regards to the environments, the unsuccessful trials were spread among all environments, with only one environment presenting 2 unsuccessful trials (see Fig. 3). The 8 unsuccessful trials were discarded for further analysis.

4.2 Distance to source

Since all the unsuccessful trials were discarded, the final *distance to source* represents a proxy variable for the accuracy of subjects to locate the source in close distances. The median final distance between the user hands and the source is 0.71 m. The minimum distance (most accurate trial) is 0.07 m, while the maximum distance (while keeping LOS with the source) is 7.33 m.

In order to investigate the potential differences between trials and subjects we first explored normality of the data in various grouping configurations: entire dataset, diffraction on/off, individual subject data, and individual trial data. A Lilliefors test revealed non-normality of all of the data groups, which is to be expected for positive bounded data. Following, we performed Lilliefors tests on the log-transformed dataset for various groupings. The results confirmed log-normality of all the data groupings, except for the datasets of subjects 1, 4, and 7. However, based on this analysis, it is a reasonable trade-off to continue the analysis under the assumption of log-normality of all data groupings. As such, we conducted ANOVA tests on the log-normalized data in order to explore statistical differences in the various grouping configurations. The data for all groupings is summarized in Fig. 4.

ANOVA analysis on the log-normalized data suggested that no statistically significant differences are present in *distance to source* between the conditions *Diffraction ON* and *Diffraction OFF* ($p = 0.53$). This is to be expected, as diffraction paths are unlikely audible in the presence of direct sound, and thus we do not expect it to affect the perceived location of sources. Similarly, we found no statistically significant effects of trial (specific mazes) on *distance to source* ($p = 0.81$), which is to be expected as well. However, we did observe significant differences in *distance to source* for different subjects ($F = 11.44$, $p = 2.76e-15$). Multiple pairwise comparison tests confirmed that each subject presented statistically significant differences to at least 2 other subjects with subjects presenting differences with as high as 6 other subjects. Since users were instructed to complete each trial as quickly as possible while trying to get as close to the source as possible, these results could be explained by two factors: a different individual strategy in approaching the trade-off between speed and accuracy, as well as intrinsic individual differences in their abilities to localize the source.

4.3 Navigation time

The median *total navigation time* was 28.9 s. In order to gain more insight, we followed the same strategy to analyze the *total navigation time* as explained in Sec. 4.2. The log-normality of each data group - full dataset, diffraction on/off, individual subjects, and individual trials - was confirmed, except for two exceptions (trials 23 and 30). As before, to preserve statistical power over the vast majority of the dataset, we continued the analysis under assumption of data log-normality.

The ANOVA tests revealed no impact of the modeling of diffraction acoustical effects on *total navigation time* ($p = 0.38$). This suggests that accurate simulation of this acoustic phenomena provides no advantage in the efficiency of navigating virtual environments using acoustic cues. This could be explained by several reasons, such as the presence of high reverberation levels

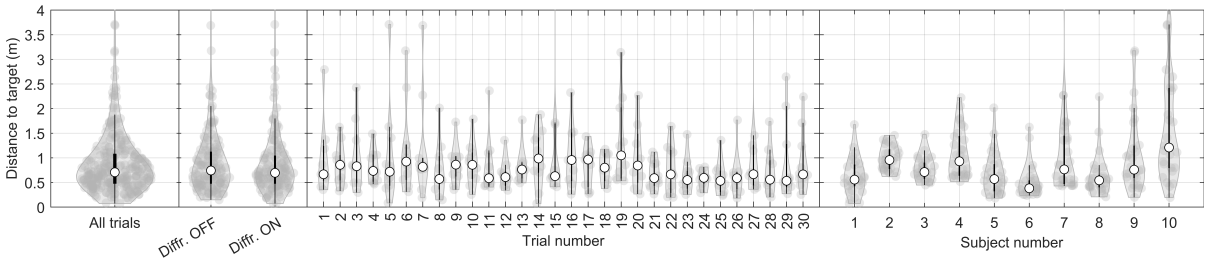

Fig. 4: *Distance to source* for various groupings of the data. White circles represent the median value, black boxplots represent interquartile ranges (IQR), and whiskers extend 1.5 times the IQR.

