

# WALK-THROUGH AURALIZATION FRAMEWORK FOR VIRTUAL REALITY ENVIRONMENTS POWERED BY GAME ENGINE ARCHITECTURES

D Castro      Wood & Grieve Engineers, Melbourne, AUSTRALIA  
E Ballesterio      London South Bank University, Department of the Built Environment, London, UK  
S Platt      Wood & Grieve Engineers, Melbourne, AUSTRALIA

## 1 ABSTRACT

The use and development of Virtual Reality (VR) has been growing steadily in recent years. Whilst most of the efforts have been focusing in the visual techniques to allow a more truthful and immersive experience, the corresponding realistic audio component has been slightly neglected. A novel approach for providing free walk-throughs in a Virtual Reality context using game engine capabilities is proposed in this paper. The latter unfolds as a Unity3D plugin combining the Oculus Audio SDK, the FMOD audio middleware and the manipulation of ambisonic room impulse responses (RIR). A 3D computer model of an existing library/venue in London has been used to demonstrate the capability of the plugin. The effect of initial guess, the distance between RIR measurement points and the user movement velocity were investigated in detail. It is found that the proposed methodology is robust and has the ability for letting the user freely move through the room within a VR scenario while experiencing the (measured or simulated) acoustic response of the space.

## 2 INTRODUCTION

Computer simulation of room acoustics has been developing since the late 60's in parallel to development of the required technology to carry it out. Since the first 2D models created by Schroeder<sup>1</sup> to the latest Round Robin in acoustic simulations<sup>2</sup>, the interest of the acoustical community has been focused toward the creation of acoustic virtual environments, i.e. aiming at the prediction and emulation techniques of acoustic behavior in virtual spaces, within a set of identified limits. This ultimately led to investigations on best-spoke audio reproduction methods; on how to recreate at best the acoustic impression of the said environment.

A significant output from virtual acoustic investigations is nowadays commonly known as auralization, defined as *"the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such way as to simulate the binaural listening experience at a given position in the modelled space"*<sup>3</sup>. Auralization is thus an artificial binaural reproduction technique that processes sound signals with Head-Related Transfer Functions (HRTFs) and synthesizes them into a dual-channel signal information (Left and Right ear channels) that can be later reproduced through headphones; a process also termed as binaural synthesis.

Auralizations are often used for prospective acoustic design of rooms as they convey the intrinsic characteristics of the space and makes them audible in a familiar way. They are a strong front-end tool allowing a hearing description of the generally back-end and highly theoretical computer prediction results. They allow a more reality-friendly approach of any kind of virtual space; may they be already existing, yet to be built or destroyed through time. Usually, the simulations have been static, providing the response of a specific room between a source and a receiver with no option for the user to freely move around.

Whilst few dynamic auralization solutions have been developed in the recent years<sup>4,5,6,7</sup>, the application herein described presents the methodology to provide an upgraded VR-friendly experience with an additional degree of freedom allowing the user to freely move within a virtual space while still experiencing the acoustics of the room in a binaural and dynamic way.

### 3 AUDIO IN VR

To properly grasp the richness of auralizations and their usefulness in the audio industry, one must first understand the common state of work for audio rendering as well as the new realm of possibilities that has been brought by the advent of VR technologies.

Audio rendering is usually associated with audiovisual activities such as cinematic or gaming content. Whilst video rendering has benefited from significant advances for the past years or decades, the standard of audio realism has remained relatively low. This is mostly due to the inherent nature of both audiovisual contents, i.e. a screen based medium. Advanced audio rendering technologies have been relatively minimized during this period due to their conflicting potential with the main medium – the video screen – as they could bring inconsistent or confusing auditory cues to the spectator, leading to a general disruption of the audiovisual immersion, e.g. strong sound localization effects for rear sound reproduction in cinematic or gaming configurations instead of ambient fill effects. However, a few improvements to 3D audio rendering were made in first-person game environments where the latter could provide tactical assistance to the player, usually being headphone-based. Such potential applications nurtured the development of better spatial sound technology.

