International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

A Survey and Taxonomy of Real-Time Spatial Audio Technologies for Interactive Game Environments

Anubhav Patil[1], Prof. Sunny Nahar[2]

[1]Master of Computer Applications, Vivekanand Education Society's Institute of Technology, Mumbai, India
[2]Assistant Professor, Vivekanand Education Society's Institute of Technology, Mumbai, India

---

Abstract - *To immerse the player fully in interactive digital environments, much more is required than realistic graphics. Music and sound are also important and building believable sound environments is critically dependent on a proper simulation of sound propagation and interaction in complex virtual environments. This paper compares and contrasts all of the spatial audio technologies that are employed in existing modern game engines, with the origins of the basic research in room acoustics and spatial audio since the early 1960s. This work is a systematic survey and conceptual taxonomy, and did not include any original empirical experiments.*

Key Words: Spatial Audio, Game Development, Adaptive Rectangular Decomposition, Beam Tracing, Ray Tracing, HRTF Personalization, SAMOSA, Presence, Head-Related Transfer Function, Sound Propagation, Wave-Based Acoustics, Binaural Rendering, Perceptual Evaluation, Deep Learning Audio.

1. INTRODUCTION

Music and sound are more than just a pretty face to adorn a game; they're a part of the total experience and the experience of a player who plays in a virtual world. In many contexts, information is conveyed prior to visual information. Someone walking down a narrow stone passage can hear approaching footsteps well before they see the person coming, and an open courtyard will immediately give the sense of scale when a player walks into a large cathedral with reverberation and echo. In the absence of these auditory signals, when exaggerated or impossible to perform, the sense of presence is diminished and gives the player the realization that they are interacting with a virtual reality, and not a believable world [1, 2].

While the computer graphics have made a remarkable development in the last few decades, the spatial audio has not progressed in the same pace. Today's game engines can produce highly detailed environments via real-time rasterization and even path-traced lighting that can rival photos on consumer hardware. Many commercial games, however, still employ much the same kind of simplified models and approximations that were developed decades ago. This mismatch has been described by several researchers as an "immersion gap", where the degree of realism in the visual domain has grown considerably while the auditory domain has been comparatively little [4, 5]. In recent surveys, similar assessments of spatial audio features in virtual environments have been investigated. More specifically, Melchior et al. [3] assessed the spatial audio formats and the possibility of spatial reproduction on immersive platforms. Although their attention is mostly on documenting the existing production workflow and consumer delivery capabilities of standard virtual reality systems, this paper extends their analysis by compiling and formalizing the prior work to develop a formal comparative taxonomy of the underlying physical propagation mechanisms (geometric, wave-based, and hybrid), and proposes a conceptual implementation of a real-time semantic-acoustic pipeline (SAMOSA+ARD), that would enable the use of dynamic scene geometry, as well as a proposed structured evaluation protocol.

The challenge isn't a lack of understanding of acoustic principals. The location and characteristics of sounds are determined by human beings with a combination of perceptual cues such as Interaural Time Difference (ITD), Interaural Level Difference (ILD), and the spectral filtering effects due to the head and torso and outer ears. All these effects are summarized in the HRTF (Head-Related Transfer Function). Accurately reproducing these cues in real-time becomes increasingly difficult as the sound sources, listeners and environments are constantly changing. Consequently, spatial audio has typically been a by-product of computational efficiency and perceptual realism, and not a key design goal [6, 7].

But this is beginning to change, with a number of technological trends bringing a resurgence of interest in spatial audio research. The swift advancement of Extended Reality (XR) technologies such as Virtual Reality (VR) and Augmented Reality (AR) are significant factors. Visually, the amount of information available to the user is limited, which makes sound a more critical form of information for