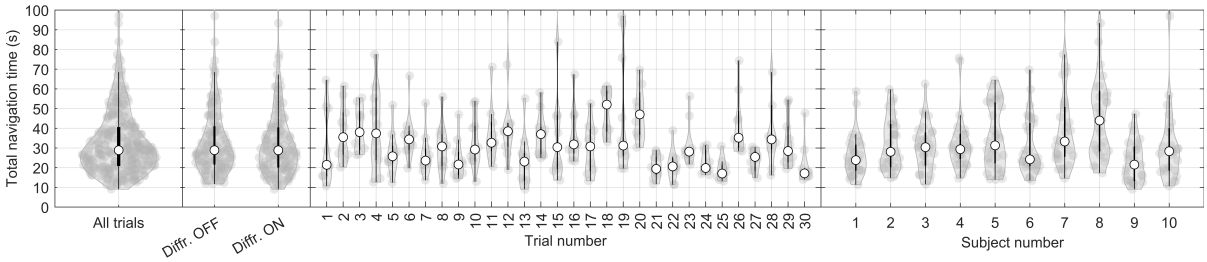

Fig. 5: *Total Navigation Time* for various groupings of the data. White circles represent the median value, black boxplots represent interquartile ranges (IQR), and whiskers extend 1.5 times the IQR.

throughout the environments, the leveraging of acoustic transmission to infer the position of sources without LOS, or the high complexity of the environments.

Contrarily, ANOVA tests revealed a statistically significant effect of both trial conditions ($F = 4.02$, $p = 6.21e-10$) and subject ($F = 5.06$, $p = 2.44e-06$). This suggests that although all the scenes were designed to present similar geometrical complexity, specific features such as room and portal distribution or the initial source and subject positions could have an important impact on the difficulty of the task. Additionally, all subjects might present varying degrees of spatial awareness as well as abilities to construct mental spatial maps, efficiently exploit acoustic information for navigation tasks, and navigate VR environments.

All the results discussed above are reported in Fig. 5. Based on individual differences, we decided to explore the results of each subject in more detail. The main goal was to determine whether specific subjects were able to leverage diffraction effects to navigate more quickly towards the sound source. Once again, we confirmed log-normality of the data for each of the subsets i.e. all trials of each individual subject for the conditions of Diffraction ON and Diffraction OFF.

The results for each individual subject are reported in Fig. 6. At first glance, it appears that only subject 8 presents a clearly visible trend, with navigation times that are higher in the condition of *Diffraction OFF*. However, ANOVA tests revealed that these differences were not statistically significant ($p = 0.09$). In fact, none of the subjects presented statistically significant differences between the conditions *Diffraction ON* and *Diffraction OFF*.

Next, we explored the *cropped navigation time* for threshold distances of 3, 2, and 1 m. The motivation behind this metric is to explore those portions of the navigation in which room acoustic effects are stronger, and thus acoustical phenomena other than direct sound would need to be leveraged to navigate efficiently. Following the same process as with previous metrics, we concluded that the results were strongly correlated to those of *total navigation time* and thus we do not report them for the sake of brevity. The use of smaller distance thresholds for the computation of *cropped navigation time* leads to an increasing number of missing

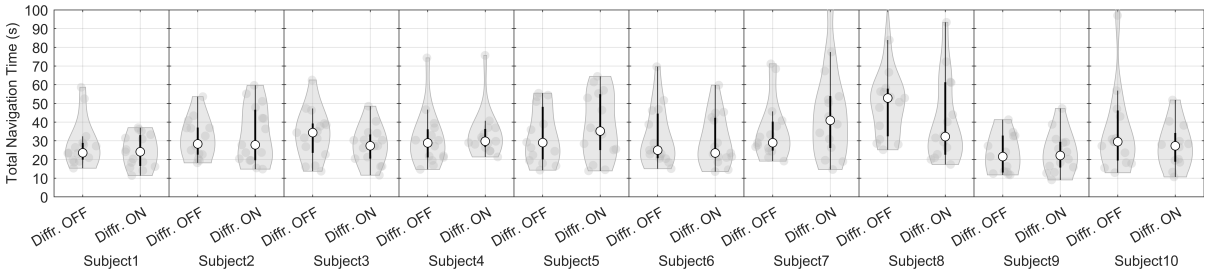

Fig. 6: *Total Navigation Time* for each individual subject in the conditions of *Diffraction ON* and *Diffraction OFF*. White circles represent the median value, black boxplots represent interquartile ranges (IQR), and whiskers extend 1.5 times the IQR.

data points, as in many cases participants did not get closer than the threshold distance.