Audio rendering in both applications is nonetheless conducted by intuitive mixing of different sound elements and effects, as one would do for a song mastering. This is conventionally more the case for cinematic content but also applies for gaming audio where it would additionally include the potential player's interactivity with the different sound elements within a certain environment. The latter mixing usually approximates the sound field by taking a crude estimation of its reverberation component and lump it to several filters for reproducing distance sound attenuation curves, frequency filtering, echoes or delays, etc. Still, such approximation of sound propagation is far from the physical world behavior, which is substantially more complex and involving a very critical human analysis of its sonic components. The approach of intuitive mixing thus overly simplifies the continuous changes of environmental acoustic effects, such as occlusion by hard boundaries, diffraction around objects, diffuse and/or specular reflections from surfaces, as well as sound absorption from boundary materials. However, the ever-increasing gaming market led to new challenges and ultimately rose the concept of pairing realistic audio to realistic video, trying to mimic at best the latter acoustic phenomena.

One of the latest advances with the most impact in the interactive industry is Virtual Reality (VR), providing the user with an increased sense of (virtual) consensus reality through a Head-Mounted Display (HMD). This new type of virtual immersion changes everything, as the user can now rotate his head in every possible direction whilst seeing a continuous visual scene updated by his movements in real-time. Such technology is sure to provide a new gateway to audio rendering applications.

Whilst VR is resolutely focused on its 'video-ready' aspect, actual audio rendering capabilities for VR are still not at the highest level they could be. Signal processing improvements have enabled the incorporation of real-time HRTF convolution of dry acoustic signals within game engine architectures<sup>8,9,10,11,12</sup> but solutions that provide accurate acoustic characteristics of spaces, in plugin format, are not commercially available. To achieve a user satisfactory immersion in VR environments, it is therefore of the utmost importance to create spatial sound that matches the visual impression – convergence of at least two senses, vision and audition. A source signal interpolated through binaural synthesis would bear perceptual cues amended by spatial components defined by the acoustic parameters of the virtual space and the HRTF characteristics being modelled; separating the sound intensities and delays that each ear needs to hear.

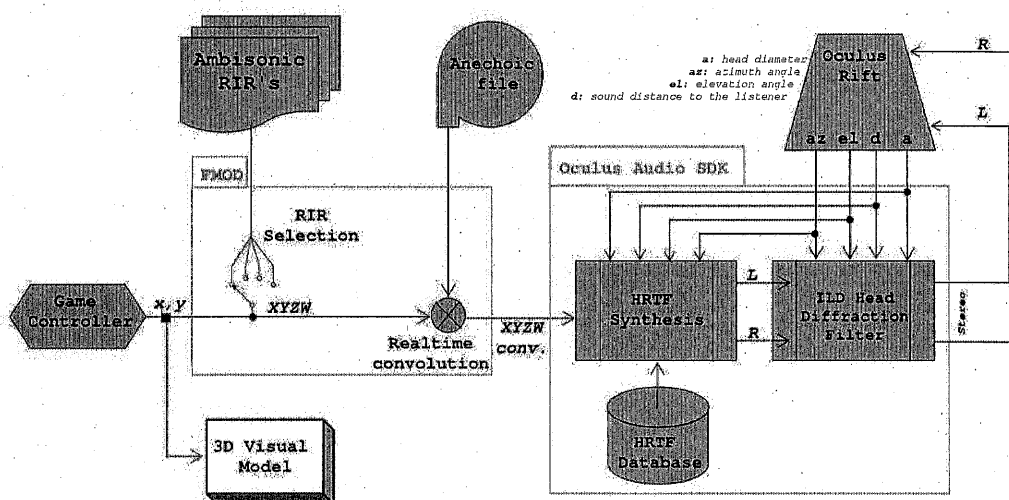
Adding movement dynamism to such real-like sound reproduction, so that it matches the user's body or head behavior, would significantly help in increasing the user immersion within the virtual environment.

The incorporation of acoustically simulated auralizations into VR thus results in a two-fold advantage, viz., (i) a more accurate capture and reproduction of the different 3D soundfields than standard audio engine capabilities (i.e. acoustic simulations vs. intuitive audio mixing), and (ii) the ability to rotate the recorded soundfields with spatial-dependent auditory cues, thus enabling audio dynamism.