navigation, awareness and environmental perception in IHMDs. Consequently, spatial audio has become a key part of the overall user experience, and not just an add-on [3, 8]. The other is the ongoing development of GPU computing. Modern graphics hardware is powerful enough to run acoustic simulations that would have been impractical and too wasteful to do real-time, previously. Until now, techniques based on wave-based methodology were only used in offline study settings, but are now becoming feasible for interactive systems and game engines [9, 10]. The third big development is the appearance of machine learning methods for spatial sound processing. In recent years, there have been significant advances in creating personalized Head-Related Transfer Functions (HRTFs) from geometric information of listeners' anatomical structures. The last few years has seen great progress in the field of personalization of Head-Related Transfer Functions (HRTFs) using geometric measurements of a listener's anatomy. They represent a convenient alternative to the conventional HRTF acquisition techniques, which involve expensive anechoic chamber recordings and specific equipment. In this paper, these developments are synthesized by means of a structured comparison of the current state of the art in spatial audio technology. The most important acoustic modelling techniques, namely geometric, wave-based and hybrid techniques will be discussed and compared in terms of performance, based on quantitative results reported in the literature. The existing commercial HRTF personalization solutions and game-engine integrations are also discussed, as are existing concepts for HRTF personalization and the demographic coverage of the available HRTF databases. Moreover, we introduce a perceptual evaluation scheme based on proven psychoacoustic principles and highlight some of the research challenges that need to be pursued in the quest for next generation immersive audio systems.

1.1 Novel Contributions

This paper synthesizes prior work and proposes four conceptual contributions:

(1) Systematic Comparative Taxonomy: A structured classification of Geometric, Wave-based, and Hybrid acoustic models with quantitative performance benchmarks derived from a systematic survey of more than 50 primary sources, identifying the immersion gap in native engine architectures [4, 13].

(2) Multimodal Simulation Framework: A proposed conceptual pipeline (SAMOSA+ARD) for fusing real-time on-device RGB-D sensor data with GPU-accelerated wave equation solvers for physically-plausible XR rendering, extending the on-device sensing framework of Xu et al. [12].

(3) Auditory Equity Analysis: A structured critique of the demographic bias in existing HRTF corpora, with recommendations for inclusive dataset construction [14, 15].

(4) Proposed Perceptual Evaluation Protocol: A proposed dual-method assessment framework for future empirical validation, integrating expert Likert-scale rating with inferential statistical verification, calibrated for spatial audio quality measurement [16, 17].

2. HISTORICAL CONTEXT AND THE ACOUSTICS OF VIRTUAL ENVIRONMENTS

The research on computational room acoustics has its origins in the concepts used in geometric optics, treating sound in a similar way to light. In this method, sound is propagated as rays which are emitted from a source, reflected from surfaces, and lose energy as they travel through space. One of the earliest and more influential works in this area was by Schroeder and Atal (1963) who created a systematic computational scheme for simulating room impulse responses. They were the precursors to much of the acoustic modelling which followed for the next few decades [18].

The game industry, however, took much easier paths to audio simulation in the late 1980s and early 90s. Most of the games used simple distance-attenuation models in which the volume of sound reduces as the distance between the sound source and the listener increases and which also sometimes involved simplified ray-casting checks to see if an object was blocking the sound path. These methods were also relatively simple to calculate and were appropriate for the hardware that was available at the time. But they failed to make any attempt to represent important acoustic features like reflections, diffraction, reverberation, and frequency-dependent absorption. Consequently, the inside of homes was frequently the same. Even though the difference in the physical materials between a large stone dungeon, a concrete warehouse, and a carpeted office is clear, each can yield almost the same acoustic responses.

In the late 1990s, there was a more significant relationship between the worlds of acoustics and interactive graphics. One of the pioneering contributions was SIGGRAPH 1998 virtual environment interactive beam-tracing system (IBS) by Funkhouser et al. They found that beam tracing was one

of the most promising methods to obtain interactive acoustic simulation of complex scenes at the time because it was efficient for simulating high order specular reflections and yet still had a reasonable computational cost. To build on this momentum, Tsingos et al. [19] proposed precomputed reverberation methods that decoupled costly computation of the geometry from the real-time rendering of the audio. This innovation enabled location-dependent reverberation effects for interactive applications, and was a significant advance towards more realistic virtual acoustics [19].

Meanwhile, commercial audio middleware had started to come into the game development marketplace. In 1998 Creative Labs enhanced its own application programming interface for environmental reverberation by launching Environmental Audio Extensions (EAX), which made the process of adding the effect to audio possible by utilizing pre-defined acoustic presets. EAX did not model the acoustics from real geometry, but instead used manually set parameters; even so, it set the precedent that games should have some sort of environmental modelling of audio, if the implementation was relatively simple [20].

