



Audio Engineering Society

Convention Express Paper 323

Presented at the AES 158th Convention
2025 May 22–24, Warsaw, Poland

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 The complete manuscript was not peer reviewed.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.

Real-Time Auralization for First-Person Vocal Interaction in Immersive Virtual Environments

Mauricio Flores Vargas¹, Enda Bates¹, and Rachel McDonnell¹

¹Trinity College Dublin

Correspondence should be addressed to Mauricio Flores Vargas (floresvm@tcd.ie)

ABSTRACT

Multimodal research and applications are becoming more commonplace as Virtual Reality (VR) technology integrates different sensory feedback, enabling the recreation of real spaces in an audio-visual context. Within VR experiences, numerous applications rely on the user's voice as a key element of interaction, including music performances and public speaking applications. Self-perception of our voice plays a crucial role in vocal production. When singing or speaking, our voice interacts with the acoustic properties of the environment, shaping the adjustment of vocal parameters in response to the perceived characteristics of the space.

This technical report presents a real-time auralization pipeline case study that leverages three-dimensional Spatial Impulse Responses (SIRs) for multimodal research applications in VR requiring first-person vocal interaction. It describes the impulse response creation and rendering workflow, the audio-visual integration, and addresses latency and computational considerations. The system enables users to explore acoustic spaces from various positions and orientations within a predefined area, supporting three and five Degrees of Freedom (3Dof and 5DoF) in audio-visual multimodal perception for both research and creative applications in VR.

1 Introduction

Multimodal research in Virtual Reality (VR) has gained significant attention due to the possibility of VR technology to provide both audio and visual feedback [1]. Studies have explored aspects such as immersion, presence, effects on users' physical and mental states, and audio-visual perception in Immersive Virtual Environments (IVEs) [2, 3, 4, 5, 6, 7, 8].

In the context of audio-visual fidelity, studies have shown that multisensory input can help increase realism and immersion, and that the quality level of one modality can influence the other and affect cross-modal perception [2, 9], highlighting the advantages of multimodality. However, for stimuli to be integrated into a multimodal event, sensory feedback of all modes involved must be congruent and synchronised [1, 10].

Moreover, it has enabled multisensory studies focused on practical applications that take advantage of this integration [10], including vocal adaptation in virtual room acoustics [11], a VR system for ensemble singing [12], and public speaking training in immersive environments [13].

Self-perception of one's voice significantly influences vocal production. Whether in singing or speaking, the acoustic characteristics of a space (e.g., early reflections and diffuse reverberation) play a major role in vocal perception, adaptation, and performance, particularly affecting timbre and sound level [11, 14].

Acoustic simulation using three-dimensional Spatial Impulse Responses (SIRs) captured using binaural and Ambisonic techniques is a popular method for sound auralization [12, 11]. Previous work has focused on the recreation of real-world spaces to create virtual immersive experiences for listeners. However, they are mostly captured with the source and receiver at different positions, meaning the listener (i.e., the receiver) is not also the performer (i.e., the source), which makes them unsuitable for self-perception in vocal interaction.

Building upon sound auralization and the tracking capabilities of current VR systems, Acoustic Virtual Reality (AVR) has emerged as a method in which both the sound source and the receiver (most often the user) can move freely in position and rotation while interacting with the IVE [15]. This responsive framework introduces design and computational challenges, as the auralization must update in real time to reflect changes in the source–receiver relationship, as well as the user's position and head orientation in the virtual space [16].

Within the scope of VR, several applications rely on the user's voice as an integral component of the experience and interaction, such as social and collaborative VR environments, where speech is essential for communication and engagement between users. Similarly, VR applications for public speaking have gained popularity as many people experience speech anxiety, a common concern in daily life, making these applications a useful training tool [13]. However, to the best of our knowledge, virtual room acoustics are typically not rendered, or are limited to objects within the virtual space, lacking the auralization of the user's voice, whether in individual or collaborative settings.

Perceptual congruence between sensory stimuli is essential, as mismatches can undermine the benefits of

multimodal integration, such as increased presence, ecological validity, and overall user experience [17, 18]. Therefore, addressing audio-visual matching can provide more plausible and coherent IVEs, particularly when vocal self-perception and acoustic feedback are central to the experience. Beyond the benefits of richer and congruent multisensory feedback, real-time vocal auralization opens up possibilities for practical scenarios and applications within VR, such as in vocal performance, training, and public speaking.