## 4 SYSTEM DESIGN

## 4.1 Overview

Figure 1 displays a block diagram consisting of the most relevant elements involved in the plugin implementation. The details of the application are described in the subsequent sections and the following describes a high-level series of tasks executed by the application. Get FMOD to load the "OculusSpatializerFMOD.dll" plugin. Read a specific directory structure (with the stored RIR), and if there are quad-channel directories found, load them as a quad-channel DSP. Read the local Impulse Response directory and creates a list of them and echoes the length of time it took to load the files to the console. Loads a byte array for every ImpulseResponse called "LoadImpulseResponseChannel", which consists of 4 "FMOD.System" channels, i.e. X, Y, Z and W. Adds the amount of total Impulse Responses per channel to itself. Creates a 2-dimensional grid of positions based on the Impulse Response directory/files, and loads them into an array (see Section 3.5). All later calculations include the base offset (where the grid begins vs the scene) on an ad-hoc basis (see Section 3.3). Check that system is capable of loading and playing 5.1 mixable multichannel speaker mode via FMOD, and test it by loading and playing the Anechoic sound file on loop, through the FMOD channel group that the reverbs are loaded into. Get current position, calculate the nearest 4 grid points and use those to calculate the gain level for the reverb (see Section 3.5). Add a new Reverb DSP applied to the sound source, for each grid point that the player is closest to, with a gain-level based on the position (see Section 3.5). Get current position & vector/heading information from Oculus Rift and let FMOD's Oculus ambisonics integration calculate the 3D audio attributes. Convert current player position & heading information to FMOD-vectors, attach the updated 3D ambisonics info, and attach them to the game engine listening agent.



*Figure 1 Block diagram of the plugin architecture*

## 4.2 Acoustic input data generation

There are two main methods of obtaining the ambisonics RIR that are relevant for this application, viz., (i) by measurement or (ii) by simulation. Measurements are better used when the space already exists and it is available to measure using adapted microphone systems<sup>13</sup>. Simulations are expected to be used when the spaces do not exist and thus are not available for physical measurements.

It should be noted that the proposed methodology does not intend to be used as a live feature, meaning that it is required to pre-compute ("pre-bake") the acoustic information of the sceneries prior to the execution of the plugin within the game engine architecture. The reason for this is that the computational time required for simulating acoustically accurate RIR does not allow for a real-time walkthrough calculation. It is acknowledged that there are several commercial and research applications that suggest that an actual real-time acoustic processing is possible such as Enscape, AMD, or other research centres<sup>14</sup>, however these solutions do not provide either the full technical description of how they work or any example to trial them.

### 4.2.1 Measurement of ambisonic RIR

In order to process the audio with a 3D spatial component, the recording of the RIR needs to be carried out with an ambisonic microphone. One of the advantages of ambisonic measurements is the ability to capture and reproduce a specific sound field regardless of the final audio reproduction configuration. The ambisonic encoding (A-format/B-format) captures the 3D spatiotemporal incidences of sound, which with the corresponding equations and the appropriate post-processing allows the user to decode and manipulate the spatial sound field of the recorded audio into any designed audio configuration (D-format), e.g., stereo, full-periphonic and pantaphonic sound reproductions, binaural, 5.1, etc.

There are number of commercially available microphones that allow recording in ambisonic format. These transducers usually consist of a tetrahedral microphone array, transcribing the acoustic field by using four near-coincident microphonic capsules, resulting into a four-channel format called A-format (LFU, RFD, LBD and RBU channels). Such encoding format is however relatively complex to use for straightforward sound decoding, and is thus commonly reformatted into B-format – the four channels W, X, Y, and Z for 1st order ambisonics consisting in the combination of A-format channels (see Figure 2) – allowing a better manipulation of the sound field reproduction. The recording of the actual RIR is the same as the recording of any RIR with the requirement of using a multichannel recording device as 4 channels are being used instead of one.<sup>15,16,17</sup>