This phase of the game sound evolution took place from 2005 to 2015.This is another major era in game audio technology development, so the 2005-2015 time frame is the period of evolution. This decade saw the arrival of a number of important middleware platforms, such as Audiokinetic's Wwise, Valve's Steam Audio (with research from Intel's RealSense projects) and Firelight Technologies' FMOD. These systems slowly evolved the development possibilities, enabling the creation of more sophisticated applications to simulate propagation, environment and spatial audio rendering. As time went on, they were expanded to include elements like real-time acoustics from ray-tracing, dynamic occlusion modelling, and HRTF-based binaural rendering, which has brought game sounds closer to physically-informed acoustics than ever before [21].

2.1 Geometric Acoustics: Image Source Method

One of the most prominent and early developed methods for simulating sound reflections in enclosed environments is the Image Source Method (ISM). The method was first introduced by Allen and Berkley (1979) and subsequently expanded to three dimensional spaces by Borish (1984), and models reflections using virtual copies of the sound source. The virtual, or "image," sources are created when the original source is reflected on reflective surfaces in the environment. The resultant room impulse response is then computed as the sum of the contributions to the sound from the real source and all of the valid image sources through a specified reflection order [22, 23].

The major benefit of ISM is the geometric accuracy. The method yields very accurate results for the reflection paths that it correctly models and accurately represents the physical behaviour of specular sound reflections. The accuracy is, however, costly because of the amount of computation required. The larger a scene's surface count, the more image sources a scene can have. The approximation of the number of image sources that will be generated at reflection order r for an environment containing N_w planar reflective surfaces is O(N_w^r) = N_w \sum_{i=1}^{r} (N_w - 1)^{i-1}. This formulation assumes that walls are flat and perfectly reflecting, and that there are no occlusions, wall boundaries, and wave diffraction. This results in that many computed image sources are not physically blocked or invalid, and have to be pruned, which is very expensive. This formula may then be considered an approximation of the number of possibilities of specular reflection path and may be useful only in the real time environment with low reflection orders in geometrically simple environments.

Due to this constraint, most practical applications only consider three or four orders of reflections. This is enough to recreate the early reflections which are important to spatial perception and source localization. But higher order reflections, which play an important role in late reverberation and the overall acoustic quality of the space, are often ignored. This constraint is especially challenging for large scale or geometrically complex virtual environments, such as a dense urban environment with thousands of building facades or a natural caves with non-uniform structures and non-uniform surfaces [13, 24].

However, the scalability issues do not deter ISM from being important in current acoustic simulation systems. It is particularly useful for reproducing early reflection paths accurately, which is crucial for producing the direct path and near-direct path acoustic cues that are the most perceptually important for listeners. Many current hybrid pipelines use ISM for modelling low-order reflections, but use other methods for the late-reverberation portion of the sound field such as stochastic ray-tracing or statistical reverberation models to approximate the later components in these reflections [25].

2.2 Ray Tracing Methods

The Ray Tracing Method (RTM) is a technique adapting the principles of ray tracing in optics, for the simulation of propagation of sound in an environment. RTM does not

explicitly calculate each and every reflection path, but rather it shoots a massive amount of "rays" of sound particles from the point of origin and traces their interaction with nearby surfaces. Evacuated rays transport an amount of acoustic energy which decreases with their travel through distance attenuation and energy losses due to surface absorption. If a ray arrives at the listener, or crosses a user-defined detection volume surrounding the listener, then it is added to the resultant impulse response [26, 27].

RTM is much more applicable to environments with complex or irregular geometry as compared to the Image Source Method. Its computational cost is essentially not influenced by the number of possible combinations of reflection of the rays, but is only influenced by the number of rays being traced. This results in performance that scales more predictably and thus allows the technique to be used in large virtual scenes like urban environments, industrial installations or large natural landscapes, where image-source methods are too computationally expensive.

Nevertheless, there are some drawbacks to the method. Since RTM is a Monte Carlo simulation, the accuracy of the simulation will depend greatly on the number of rays generated. If the number of samples is not adequate, there will be statistical noise in the impulse response, especially in the late reverberation region where there are many reflections from the same source but each reflects a small amount of energy. To get a smooth and reliable result one may need to trace a very large number of rays, thus adding to the computational cost.