This work is a case study on the design and implementation of a pipeline that enables first-person real-time sound auralization in VR. The system allows users to perceive the acoustic response at different positions and directions within a defined area, providing audio-visual feedback for multimodal studies and practical applications. The pipeline has been split into two systems: the production of SIR from a first-person perspective and the audio signal processing to auralize the user's voice in real time within IVEs.

2 SIR Production

The SIR production closely follows the methods proposed by Kearney et al. [19] and Mroz et al. [20]. While the dataset from Kearney et al. [19] includes a small number of SIRs at the performer's position, they were recorded from non-adjacent locations, limiting their application to three degrees of freedom (3DoF). Therefore, the proposed system is an adaptation designed to capture various adjacent positions and directions, providing the flexibility for users to teleport or move freely within an area, thus enabling five degrees of freedom (5DoF): three rotational (roll, pitch, and yaw) and two translational (left-right and forward-back) movements.

The system is limited to two translational degrees of freedom, as the height remains fixed across all recordings. Capturing the third translational DoF would require additional layers of impulse responses recorded at varying heights to account for elevation [21], which would incur significant costs in terms of capture and reproduction. Therefore, we deemed it unnecessary, as performers typically maintain a relatively stable vertical position, and this approach has been shown to be effective in similar work [12, 20].

Pre-recording Setup To allow for translational DoF, a grid comprising 20 recording positions spaced 1 m apart was defined, providing a total area of 12 m² (4 x 3 m) for user exploration (see Figure 1). Initially, a triangular distribution, as used by [20], was considered since it is less computationally intensive; however, this arrangement results in a smaller area for the same number of positions (approximately 9 m² for 20 positions spaced 1 m apart) with an irregular shape.

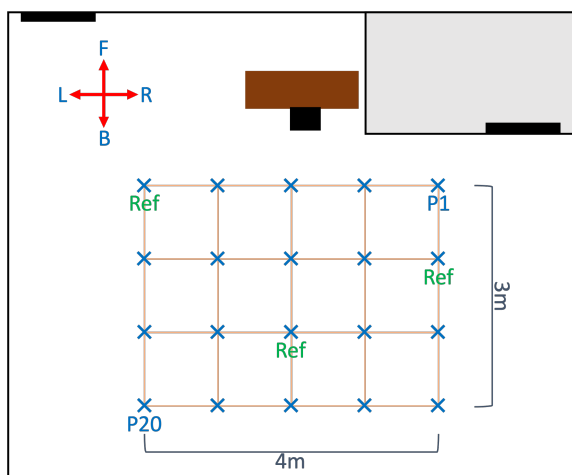


Fig. 1: Grid of recording positions.

To enable rotational DoF, four directions were considered for each position: front, right, back, and left. This approach was necessary because the speaker used (Genelec 1029A) to excite the environment has a somewhat subcardioid pattern [22], comparable to the directivity of the human voice. As a result, the acoustic response varies depending on the speaker's orientation within the acoustic space (see Figure 2). A Digital Audio Workstation (DAW) project was set up to streamline the

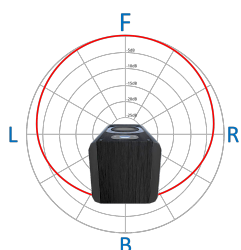


Fig. 2: Speaker directivity pattern.

recording process. The project included a mono audio track set at -20 dBFS containing 80 sine sweeps, each 20 seconds long, corresponding to 20 positions with 4 directions each. Preceding each sine sweep, the project included a voice identifier that specified the respective position and direction, followed by a sync-pop to ensure synchronisation during subsequent processing. Additionally, three reference recordings were included for comparison against the convolved signal outputs. A 5-second interval followed each sine sweep to allow the reverberation to decay before the next recording. In addition, the project included a 20-channel track to record the input signal from the 19 omnidirectional condenser microphone capsules of the Zylia ZM-1 microphone [23]. These recordings were subsequently encoded into third-order Ambisonics (3OA) using the Zylia Ambisonics Converter plugin.

Recording Process Acoustic recordings were performed with a 3OA microphone (Zylia ZM1) to capture acoustic responses within a 360° sound field. The speaker was set at a fixed height of 1.5 m, with the microphone placed 2 cm above, simulating the relative position of the mouth and ears of a person and representing a first-person perspective. The speaker output was calibrated to 85 dBA using an SPL meter placed 1 m away. An SPL reading was also taken at the microphone position as a reference. The microphone gain was then adjusted to peak around -3 dBFS to prevent clipping caused by resonances in the acoustic space. During the recording process, both the speaker and microphone were placed at each designated position. The microphone was consistently oriented to the front of the stage, while the speaker was rotated sequentially to face each of the four directions: front, right, back, and left.