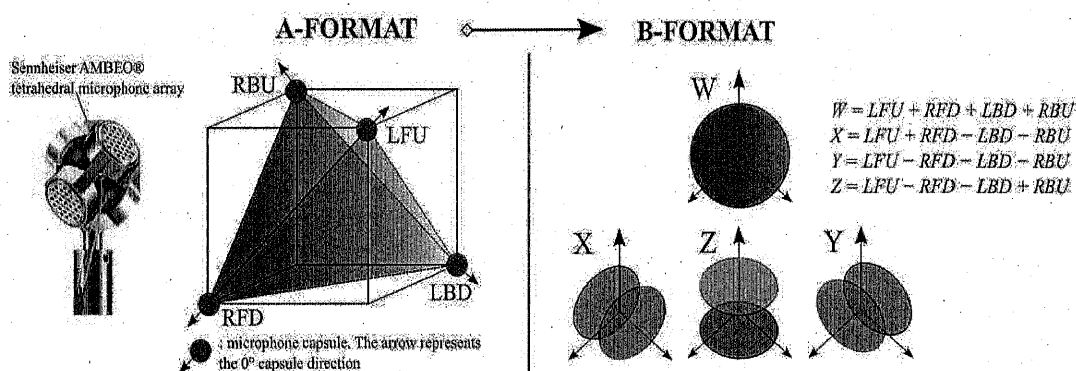


Figure 2 A-format microphone capsule configuration and subsequent B-format encoding equations<sup>18</sup> using tetrahedral microphone capsule information. LFU = Left Front Up; LBD = Left Back Down; RFD = Right Front Down; RBU = Right Back Up.

#### 4.2.2 Computer modelling

As previously mentioned, acoustic simulation of the virtual space can be also used for obtaining RIR<sup>19</sup>. This requires dedicated software that can calculate and export RIRs in ambisonic format, such as CATT-Acoustic, ODEON or EASE, enabling the creation or import of a virtual 3D CAD model of the space and the ability to assign specific acoustic properties to the model boundaries, viz., absorption and diffusion. It is to be noted that most of these software work under the statistical and geometrical formulations of acoustics, therefore limiting the extent of acoustic simulations to a certain frequency range and to airborne sound. Although it is obvious that the virtual environment would be as good as the simulations that are generated, there is enough evidence that, when used appropriately, the acoustic software can accurately reproduce the acoustics at any given point<sup>20</sup>. In the case study presented in section 5 no simulation was used and only measured RIR were used.

#### 4.3 Spatial colocation

One of the key points of the proposed plugin is the use of the superimposition theorem, where the visual and acoustical information are superimposed in space and time. The only common feature of those two otherwise independent models (visual and acoustic) is their absolute location in space (base offset coordinates). These absolute 'x' and 'y' coordinate values must be shared by the two models which will be run concurrently within Unity3D. This will provide a sense of comfort for the user as the visual and acoustical clues are expected to be in total synchronicity.

#### 4.4 FMOD & Oculus Audio SDK

The core processing of the plugin lays within the FMOD middleware, an audio rendering API which can be used for auralizations or audio spatialization purposes<sup>21</sup>. FMOD has an advanced plug-in architecture that can be used to extend the support of audio formats or to develop new output types, e.g. for streaming. The main function of this audio library is to carry out the real-time convolutions between the input RIR ambisonic files and the anechoic files, thus convolving the chosen anechoic file with the four B-format channels per position.

One of the out-coming challenges of such operation unfolds when one tries to integrate and overlay the real-time convolutions applied to a sampling grid of RIRs throughout space with the user's position and head orientation information when navigating within the model. To perform this, the information continuously recorded by the VR headset in the game engine is used, i.e. the gyroscopic angular values of azimuth and elevation along with the player's location coordinates are constantly updated and sent via UDP communication to FMOD, therefore allowing a real-time decoding of the location dependent RIRs as they are triggered by the player's movement<sup>22</sup>. This unveils another problem, consisting in the RIR selection as the player moves in between different RIR measured locations.

#### 4.5 RIR selection

A game controller is used to allow the user to move around the space, whilst the VR headset allows for a 3D dynamic perception of the said environment. The location of the user provides the x and y coordinates within the room grid, where the RIR's have been previously measured or simulated. The specific (x, y) coordinates of the user in relation to the RIR coordinates will determine which RIR are to be used at any moment in time. The plugin will always consider four RIRs corresponding to the four nearest RIRs with respect to the instantaneous user location. Each of these four RIRs is assigned with a gain weighting – equivalent to a fader value in a mixing desk – and simultaneously convolved with the anechoic file. It should be noted that, in fact, there are 16 files being simultaneously processed at any given time as each of the active RIR will consist of the four X, Y, Z and W ambisonics channels.