4.4 Total path length

To investigate the navigation path length under each condition, we followed the same procedure as with the variables described earlier. Once again, we confirmed log-normality of all of the groupings. Unsurprisingly, we found that the navigation path length was strongly correlated with the *total navigation time* and *cropped navigation time*. This can be explained by the fact that since the locomotion speed is fixed, unless a subject stops and performs only head rotations, the navigation time and the path length will be proportional.

The same analysis as with previous variables was performed. The median *total path length* was 51.3 m, and it was significantly influenced by trials ($F = 5.72$, $p = 1.36e-15$) and subjects($F = 2.44$, $p = 0.0110$). The conditions of diffraction simulation did not affect the *total path length* in any meaningful way.

Regarding individual subjects (see Fig. 8), although once again subject 8 showed a trend towards decreased *total path length* in conditions of *diffraction ON*, it was not statistically significant ($p = 0.13$).

4.5 Navigation examples

As a final analysis step, we mapped all the navigated paths on top of each floorplan of the scenes. These can be utilized these to form hypotheses regarding the behaviors and potential strategies used by the participants in the navigation task. However, a detailed analysis of these is out of the scope of this paper. In Fig. 9 we present 3 examples comparing the navigation of two subjects.

4.6 Post experiment survey

After the test we conducted a survey that was completed by 9 out of 10 participants. The survey assessed the perceived diffulty of the test, the degree of motion sickness experienced, and whether subjects would participate again in a similar test. The average perceived difficulty was 2.3/5. However, 6 out of 9 subjects reported very high to extreme motion sickness. This suggests that improvements to the locomotion system should be considered in further studies. In any case, a majority of participants (6 out of 9) reported that they would participate again in a similar experiment.

5 Summary and conclusions

In this contribution we presented a framework for the conduction of navigation experiments in VR. The framework allows the creation of procedural environments of arbitrary geometrical complexity, and is paired with a real-time geometrical acoustic simulation engine that presents head-tracked audio in 6 DoF. We conducted a pilot experiment in which we asked listeners to locate an audio source in a remote location. From the results, we can form the following conclusions:

- Listeners are successful in navigating to a position in direct LOS with the source in the vast majority of trials (97.33%). This suggests that navigation tasks in VR could be leveraged for other applications that require acoustically guided spatial awareness (such as evacuations or emergency situations) and that the use of remote acoustic beacons in complex environments could prove successful in spatial navigation.

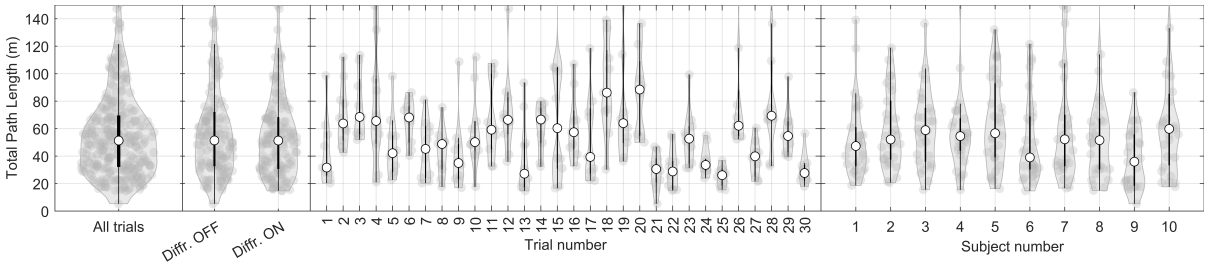

Fig. 7: *Total Path Length (TPL)* for various groupings of the data. White circles represent the median value, black boxplots represent interquartile ranges (IQR), and whiskers extend 1.5 times the IQR.