SIRs Creation After completing the measurement recordings, a single 16-channel WAV file (3OA) was exported. A MATLAB script was developed to segment the file into 80 individual measurements, using the sync-pops for precise segmentation. Each channel was exported separately, resulting in 16 mono audio files per measurement. These mono files were deconvolved using Voxengo's Deconvolver [24], chosen for its reversed test tone deconvolution technique, which effectively mitigates artefacts and improves the quality of impulse responses. Processing was conducted on mono files as the software is limited to a maximum of

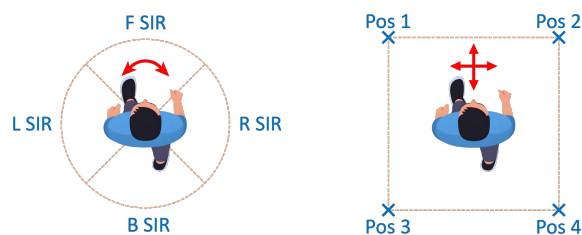


Fig. 3: Directional interpolation (left) and Translational interpolation (right).

8 channels. Finally, all deconvolved files were recombined into 16-channel SIRs.

3 Auralization Pipeline

This pipeline is responsible for auralizing the input signal, specifically the performer's voice, based on the user's position and rotation within the virtual environment. To achieve the necessary level of freedom, the system interpolates between different SIRs, which are convolved in real time with the input signal from the microphone. Therefore, the interpolation happens across different positions and four different directions (see Figure 3).

3.1 System Setup

This section outlines the system setup essential for real-time vocal auralization. It encompasses the signal flow configuration, the amplitude compensation and gain calibration procedures, and the processing and alignment of SIRs for effective latency management.

DAW Project Setup The audio signal processing was handled within Reaper DAW [25]. The project was organised using folder tracks. Each folder track corresponded to one position on the area grid and contained four 16-channel audio tracks representing each direction (front, right, back, left). Each track corresponding to a direction included a time adjustment delay VST plugin and a multichannel convolver VST plugin loaded with the corresponding SIRs as inserts. The microphone input was assigned to a single 16-channel track containing an Ambisonic encoder VST plugin to encode the mono signal into 3OA, which was then routed to the directional tracks of all positions for convolution. Finally, all position folder tracks were routed

to the Master track, which contained a binaural decoder VST plugin to decode the Ambisonic signal into a binaural headphone signal (see Figure 4).

Direct Signal Amplitude Compensation Noise-cancelling headphones were used to block the existing acoustic response of the physical environment. While some studies propose using open-back headphones to allow direct sound to reach the musician [26], we opted for noise-cancelling headphones based on the fact that, unless users were in an acoustically transparent space such as an anechoic chamber or heavily treated room, open-back headphones would result in hearing a mix of both the virtual auralization and the physical room's acoustic response. However, using headphones, especially noise-cancelling ones, results in a perceived amplitude loss of the user's voice. To account for this, the amplitude loss was measured by comparing white noise recordings made with a Sennheiser KU100 dummy head microphone, with and without headphones placed on its ears, and a speaker positioned 1 m away. The input of the microphone was then fed directly to the headphones via the audio interface's direct monitoring feature, which allows listening to the input signal with near-zero latency. The monitoring level was adjusted based on the measured amplitude loss to compensate for the attenuation and ensure that the direct voice signal reproduced over the headphones matched the perceived level without headphones.

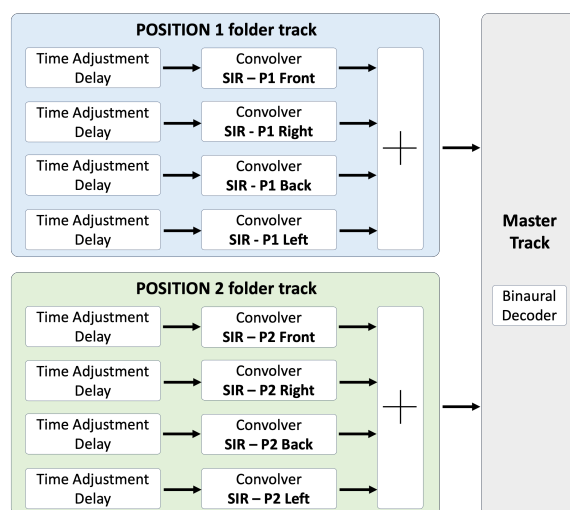


Fig. 4: Signal flow setup for two of the twenty folder tracks in the Reaper project.