Another drawback is the fact that some reflection paths can be overlooked. RTM only propagates a subset of all possible propagation paths, so some acoustically significant specular reflections might not be traced. This is known as the "missing-path" problem, and can result in noticeable differences in the sound field audible to listeners, particularly in settings where the reflections are loud and distinct. In these situations listeners may hear an echo or lack of consistency in what they hear, but may not completely suspect the method is not working efficiently [28].

Despite these challenges, ray tracing remains one of the most widely used approaches for real-time acoustic simulation. Its ability to handle large and geometrically complex environments has made it a key component of modern game engines and spatial audio middleware. Many contemporary systems further improve its performance and accuracy by combining ray tracing with complementary techniques, creating hybrid acoustic models that balance computational efficiency with perceptual realism.

2.3 Beam Tracing

Using deterministic pyramidal beams instead of stochastic rays, Beam Tracing, formalized by Funkhouser et al. (1998), is an improvement over RTM. The energy in each beam is in a coherent cone. The beam tracing method is an efficient method that reduces the amount of computation required compared to exhaustive ray casting by taking advantage of the spatial coherence of neighbouring beams that share the processing surfaces. A technique such as Binary Space Partitioning (BSP) trees or other spatial acceleration structures enable identification of surfaces intersected by each beam in a short amount of time [13, 29].

The main benefit of beam tracing is that it can produce the second, third and higher order specular reflections at interactive rate for moderate scene complexity. The original implementation by Funkhouser was successful in the auralization of architectural scenes having several hundred polygons. There have been more recent implementations that extend this to tens of thousands of polygons, based on hierarchical visibility pruning and GPU parallelization [30, 31].

3. WAVE-BASED ACOUSTIC MODELLING

Geometrical acoustics is based on the assumption that sound travels as rays; this is a valid assumption when the wavelengths of the sound are small compared to the size of the geometry around which sound travels and any objects in it. In this case, sound reflections can be represented by a similar model to light rays, which allows for geometric modelling to be used efficiently to a good approximation in a large number of applications.

This assumption is not as true at lower frequencies, though. Typical indoor environments usually have frequencies below a certain threshold (physically defined as the Schroeder crossover frequency fs = 2000 sqrt(RT60 / V) where the reverberation time, RT60, is in seconds and the room volume, V, is in cubic metres), where wavelengths are comparable or greater than many architectural features. This formula is valid for a statistically uniform sound field, where the sound is spread out. With this assumption, the transition between low frequency discrete wave modes and high frequency overlapping specular reflections is taken to be at f = fs. The crossover frequency for normal domestic rooms or in small studios ranges from 300 to 500 Hz. In these spaces

with very strong definition (coupled volumes, long corridors etc.) this transition is gradual, and the formula is only approximate. At frequencies lower than this, sound does not appear to be made up of independent rays. Rather, it has been found to demonstrate complex wave phenomena like diffusion, interference and modal resonances [24, 28, 32].

Because these effects arise from the wave nature of sound, they cannot be accurately represented using purely geometric approaches. Consequently, geometrical acoustic models often struggle to reproduce low-frequency behaviour, leading to inaccuracies in the simulation of bass response, room resonances, and sound propagation around obstacles. Addressing these limitations requires the use of wave-based simulation techniques that directly model the underlying physics of sound propagation, particularly in environments where low-frequency accuracy is important.

3.1 Finite-Difference Time-Domain Methods

A wave-based approach to the acoustic simulation that is widely studied is the Finite-Difference Time-Domain (FDTD) method. In particular, FDTD is based on the 3D acoustic wave equation: \partial^2 p / \partial t^2 - c^2 \nabla^2 p = 0, with the acoustic pressure p(x, t) measured in Pascals, the speed of sound in metres per second, and t being time. The 3D Laplacian operator is \nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2. This equation is based on a linear propagation of sound in a homogeneous and lossless air medium. It is very useful for modelling low-frequency wave phenomena (diffraction and interference), but has a number of limitations: it doesn't take account of the viscosity of the air, heat losses and propagation effects that are not linear at high sound pressure levels. In contrast, instead of assuming that sound is propagated as rays, FDTD numerically solves this wave equation, by dividing the space around the ear into a regular lattice and calculating the time evolution of sound pressure at every lattice point. Each time the simulation is stepped forward, the pressure values are calculated following finite-difference equations which represent physical behaviour of sound waves.