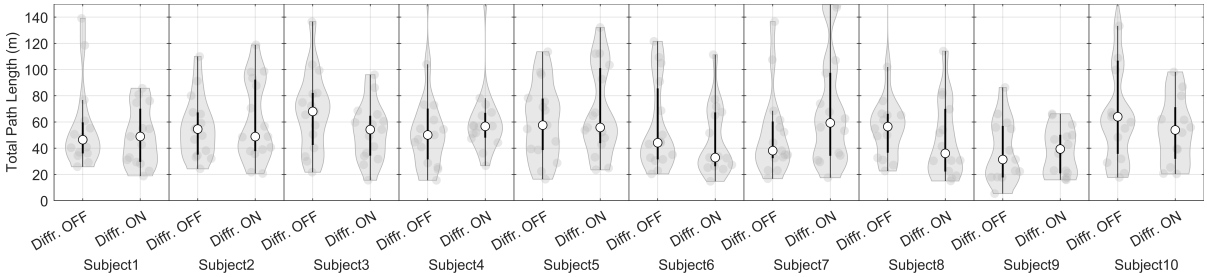

Fig. 8: *Total Path Length (TPL)* for each individual subject in the conditions of *Diffraction ON* and *Diffraction OFF*. White circles represent the median value, black boxplots represent interquartile ranges (IQR), and whiskers extend 1.5 times the IQR.

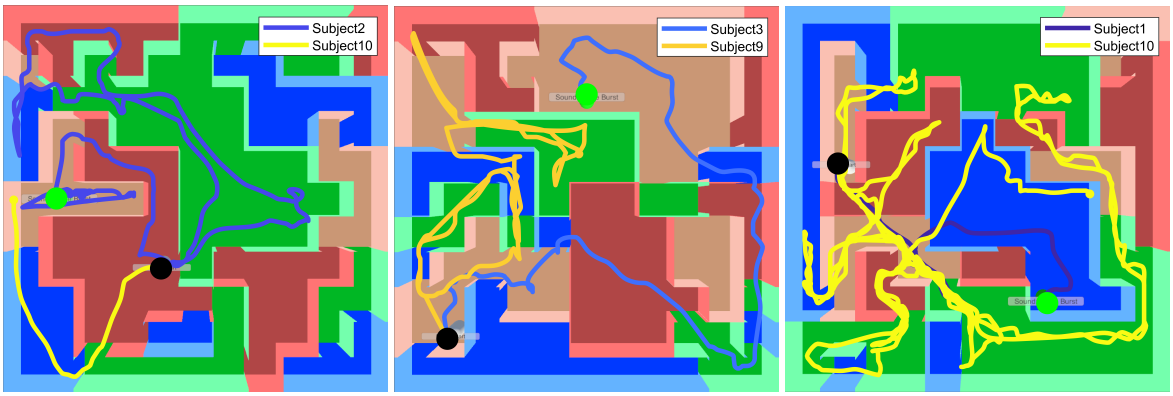

Fig. 9: Example of three paths from the experiment. (Left) The image shows two subjects successfully arriving at the final position, with Subject 10 navigating much more efficiently than subject 2. (Middle) In this case, Subject 3 completed a straightforward path, while Subject 9 did not arrive successfully to the target position, but instead marked the final position at a point with a wall between them and the source. It is likely that the transmission path was strong enough for the subject to perceive the presence of a source in that specific direction. (Right) In this case, both subjects finish at a position with LOS to the source. However, Subject 10 navigated the environment in a redundant path, and the final position represents the successful trial with the highest distance between the subject and the source (7.33 m).

- Specific scenes and subjects results in statistically significant differences in the metrics of *total navigation time*, *cropped navigation time*, and *total path length*.
- Specific subjects presented statistically significant differences in the final *distance to source*. These could be attributed either to trade-off strategies during the task or to intrinsic differences in their abilities to complete the task.
- We did not observe a statistically significant difference in navigation due to diffraction in the tested scenes. We hypothesize that the high geometrical complexity, paired with strong energy from other acoustical phenomena (reverberation level, transmission) would allow users to form spatial mental maps without the need of diffraction. Further tests in simplified and specifically crafted environments could unveil more information regarding the advantages of diffraction in specific cases.
- A large amount of participants reported strong motion sickness, likely due to the continuous locomotion at a fixed speed. This is a long standing problem in VR, and exploring solutions like including light dimming, acceleration, continuous rotation, and a proprioceptive component to induce movement could help alleviate these issues.