Latency Compensation One of the biggest challenges in self-produced sound processing is the inherent latency between producing a sound and hearing its processed output. This latency is not fixed and arises from multiple factors, including the audio interface and the system’s digital processing and buffering before it reaches the headphone output. Therefore, we aimed for the lowest buffer size possible without performance glitches or dropouts.

Since the direct sound is fed directly to the user via direct monitoring, the SIRs were trimmed up to the first reflection, removing both the direct sound and the Initial Time Delay Gap (ITDG). The trimmed SIRs were then aligned within the Reaper project using a delay plugin, with the delay time compensation determined by subtracting the round-trip latency from the removed portion. The ITDG to the nearest surface (the ceiling, in this case) was calculated to accurately estimate the arrival time of the first reflection, using the following expression:

$$\Delta t = t_1 - t_0 = \frac{(L_1 + L_2) - L}{c}$$

where t_0 and t_1 are the arrival times of the direct and reflected paths respectively, L is the direct path length, $L_1 + L_2$ is the reflected path length, and $c \approx 344 \text{ m/s}$ is the speed of sound [27].

Most commonly, the first arriving reflection comes from the floor as it is the closest surface to the performer. However, similar to Kearney et al. [12], we did not consider this reflection due to its negligible significance, attributed to the shadowing effect of the singer’s torso and the voice’s directivity. Nonetheless, we acknowledge that different scenarios may require estimating the ITDG based on the floor reflection.

Input Signal Calibration Accurately reproducing the auditory experience in IVEs requires matching the loudness of the auralized signal reproduced over headphones to the perceived amplitude of the physical acoustic space. Thus, calibrating the input level ensures that the convolved signal results in a direct-to-reverberant ratio (DRR) consistent with the acoustic response of the physical space. To achieve this, white noise was played through the speaker at 85 dBA, measured at a distance of 1 m, matching the SIR production setup. With the microphone positioned 5 cm from the speaker, corresponding to its distance from the user’s

mouth during auralization, the signal level was calibrated within Reaper to match the RMS amplitude of a recorded sine sweep from the SIR production. This approach allowed the convolved signal to be reproduced at the intended direct-to-reverberant balance between the direct and auralized signals reproduced over headphones.

3.2 Real-time Signal Auralization

Following the system setup and calibration for direct monitoring and auralization, we now describe the real-time interpolation achieved by adjusting the output levels of the convolved signal with different SIRs across two independent stages: Directional and Translational interpolation.

Directional Interpolation Interpolation between the four directional SIRs (front, right, back, left) of each position was performed using constant power panning to adjust the gain of each corresponding audio track in Reaper, based on the Head Mounted Display’s (HMD) orientation relative to the IVE.

The VR experience was rendered using the Unity game engine [28]. A C# script was developed to manage the interpolation, continuously tracking the HMD’s rotation each frame, wrapping the values to a 0–360° range to avoid negative angles, and applying 4-way constant-power panning based on the yaw angle. Afterwards, the resulting four gain values were transmitted to Reaper via the OSC protocol, controlling the directional tracks. The gain G_d for each direction d was calculated using cosine-based constant power panning. Since power is proportional to the squared amplitude, and $\cos^2(\theta) + \sin^2(\theta) = 1$, assigning gains as $L(\theta) = \cos(\theta)$ and $R(\theta) = \sin(\theta)$ provides constant power panning [29].

Translational Interpolation Previous work has explored different interpolation strategies for 6DoF audio rendering [21, 30, 31]. In the translational interpolation stage of this pipeline, we opted to implement Inverse Distance Weighting (IDW) due to its flexibility with node distribution, perceptual robustness, and computational efficiency. IDW computes a weighted average of surrounding node values, where each weight is inversely proportional to the node’s distance from the listener [30]. The gain interpolation was defined as:

$$w_i = \frac{\frac{1}{d_i}}{\sum_{i=1}^M \frac{1}{d_i}}$$

where d_i is the distance from the performer to the i th impulse response position, and M is the number of contributing nodes.