Figure 3 displays the RIR selection process. At any moment for any position of the user, the distances between a user location and the predefined grid locations are used to apply certain gain to each RIR.

Table 1 displays an example with specific values for a user location and its resulting gain. This gain is then applied to the real-time convolution in FMOD.

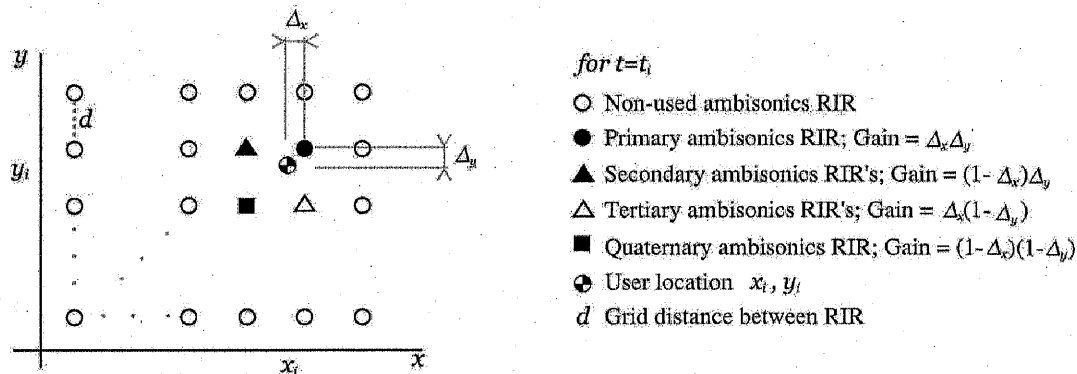


Figure 3 Example of the RIR selection based on the user location

Table 1 Example of the coordinate values for a user position and its corresponding gain value

	X (m)	Y (m)	Gain
User location	8.75	4.75	-
Primary RIR	9	5	0.5625
Secondary RIR	9	4	0.1875
Tertiary RIR	8	5	0.1875
Quaternary RIR	8	4	0.0625
			Total Gain = 1

## 5 CASE STUDY

An example was created to test the viability of the proposed plugin. As a first trial, and to minimise the potential errors due to simulation, a model based on measurements of an existing room was preferred. The Octagon at the Mile End campus of Queen Mary University of London was chosen. The main reason for such selection was that a series of ambisonic RIR measurement were already available online<sup>23</sup>, as well as enough information about the room to be able to generate the 3D visual model.

The Octagon, is one of the most significant buildings of architectural and historical importance in East London. It is a triple height, single volume, symmetric eight-sided hall with plain stock brick walls rising to prospect level, behind which the slate roof terminates in the glazed roof lantern. During its latest refurbishment (2006), the decorative plasterwork was restored, as was the polychromatic high-Victorian style color scheme. It is currently used as a meeting, exhibition and wedding venue<sup>24</sup>.

### 5.1 Measurements

A set of Ambisonic measurements were conducted using the sine sweep technique with a Genelec 8250A loudspeaker and a Soundfield SPS422B microphone<sup>13</sup>. A total of 169 receiver positions were measured.

The filenames for each position indicate the (microphone/channel/position), e.g., 'Yx04y10.wav' corresponds to the Y channel, up-down bidirectional, from the Soundfield microphone at the 4th column from the right of the 10th row from the front when facing the loudspeaker. This nomenclature is crucial to overlap both the visual model with the acoustical one. Figure 4 displays a diagram of the RIR distribution within the Octagon. These RIRs are released under the Creative Commons Attribution-Noncommercial-Share-Alike license with attribution to the Centre for Digital Music, Queen Mary University of London.