FDTD has several main benefits over other geometric methods in that it is able to accurately capture wave phenomena that are not captured by geometric methods. As it is based on the direct modelling of sound propagation from the wave equation, it is natural that it is able to model diffraction around obstacles, constructive and destructive interference, scattering and room resonances. The method is also very flexible and allows for complex boundary conditions as well as a multitude of geometries of the environment that can be handled [32, 33].

Even though very accurate, FDTD is a very costly computational method. The resolution of the spatial grid used should be fine enough compared to the shortest wavelength simulated to ensure numerical stability and accurate simulation. This is normally done by specifying a grid size, (space) $\Delta x \approx 1/10\ \lambda$ (or $\Delta x \geq 10$ cells/λ), where λ is the smallest wavelength of interest, following the standard FDTD stability rules. But other low dispersion FDs are able to be used closer, closer to 6-8 cells/wavelength value. Hence, even a small-scale simulation, with frequencies up to 1 kHz, could entail tens of millions of grid cells, where each cell would have to be updated per simulation time step [32, 33, 34].

Numerical dispersion is a difficulty that occurs in the process of discretization and is caused by the grid. Where the speed of different frequency components of the pulse is slightly different, depending on their frequency, as they would be in a real physical environment, numerical dispersion will cause distortions which will become more pronounced with increasing frequency. These errors can be generally reduced by using a smaller spatial grid to increase the number of calculations needed or by using a higher-order finite difference scheme with an increase in the complexity of implementation [33].

Due to these restrictions, full-spectrum FDTD simulation of large-scale indoor situations is still time consuming. While the practicality of wave-based acoustic simulation has advanced greatly with today's GPUs, most current hardware is unable to accommodate real-time FDTD modelling of room-scale spaces over the entire audible range. For this reason, FDTD is most often applied in research applications, in offline acoustic analysis or in hybrid simulation approaches in place of real-time simulation [33, 35].

3.2 Adaptive Rectangular Decomposition

Traditional wave-based methods are often complex and time consuming due to their computational requirements, so Raghuvanshi et al. [9] proposed the Adaptive Rectangular Decomposition (ARD) approach to overcome these difficulties and retain the physical accuracy of the wave simulation results, but with reduced computational cost [9]. ARD was conceived as an efficient tool for simulation of low-frequency acoustic propagation in architectural settings, which is hardly feasible for Finite-Difference Time-Domain (FDTD) methods.

The fundamental idea on which ARD is based is the simple observation that in many of the indoor spaces there are few or no obstacles that block the path of the air and so it can be modeled as rectangular areas. Unlike FDTD, ARD divides the space into a set of 'sub-domains' that are rectangular and are connected together. The pressure p is expanded in the sub-domains [36] x [36] x [36] as: p(x, y, z, t) = \sum_{l, m, n} q_{l,m,n}(t) \cos(l\pi x / L_x) \cos(m\pi y / L_y) \cos(n\pi z / L_z), where the modal indices l, m, n, and the time varying modal amplitudes q_{l, m, n}(t) are defined. The expansion is in the assumption that the inner boundaries of each rectangular box are rigid (reflecting). Then pressure and velocity continuity boundary conditions are applied to the common boundaries of the adjacent sub-domains: p_i = p_j and \partial p_i / \partial n = \partial p_j / \partial n with normal n pointing out of the interface. The major drawback is that this method can only be applied to rectangular sub-domains; curved or irregular boundaries have to be subdivided in steps or staircases that can cause errors. This makes the problem of solving this propagation in a domain one of computing Discrete Cosine Transforms (DCTs), a task that is well optimized in modern GPUs.

Such an approach is extremely cost-effective in terms of computation time while maintaining the ability to model key wave phenomena like diffraction, interference and room resonances. ARD takes advantage of the mathematical structure of rectangular spaces and the efficient implementations of DCT on GPUs to achieve performance levels, which would be hard to come by with conventional wave-equation solvers.