A second C# script was developed to perform IDW-based interpolation between positions, using the HMD's location within the immersive virtual environment (IVE). The script dynamically tracked the user's position and computed normalised weights based on the inverse distance to active IR nodes. These values were then transmitted to Reaper via OSC to control the amplitude of the corresponding folder tracks. To improve functionality and performance, the script also limited the number of active nodes and disabled audio plugins in the Reaper project when a node was inactive, using hysteresis thresholds to prevent unintended on/off toggling.

4 Insights and Practical Considerations

Several considerations have been identified throughout the development and implementation of this work.

Time constraints and limited venue access are significant considerations when conducting acoustic recordings, as many venues operate under tight schedules. To minimise setup time, a recording grid can be pre-traced onto an acoustically transparent fabric in advance, allowing for efficient setup and requiring only the placement of the recording area on stage.

In addition to setup efficiency, the placement of the recording grid is a key consideration, as the perceived acoustic response within an acoustic space varies significantly depending on the performer's (source/listener) location. For example, on a performance stage, the centre tends to provide stronger early reflections from overhead canopies and walls, improving clarity and vocal support acoustic parameters [32, 33]. In contrast, positions toward the front of the stage present lower early reflection energy, while diffuse reverberation becomes more apparent, reducing vocal feedback but enhancing projection toward the audience [34]. Given the translational DoF allowed by the system, a recording placement that captures acoustic variation makes the most practical sense.

However, expanding the spatial scope of the system presents practical challenges, as it is inherently constrained by the number of acoustic measurements required to capture the defined exploration area, making measurements in larger spaces potentially unfeasible.

Alternative strategies may accommodate larger spaces or those with different spatial and architectural characteristics. For instance, designing the recording grid with greater spacing or irregular distribution could reduce the total number of required nodes or allow for a larger exploration area, as reasonable increments might not be perceptually salient. This approach is particularly suited to IDW interpolation, which performs well with sparse and non-uniform grids [30]. Furthermore, incorporating satellite positions or grids recorded at key spatial locations could further extend the exploration capabilities of the system, enabling direct investigation through teleportation-based locomotion within the IVE. This method could enable the exploration of acoustic variation derived from specific architectural features or spatial attributes, such as ceiling height and geometry variations.

Lastly, higher-order Ambisonic SIRs provide greater rendering accuracy but come at the cost of significantly increased computational load due to the number of impulse response channels required [30]. Although 3OA offers a balanced compromise between perceptual fidelity and computational efficiency [35], depending on system resources, alternative formats such as first-order Ambisonics or binaural SIRs may provide a suitable trade-off, delivering plausible results at a fraction of the computational cost required for convolution.

5 Discussion

While this work did not include a formal user evaluation, the pipeline was tested by several audio researchers, who reported that the system delivered a convincing spatial impression and that the interpolated auralization felt responsive and perceptually plausible for real-time vocal feedback in IVEs. Additionally, feedback from users with professional singing experience was positive, commenting on the enjoyment of the immersive singing experience and recognising its potential value for musical applications such as rehearsing and performing in virtual venues.

Moreover, the system has been successfully implemented in perceptual studies on audio-visual perception and rendering quality, enabling the integration of real-time first-person auralization into multimodal VR research, vocal practice and performances, and public speaking applications. With 3DoF interpolation, the system delivers natural and seamless auralization, while the addition of translational rendering enhances

spatial continuity, granting users greater agency to explore acoustic virtual spaces in a more immersive and dynamic way.

The proposed system supports real-time auralization through convolution by leveraging the ITDG to account for inherent system latency. Although latency constrains the minimum distance to the nearest reflective surface, it still enables accurate acoustic reproduction of realistic environments. In our case study, system latency of approximately 13.1 ms remained below the ITDG threshold for performance stages, achieving perceptually coherent results. In a room with a more standard ceiling height, the ITDG was slightly shorter than the system's latency by approximately 2.6 ms; however, this difference remained below the perceptual just noticeable difference for ITDG [27, 36], which is unlikely to affect the auditory experience. Latency can be further improved through hardware optimisation, using audio interfaces with lower latency performance and faster data transfer protocols than the Focusrite Scarlett 2i2 [37] used in this case study.

While audio latency was the main focus in the system design, it is also important to consider motion-to-data latency caused by the transfer of tracking data from the HMD to the resulting audio update within the audio rendering software. In our setup, directional and translational motion-to-data latency using a tethered Meta Quest 3 [38] was approximately 43 ms, remaining below the 60 ms threshold identified as perceptually acceptable in previous research [39], making it suitable for immersive vocal auralization.