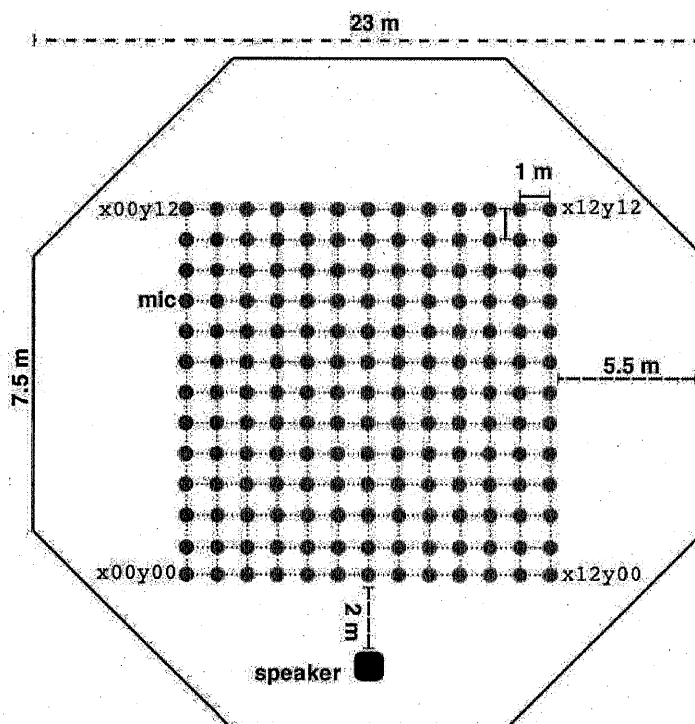


Figure 4 Example of the measured RIR grid for the Octagon room

## 5.2 Unity3D model

The required visual model was generated with Unity3D, a multipurpose game engine that can be used to create both 2D and 3D virtual environments for computers, mobile devices, and video game consoles. This software possesses a wide range of tools for both visual and aural features involved in today's gaming industry. Its scripting capabilities (C# and Javascript) also allow endless custom developments to be made alongside cross-software connections (e.g. FMOD). With such tools at one's disposition, Unity3D proves to be an ideal environment to combine audio and VR technologies

With the aim to create the 3D model of the space, the publicly available information about the venue and google photos<sup>25</sup> were used. An initial mesh was created with 3ds Max and then later exported to Unity3D where the appropriate textures and lighting were applied to the model. Figure 5 shows a comparison between the 3D model and real images from the venue.

The level of detail and refinement of the visual model is proportional to the time spent doing it, meaning that the realism applied to the visual model could be as precise and detailed as required. Notwithstanding that, the current hardware imposes a limitation on the realism of the textures and lighting processes. It is expected that with the rapid advance in technology in the forthcoming years, the visual aspect of this application could dramatically improve.

## 5.3 Output

The output of the work here presented takes form of an executable file that delivers an immersive VR experience where the user can freely walk through the space, and experience visually and aurally the room based on his/her location in the virtual space. It is noted that the executable could run in a computer with or without a VR set, however, the latter would reduce substantially the immersive experience. The following link<sup>26</sup> is an example of the of a screen capture video of the plugin run within the Octagon room using as a source file a flamenco guitar anechoic recording<sup>27</sup>.

In terms of future works, the next step will be to create a 3D model and develop the same process but for a fully virtual environment. In this instance, the main challenge would rely on the methodology that should allow us to obtain the RIRs in a semi-automatic manner, i.e., one does not expect to calculate over 200 RIR for a single room and do it manually. A scripted system is likely to be required so that it could read the coordinates, perform the calculations and store the results without user interaction. This way the expected long time of computational time won't depend on any user interaction. Once the RIR are obtained, the rest of the process will be the same as the one described in this document.

## 7 ACKNOWLEDGEMENTS

Thanks to Mark Crain from the Digital Engineering Department at Wood & Grieve Engineers for his introduction to the VR world and to Jon McCormack (Monash Uni) and Brett Paterson (Firelight Technologies) for sharing their network of contacts. Kudos to the amazing Wood & Grieve Acoustics Team for their most valuable input too.