Future work comprises the investigation of the effect of other simulation parameters on navigation e.g. BDPT, HRTF personalization, transmission, reverb - in several environments of varying complexity. Additionally, a natural extension of such an experiment could consist of a similar task in an untethered 6 DoF setup, where listeners would be able walk in an open space while navigating the virtual environments.

Acknowledgements

We want to thank Carl Schissler for providing support on the audio simulation engine, Ben McLean and Jordan Bejar for Unity development support, and James Morrow for data collection support.

References

[1] Kolarik, A. J., Cirstea, S., Pardhan, S., and Moore, B. C., "A summary of research investigating echolocation abilities of blind and sighted humans," *Hearing research*, 310, pp. 60–68, 2014.

[2] Dodsworth, C., Norman, L., and Thaler, L., "Navigation and perception of spatial layout in virtual echo-acoustic space," *Cognition*, 197, p. 104185, 2020, doi:https://doi.org/10.1016/j.cognition.2020.104185.

[3] Picinali, L., Afonso, A., Denis, M., and Katz, B. F. G., "Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge," *International Journal of Human-Computer Studies*, 72(4), pp. 393–407, 2014, doi:10.1016/j.ijhcs.2013.12.008.

[4] Merabet, L. B., Connors, E. C., Halko, M. A., and Sánchez, J., "Teaching the Blind to Find Their Way by Playing Video Games," *PLOS ONE*, 7(9), pp. 1–5, 2012, doi:10.1371/journal.pone.0044958.

[5] Balan, O., Moldoveanu, A., and Moldoveanu, F., "Navigational audio games: an effective approach toward improving spatial contextual learning for blind people," *International Journal on Disability and Human Development*, 14(2), 2015, doi:10.1515/ijdhd-2014-0018.

[6] Lokki, T., Grohn, M., Savioja, L., and Takala, T., "A Case Study of Auditory Navigation in Virtual Acoustic Environments," in *International Conference on Auditory Display*, 2000.

[7] Levy-Tzedek, S., Maidenbaum, S., Amedi, A., and Lackner, J., "Aging and Sensory Substitution in a Virtual Navigation Task," *PLOS ONE*, 11(3), p. e0151593, 2016, doi:10.1371/journal.pone.0151593.

[8] Steffens, H., Schutte, M., and Ewert, S. D., "Acoustically driven orientation and navigation in enclosed spaces," *The Journal of the Acoustical Society of America*, 152(3), pp. 1767–1782, 2022, doi:10.1121/10.0013702.

[9] Seki, Y. and Sato, T., "A Training System of Orientation and Mobility for Blind People Using Acoustic Virtual Reality," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(1), pp. 95–104, 2011, doi:10.1109/TNSRE.2010.2064791.

[10] Zhao, Y., Bennett, C. L., Benko, H., Cutrell, E., Holz, C., Morris, M. R., and Sinclair, M., "Enabling People with Visual Impairments to Navigate Virtual Reality with a Haptic and Auditory

Cane Simulation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, doi:10.1145/3173574.3173690.

[11] Griffin, E., Picinali, L., and Scase, M., "The effectiveness of an interactive audio-tactile map for the process of cognitive mapping and recall among people with visual impairments," *Brain and Behavior*, 10(7), p. e01650, 2020, doi:https://doi.org/10.1002/brb3.1650.

[12] Cao, C., Ren, Z., Schissler, C., Manocha, D., and Zhou, K., "Interactive sound propagation with bidirectional path tracing," *ACM Transactions on Graphics (TOG)*, 35(6), pp. 1–11, 2016.

[13] Schissler, C., Mückl, G., and Calamia, P., "Fast Diffraction Pathfinding for Dynamic Sound Propagation," *ACM Trans. Graph.*, 40(4), 2021, doi:10.1145/3450626.3459751.

[14] Christian Schörkhuber, R. H., Markus Zaunschirm, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte der Akustik - DAGA 2018, 44. Deutsche Jahrestagung für Akustik*, DEGA, Munich, 2018.