Some areas for optimisation were identified during implementation, highlighting opportunities for performance enhancement. The system used directional SIRs for rotational interpolation, which significantly increased the number of impulse responses required at each grid position. This number could be reduced by producing 360° SIRs using an omnidirectional speaker instead. During auralization, the system could implement a directivity shaper to approximate the natural directivity of the human voice, oriented in real-time based on the user's head direction. In addition, interpolation methods present opportunities for further optimisation. IDW was suitable for our case study due to its flexibility and ability to perform gain-based interpolation without frequency-dependent features. However, alternative interpolation methods, such as dynamic SIR selection across transitions [40] or object-based rendering approaches [31], may offer better perceptual

realism and computational performance in more complex or dynamic environments.

Beyond the technical design and the benefits of richer, congruent multisensory feedback, the system demonstrates clear relevance for practical scenarios and applications within VR that leverage audio-visual integration. Among these, the development of audio-visual digital twins, which are high-fidelity virtual replicas of physical spaces [41], is a significant area of focus due to its broad range of potential applications. For instance, acoustical heritage has gained recognition as an essential component of cultural identity and memory, promoting interest in preserving the unique acoustic qualities of historical buildings and concert halls [42, 43]. Emerging digital tools and methods, such as VR and AR, enable the integration of coherent, immersive auditory scenes with the visual presentation of historical sites [43]. Therefore, the production of audio-visual digital twins, particularly of performance venues that allow for first-person vocal interaction, contributes to the preservation of these acoustic sites, capturing how performers perceived the stage's acoustic response during performances.

In addition to cultural heritage, audio-visual IVEs could also enable real-time vocal auralization for actors and singers, allowing for live performances within the metaverse. Since the rendering of acoustic features enhances both the sense of presence and the quality of vocal performance, providing performers with a coherent acoustic representation can potentially support their performance and overall experience. Acoustic parameters, such as room supportiveness and noiselessness, have been shown to influence vocal performance and preference [14]. Yet, research has found that music practice rooms often lack adequate acoustic conditions, forcing musicians to adapt their playing according to their present practice room [44]. In this context, the production of virtual audio-visual rehearsal rooms offers the potential for more acoustically appropriate rehearsal conditions, closely aligned with those of performance venues. Furthermore, similar to the effects highlighted in public speaking research, such environments may help performers reduce anxiety and better adapt to the acoustics of the actual venue.

Overall, the system demonstrates strong potential for real-time vocal auralization in IVEs, supporting both 3DoF and 5DoF. The implementation shortcomings identified highlight opportunities for further refinement

in terms of hardware optimisation and computational efficiency, while also opening pathways for practical implementation and future research. Future work will explore alternative translational interpolation methods, formal user evaluation, and broader integration into interactive or performance-based VR applications.

6 Conclusion

In this paper, we presented the design and implementation of a first-person vocal auralization system that enables real-time 5DoF vocal interaction within immersive IVEs. The pipeline uses 3OA SIRs captured from a first-person perspective to deliver 360° acoustic representation of the space. We described in detail the process of impulse response generation as well as the real-time amplitude interpolation across both rotational and translational dimensions. This system allows users to explore acoustic spaces in 5DoF, successfully supporting audio-visual perception in both multimodal research and creative applications within IVEs.

7 Acknowledgements

This work was conducted with the financial support of Science Foundation Ireland under the ADAPT Centre for Digital Content Technology (Grant No. 13/RC/2106 P2).