One possible source of error is at the interfaces between the neighboring rectangular domains. In practical architectural spaces, which generally are not rectangles, the transmission of sound through shared boundaries can only be approximated and not exactly represented. As a result, there will be some error in transferring acoustic energy between neighbouring sub-domains. But experiments done by Raghuvanshi and his co-workers demonstrated that most of the errors were small and are perceptually insignificant for most of the architectural environments faced in practice [9].
The performance gain from ARD is very scene dependent and architecture dependent for GPUs. In the particular experimental situations presented by the original authors on 2009-era hardware (NVIDIA GeForce 8800 GTX), ARD showed a factor of ~200 times faster than equivalent FDTD simulations. This was a significant improvement in computational efficiency for the simple rectangular domain of moderate size (about 100-fold less) allowing low-frequency room impulse responses to be generated in near real-time. Many subsequent studies and optimizations for the GPU have boosted the efficiency, making ARD one of the most convenient wave-based methods for interactive simulation [9, 37].

Hence, ARD is seen as a vital link between the high level of accuracy of the wave solvers, but also as one of the necessary criteria for real-time applications. It shows that, by using the principles of intelligent domain decomposition and using numerical techniques that are accelerated by hardware, physically based acoustic simulation can be made much more useful and more viable, and much more useful for more advanced spatial audio systems in games, virtual reality, and architectural acoustics.

3.3 Precomputed Wave Simulation

The more accurate wave-based acoustic simulation methods are only applicable if most of the computation is performed at a precomputation stage, outside the run time. On the basis of the previous work of Adaptive Rectangular Decomposition (ARD), Raghuvanshi et al. [37] proposed a method allowing for the acoustic rendering of a dynamic listener and/or sound source position with high realism by employing precomputed acoustic data [37].

The main innovation in this method is the preparation of a seven dimensional (7D) acoustic transfer function. The transfer function is the propagation of sound in an environment as a function of three coordinates of the source position, three coordinates of the listener position and one frequency dimension. That is, it stores the acoustical connection between any source and listener position in a certain scene, over a frequency range. This information is calculated beforehand, so during the actual running of the system, the system doesn't need to solve the complex wave equations.

The 7D acoustic field is created and saved as a compressed data structure. The required acoustic response can then be computed during run time, through a query of this database, and by performing some interpolation between cached precomputed responses. These operations are also easily performed in the computer, so that very little processing overhead is required to achieve wave-accurate spatial audio. This allows for a high degree of acoustic realism to be achieved that would not be possible with interactive applications with only real-time wave simulation.

The chief drawback of this is that it requires a certain amount of constancy in the environment. The acoustic transfer function is precomputed for a given scene geometry and will not reflect changes in the geometry, if that happens, the stored data will not be valid. This means that the whole acoustic field needs to be recalculated in the field every time the walls, obstacles or significant environmental elements change. This requirement limits the use of this technique to static environments or scenes where the acoustic boundaries do not change very often.

Under this constraint, precomputed wave simulation is one of the best solutions to provide physically plausible spatial audio in an interactive system. It takes the approach of sacrificing storage space and offline computation in exchange for being efficient in real-time, and ends up achieving a compromise between acoustic realism and performance that is hard to replicate with purely real-time wave-based approaches. Precomputed acoustic fields are therefore becoming a part of the very sophisticated virtual reality, architectural acoustics or interactive simulation systems in which the geometry of the environment varies only to a small degree.

4. HYBRID ACOUSTIC MODELLING AND GAME ENGINE INTEGRATION

Many researchers have come to the consensus that neither geometric nor wave-based modelling approaches for acoustics can be used separately to meet all the needs of real-time spatial modelling. While each one has some major strengths, none of them has any major weakness that would prevent its use alone.

Geometric methods (e.g. Image Source Method, Ray Tracing) are very useful for the modelling of propagation of sound over long distances and for capturing high-frequency phenomena like specular reflections. They are also very efficient to compute and can be used in real time and virtual environments with complex needs. These techniques, however, are unable to accurately represent important low frequency phenomena like diffraction, resonance and interference in the room due to the fact that they were grounded on rays and not waves.