## 8 REFERENCES

1. Schroeder MR. Digital Simulation of Sound Transmission in Reverberant Spaces. *J Acoust Soc Am*. 1970;47(2A):424-431.
2. Brinkmann F, Ackermann D, Aspöck L, Weinzierl S. First international round robin on auralization: Results of the perceptual evaluation. *J Acoust Soc Am*. 2017;141(5):3996.
3. Kleiner M, Dalenbäck B-I, Svensson P. Auralization-An Overview. *J Audio Eng Soc*. 1993;41(11):861-875.
4. Southern A, Wells J, Murphy D. Rendering walk-through auralisations using wave-based acoustical models. *Eur Signal Process Conf*. 2009;(Eusipco):715-719.
5. James A. Demonstration of CATT-Acoustic Walker module. <https://goo.gl/4xqyu1>. Published 2011. Accessed May 1, 2018.
6. Hodgson M, York N, Yang W, Bliss M. Comparison of Predicted, Measured and Auralized Sound Fields with Respect to Speech Intelligibility in Classrooms Using CATT-Acoustic and ODEON. *Acta Acust united with Acust*. 2008;94(6):883-890.
7. Lindebrink J, Nätterlund J. An engine for real-time audiovisual rendering in the building design process. *Acoust 2015 Hunt Val*. 2015:1-8.
8. Rungta A, Hill C. Dissertation Proposal : Simulation and Evaluation of Sound Propagation Effects for Virtual Environments. 2017:1-12.
9. Rungta A, Rewkowski N, Klatzky R, Lin M, Manocha D. Effects of virtual acoustics on dynamic auditory distance perception. 2017;(c).
10. Bagnoli M, Cirstea S. Dynamic geometry and material-dependent simulation of room impulse responses in a virtual gaming environment. *Proc - 2017 Int Conf Optim Electr Electron Equipment, OPTIM 2017 2017 Intl Aegean Conf Electr Mach Power Electron ACEMP 2017*. 2017:1095-1101.
11. Raghuvanshi N, Tennant J, Snyder J. Triton: Practical pre-computed sound propagation for games and virtual reality. *J Acoust Soc Am*. 2017;141:3455.
12. Katz B, Postma B, Poirier-Quinot D, Meyer J. Experience with a virtual reality auralization of Notre-Dame Cathedral. *J Acoust Soc Am*. 2017;141:3454.
13. Farina A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. *Proc AES 108th conv, Paris, Fr*. 2000;(I):1-15.
14. Fu Z hua, Li J wei. GPU-based image method for room impulse response calculation. *Multimed Tools Appl*. 2016;75(9):5205-5221.
15. Hong D, Lee T, Chung W, et al. SoundTracing : Real-time Sound Propagation Hardware Accelerator Hot Chips : A Symposium on High Performance Chips , August 20-22 , 2017. 2017;(2016):2017.
16. Guy SJ. Simulating Sound in Virtual Environments.
17. Coding B. Convention Paper. *Source*. 2010;(October):1-6.