References

- [1] Malpica, S., Serrano, A., Allue, M., Bedia, M. G., and Masia, B., “Crossmodal perception in virtual reality,” *Multimedia Tools and Applications*, 79(5), 2020.
- [2] Martin, D., Malpica, S., Gutierrez, D., Masia, B., and Serrano, A., “Multimodality in VR: A Survey,” *ACM Computing Surveys*, 54(10s), 2022, doi:10.1145/3508361.
- [3] Kern, A. C. and Ellermeier, W., “Audio in VR: Effects of a Soundscape and Movement-Triggered Step Sounds on Presence,” *Frontiers in Robotics and AI*, 7, 2020, doi:10.3389/frobt.2020.00020.
- [4] Buetler, K. A., Penalver-Andres, J., Ozen, O., Ferriroli, L., Müri, R. M., Cazzoli, D., and Marchal-Crespo, L., ““Tricking the Brain” Using Immersive Virtual Reality: Modifying the Self-Perception Over Embodied Avatar Influences Motor Cortical Excitability and Action Initiation,” *Frontiers in Human Neuroscience*, 2022, doi:10.3389/fnhum.2021.787487.
- [5] Mahmud, M. R., Stewart, M., Cordova, A., and Quarles, J., “Auditory Feedback to Make Walking in Virtual Reality More Accessible,” in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 847–856, 2022, doi:10.1109/ISMAR55827.2022.00103.
- [6] Tajadura-Jiménez, A., Deroy, O., Marquardt, T., Bianchi-Berthouze, N., Asai, T., Kimura, T., and Kitagawa, N., “Audio-tactile cues from an object’s fall change estimates of one’s body height,” *PLOS ONE*, p. e0199354, 2018, doi:10.1371/journal.pone.0199354.
- [7] Senna, I., Maravita, A., Bolognini, N., and Parise, C. V., “The Marble-Hand Illusion,” *PLoS ONE*, 9(3), 2014, doi:10.1371/journal.pone.0091688.
- [8] Tajadura-Jiménez, A., Basia, M., Deroy, O., Fairhurst, M., Marquardt, N., and Bianchi-Berthouze, N., “As Light as Your Footsteps: Altering Walking Sounds to Change Perceived Body Weight, Emotional State and Gait,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, p. 2943–2952, 2015, doi:10.1145/2702123.2702374.
- [9] Potter, T., Cvetković, Z., and Sena, E. D., “On the Relative Importance of Visual and Spatial Audio Rendering on VR Immersion,” *Frontiers in Signal Processing*, 2022, doi:10.3389/frsip.2022.904866.
- [10] Geronazzo, M. and Serafin, S., *Sonic Interactions in Virtual Environments*, Springer VS, 2022.
- [11] Luizard, P., Steffens, J., and Weinzierl, S., “Adaptation of singers to physical and virtual room acoustics,” in *Proceedings of the International Symposium on Room Acoustics*, 2019.
- [12] Kearney, G., Daffern, H., Thresh, L., Omodudu, H., Armstrong, C., and Brereton, J., “Design of an Interactive Virtual Reality System for Ensemble Singing,” *Proceedings of the Interactive Audio Systems Symposium*, 2016.

- [13] Poeschl, S., “Virtual reality training for public speaking-A QUESTVR framework validation,” *Frontiers in ICT*, 4, 2017, ISSN 2297198X, doi: 10.3389/fict.2017.00013.
- [14] Redman, Y. G., Glasner, J. D., D’Orazio, D., and Bottalico, P., “Singing in different performance spaces: The effect of room acoustics on singers’ perception,” *The Journal of the Acoustical Society of America*, 154(4), pp. 2256–2264, 2023, ISSN 0001-4966, doi:10.1121/10.0021331.
- [15] Pind Jörgensson, F., Jeong, C.-H., Llopis, H., Kosikowski, K., and Strømman-Andersen, J., “Acoustic Virtual Reality – Methods and challenges,” in *Proceedings of BNAM 2018*, 2018.
- [16] Sauer, J., *Acoustic Virtual Reality: An Introductory Research For Application*, Ph.D. thesis, Delft University of Technology, 2020.
- [17] Larsson, P. and Västfjäll, D., “When what you hear is what you see: Presence and auditory-visual integration in virtual environments,” *Proceedings of the 10th Annual International Workshop on Presence*, 2007.
- [18] Khenak, N., Vézien, J.-M., Thery, D., and Bourdot, P., “Spatial Presence in Real and Remote Immersive Environments and the Effect of Multisensory Stimulation,” *PRESENCE: Virtual and Augmented Reality*, 27, pp. 287–308, 2018.
- [19] Kearney, G., Daffern, H., Cairns, P., Hunt, A., Lee, B., Cooper, J., Tsagkarakis, P., Rudzki, T., and Johnston, D., “Measuring the Acoustical Properties of the BBC Maida Vale Recording Studios for Virtual Reality,” *Acoustics*, 4(3), pp. 783–799, 2022, ISSN 2624-599X.
- [20] Mróz, B., Kabaciński, M., Ciotucha, T., Rumiński, A., and Żernicki, T., “Production of six-degrees-of-freedom (6DoF) navigable audio using 30 Ambisonic microphones,” in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021.
- [21] Patricio, E., “Toward Six Degrees of Freedom Audio Recording and Playback Using Multiple Ambisonics Sound Fields,” *AES 146th Convention*, 2019.
- [22] Canclini, A., Mucci, L., Antonacci, F., Sarti, A., and Tubaro, S., “A Methodology for Estimating the Radiation Pattern of a Violin During the Performance,” in *23rd European Signal Processing Conference (EUSIPCO)*, pp. 1286–1290, 2015.
- [23] sp. z o.o., Z., *Zylia ZM-1 Microphone Reference Manual v.1.5.0*, n.d., accessed: 2025-02-31.
- [24] Voxengo, “Voxengo Deconvolver Manual,” n.d.
- [25] Cockos Incorporated, “REAPER Digital Audio Workstation,” <https://www.reaper.fm/>, 2025, accessed: 2025-02-26.
- [26] Accolti, E., Aspöck, L., Yadav, M., and Vorländer, M., “Real-time auralization for performers on virtual stages,” 2023.
- [27] Epifani, E., *Perception of Initial Time Delay Gap, Reverberation time, and density of impulse response of a room*, Master’s thesis, Scuola di Ingegneria dell’Informazione, Politecnico Di Milano, 2021.
- [28] Unity Technologies, “Unity Real-Time Development Platform,” <https://unity.com/>, 2025, accessed: 2025-02-20.
- [29] Øland, A., “Loudness Concepts & Pan Laws,” 2019, accessed: 2024-06-30.
- [30] Alary, B. and Valimaki, V., “A Method for Capturing and Reproducing Directional Reverberation in Six Degrees of Freedom,” in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021.
- [31] McCormack, L., Politis, A., McKenzie, T., Hold, C., and Pulkki, V., “Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured with Multiple Ambisonic Receivers,” *Journal of the Audio Engineering Society*, 2022.
- [32] Dammerud, J. and Barron, M., “Concert hall stage acoustics from the perspective of performers and physical reality,” *Auditorium Acoustics 2008*, 2008, doi:10.25144/17503.
- [33] Meyer, J., *Introduction to Acoustics*, p. 1–22, Springer New York, 2009, doi: 10.1007/978-0-387-09517-2_1.