[15] Chen, C., Jain, U., Schissler, C., Gari, S. V. A., Al-Halah, Z., Ithapu, V. K., Robinson, P., and Grauman, K., "Soundspaces: Audio-visual navigation in 3d environments," in *European Conference on Computer Vision*, pp. 17–36, 2020.

[16] Purushwalkam, S., Amengual Gari, S. V., Ithapu, V. K., Schissler, C., Robinson, P., Gupta, A., and Grauman, K., "Audio-visual floorplan reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1183–1192, 2021.

[17] Gao, R., Chen, C., Al-Halah, Z., Schissler, C., and Grauman, K., "Visualechoes: Spatial image representation learning through echolocation," in *European Conference on Computer Vision*, pp. 658–676, Springer, 2020.

[18] Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P. W., and Grauman, K., "Soundspaces 2.0: A simulation platform for visual-acoustic learning," *arXiv preprint arXiv:2206.08312*, 2022.

[19] Schissler, C., Loftin, C., and Manocha, D., "Acoustic classification and optimization for multimodal rendering of real-world scenes," *IEEE transactions on visualization and computer graphics*, 24(3), pp. 1246–1259, 2017.

[20] Klein, F., Gari, S. V. A., Arend, J. M., and Robinson, P. W., "Towards determining thresholds for room divergence: A pilot study on detection thresholds," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–7, 2021, doi:10.1109/I3DA48870.2021.9610876.

[21] Larsen, E., Iyer, N., Lansing, C. R., and Feng, A. S., "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, 124(1), pp. 450–461, 2008.

A Appendix: Acoustic materials

				Absorption		
Material	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz
Concrete rough	0.01	0.02	0.04	0.06	0.08	0.10
Concrete block	0.36	0.44	0.31	0.29	0.39	0.21
Wood floor	0.15	0.11	0.10	0.07	0.06	0.07
Wood thick	0.19	0.14	0.19	0.06	0.06	0.05
Wood thin	0.42	0.21	0.10	0.08	0.06	0.06
Carpet heavy padded	0.08	0.24	0.57	0.69	0.71	0.73
Carpet heavy	0.02	0.06	0.14	0.37	0.48	0.63
Carpet	0.01	0.05	0.10	0.20	0.45	0.65
Steel	0.05	0.10	0.10	0.10	0.07	0.02
Glass heavy	0.18	0.06	0.04	0.03	0.02	0.02
				Scattering		
Material	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz
Concrete rough	0.10	0.12	0.15	0.20	0.25	0.30
Concrete block	0.10	0.12	0.15	0.20	0.30	0.40
Wood floor	0.10	0.10	0.10	0.10	0.10	0.15
Wood thick	0.10	0.10	0.10	0.10	0.10	0.15
Wood thin	0.10	0.10	0.10	0.10	0.10	0.15
Carpet heavy padded	0.10	0.15	0.20	0.25	0.35	0.50
Carpet heavy	0.10	0.15	0.20	0.25	0.35	0.50
Carpet	0.10	0.10	0.15	0.20	0.30	0.45
Steel	0.10	0.10	0.10	0.10	0.10	0.10
Glass heavy	0.05	0.05	0.05	0.05	0.05	0.05
				Transmission		
Material	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz
Concrete rough	0.004	0.0079	0.0056	0.0016	0.0014	0.0005
Concrete block	0.020	0.01	0.0063	0.0035	0.00011	0.00063
Wood floor	0.071	0.025	0.0158	0.0056	0.0035	0.0016
Wood thick	0.035	0.028	0.028	0.028	0.011	0.0071
Wood thin	0.20	0.125	0.079	0.1	0.089	0.05
Carpet heavy padded	0.004	0.0079	0.0056	0.0016	0.0014	0.0005
Carpet heavy	0.004	0.0079	0.0056	0.0016	0.0014	0.0005
Carpet	0.004	0.0079	0.0056	0.0016	0.0014	0.0005
Steel	0.25	0.20	0.17	0.089	0.089	0.0056
Glass heavy	0.056	0.039	0.028	0.02	0.032	0.014

Table 3: Absorption, scattering, and transmission coefficients of the materials used in the mazes.