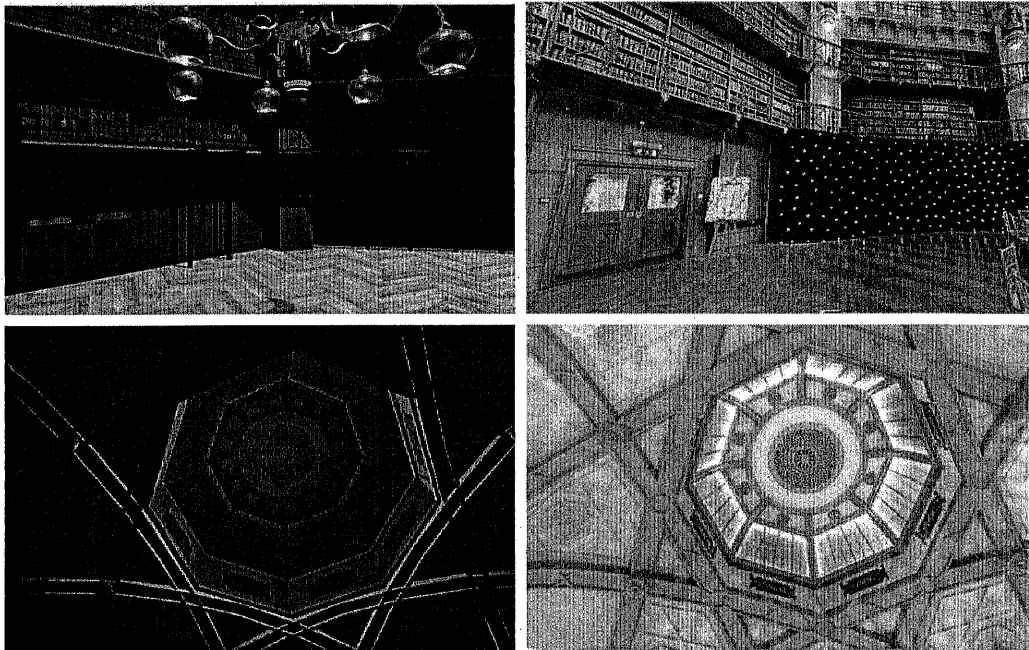


Figure 5 Comparison of the 3D model of the Octagon room (left) with the real room (right)

## 6 LIMITATIONS AND FUTURE WORKS

When using a computer simulation, the effectiveness of the result will be proportional to the quality of the auralizations and the accuracy of the model<sup>28</sup>, hence, it is required certain degree of experience in standard auralization techniques before starting using this application.

When using previously recorded RIR, it should be defined what happens when the user wanders outside the grid area. In this case, similarly to the standard case, 4 RIRs will still be used, however, the results would likely be meaningless as the user's location won't correspond to the auditory clues; this is likely to be experienced in the Octagon case when walking near the walls in the perimeter of the room. This issue could be avoided if a computer model is entirely thought with such VR reproduction as its ultimate goal, where the acoustician could define as many locations as possible to provide a good spatial coverage of the acoustic responses.

In theory, if using a computer model, it is possible to create a cloud of RIRs and having different RIRs for different heights. However, the current version of the plugin does not allow for moving in the vertical z-axis as the RIR grid is obtained at a fixed specific height (in our case at 1.7m).

This application could work with concurrent multiple sound sources – as many as required, in theory – but these sources yet must be stationary, i.e., the RIR is always between a fix source and a number of receivers; the source(s) should be always at the same static location.

It should also be mentioned that one of the constraints of this application is the speed of movement of the user in the VR environment. Due to the amount of transferred data in/out of the plugin occurring at real-time, it took a while to find an empirical compromise between acoustical comfort (i.e. something that sounds real) and the speed of movement within the VR scenario. In summary, the user can walk at an acceptable pace but is not allowed to run.

18. Hirst JM. Spatial Impression in Multichannel Surround Sound Systems. *PhD Thesis*. 2006.
19. Vorlaender M, Summers J. Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms, and Acoustic Virtual Reality. *J Acoust Soc Am*. 2008;123:4028.
20. Vorlaender M. Computer simulations in room acoustics: Concepts and uncertainties. *JAcoust Soc Am*. 2013;133:1203-1213.
21. Nuora J. Introduction to Sound Design for Virtual Reality Games implementation in Unity game engine. 2018;(March).
22. Neidhardt A, Klein F, Knoop N, Köllmer T. Flexible Python Tool for Dynamic Binaural Synthesis Applications. In: *Audio Engineering Society Convention 142*. ; 2017.
23. Sandler RS and M. Database of omnidirectional and B-format room impulse responses. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. ; 2010:165-168.
24. Memon A. An Octagonal Conundrum. *Quad (Magazine QMUL Alumni)* 13. 2004.
25. Google. The Octagon at Queen Mary Univeristy of London 360 Tour. <https://goo.gl/38B6cQ>. Accessed May 1, 2018.
26. Castro D. VR Auralization plugin Example v1. <https://goo.gl/phkxBL>. Published 2018. Accessed May 1, 2018.
27. Michio Woirgardt, Philipp Stade, Jeffrey Amankwor BB and J, Arend. Cologne University of Applied Sciences - Anechoic Recordings. 2012.
28. Witew I, Dietrich P, Vorländer M. Error and uncertainty of IACC measurements introduced by dummy head orientation using Monte Carlo simulations. *20th Int Congr Acoust*. 2010;(August):1-5.