- [34] Ueno, K. and Tachibana, H., “A consideration on acoustic properties on concert-hall stages,” in *Proceedings of the International Symposium on Room Acoustics (ISRA 2010)*, Melbourne, Australia, 2010.
- [35] Enge, K., Frank, M., and Höldrich, R., “Listening experiment on the plausibility of acoustic modeling in virtual reality,” 2020.
- [36] Chen, C.-Y., “Delving into Hearing Threshold of the Delay Gap of Initial Reflection in a Room by Using the Response of Cortical Brainwaves,” *Applied Sciences*, 13(21), 2023, doi:10.3390/app132111856.
- [37] Focusrite Audio Engineering Ltd., “Scarlett 2i2,” 2024, accessed: 2025-02-26.
- [38] Meta Platforms, Inc., “Meta Quest 3,” 2024, accessed: 2025-02-26.
- [39] Nils, M.-K., Miranda, K., J., S. S., and Tapio, L., “measuring motion-to-sound latency in virtual acoustic rendering systems,” *journal of the audio engineering society*, 71, 2023.
- [40] Thomas, M., Nils, M.-K., Christoph, H., J., S. S., and Ville, P., “Auralization of Measured Room Transitions in Virtual Reality,” *Journal of the Audio Engineering Society*, 71, pp. 326–337, 2023.
- [41] Zou, C., Rhee, S.-Y., He, L., Chen, D., and Yang, X., “Sounds of History: A Digital Twin Approach to Musical Heritage Preservation in Virtual Museums,” *Electronics*, 13(12), 2024, doi: 10.3390/electronics13122388.
- [42] Zhu, X., Oberman, T., and Aletta, F., “Defining acoustical heritage: A qualitative approach based on expert interviews,” *Applied Acoustics*, 216, 2024, doi:10.1016/j.apacoust.2023.109754.
- [43] Katz, B. F. G., Murphy, D., and Farina, A., “The Past Has Ears (PHE): XR Explorations of Acoustic Spaces as Cultural Heritage,” in L. T. De Paolis and P. Bourdot, editors, *Augmented Reality, Virtual Reality, and Computer Graphics*, 2020.
- [44] Knöfel, B., Troge, J., and Weisheit, L., “Musicians and their Practice Rooms: What do they Think about Present Room Acoustics and What would they Prefer?” in *Euronoise 2018: Proceedings of the 11th European Congress and Exposition on Noise Control Engineering*, 2018.