Wave-based methods tackle just these issues. They can directly solve the acoustic wave equation, thus faithfully representing the wide variety of acoustic phenomena, such as representing the behaviour of low-frequency sound and complex acoustic interactions. Unfortunately, this physical accuracy is very expensive to compute and would be impractical in most interactive applications and in large environments in real time, where a full-spectrum wave can be calculated [28].

The limitations of these approaches have stimulated the growing interest in research on hybrid systems, combining the best of both worlds in spatial audio. A typical approach is to partition the acoustic spectrum to different frequency bands, and use different simulation methods for each band. Wave-based methods are employed to the low-frequency components where the effects of waves are prominent, and the high-frequency components are simulated by the geometric approach, which can effectively simulate the reflection and propagation behaviour. The crossover frequency between these two regions is usually set at the Schroeder frequency fs, which is a function of the room volume and reverberation time and typically falls in the range 300-500 Hz in typical rooms, but varies according to the room volume and reverberation time [24, 28, 38].

This is a frequency domain partitioning that offers an ideal balance between the physical realism and computational efficiency. Hybrid systems can be used to model a wider variety of phenomena than either of the two methods could, when applied individually. This has led to the development of one of the most promising approaches for next generation spatial audio, geometric-wave frameworks that are hybrid, meaning that they combine both geometric and wave front methods.This has paved the way for the development of one of the most promising frameworks for the next generation of spatial audio, hybrid geometric-wave frameworks, which merge both geometric and wave front approaches in a scalable way toward more realistic and immersive sound simulation for games, virtual reality and other interactive contexts.

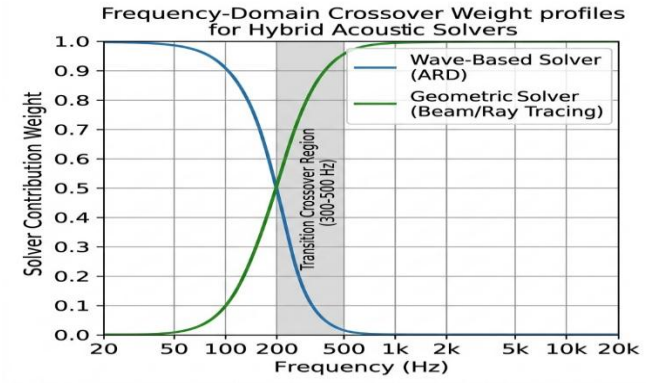

Fig -2: Crossover weight profiles for hybrid room-acoustic solvers, demonstrating the transition region (300–500 Hz) where wave-based methods (ARD) and geometric methods (Beam/Ray Tracing) are blended.

4.1 Hybrid Beam-Ray Pipelines

Since the acoustic phenomena differ with respect to the requirements of the algorithms used for simulation, a few works have investigated hybrid implementations in which multiple modelling methods are used in the same pipeline. A particularly interesting one is that proposed by Sikora et al. (2018), which combines beam tracing for early reflections and ray tracing for late reverberations, and utilizes the best attributes of both while avoiding their shortcomings [25].

The goal of this division is to draw inspiration from the differing perceptual roles of early and late acoustic energy. The first reflections (lasting less than 100 milliseconds after the direct sound) are important for spatial perception. They are crucial for directing listeners' attention to the sound's origin, for making listeners aware of the spatial context of the sound, and for giving the listener a sense of externalization of the sound source. These reflections are extremely sensitive to propagation paths, and beam tracing can help determine the paths and provide a geometrically precise representation of the paths that are important for these reflections, thus enabling these reflections to be replicated.

By contrast, late reverberation is the result of an enormous amount of higher order reflections that have bounced many times off the walls. At this point, listeners tend to be less attentive to the precise direction in which each reflection is travelling and more sensitive to how the energy fades and spreads out as a whole in the environment. Stochastic ray tracing is thus a good alternative for statistical approximation of the perceptual properties of late reverberation with much lower computation cost than deterministic approaches.

Based on these concepts, Tsingos [39] suggested a method which still further decreases the need for real-time computation by adding precomputation [39]. In this schema, the information of early echoes/late reverberation is pre-calculated offline and stored in small databases. Early reflections are modeled by precomputed image-source gradients and late reverberation is modeled by directional diffuse-energy decay profiles corresponding to different listener locations in the environment.

At runtime, the system just uses the most suitable acoustic data, depending on the location of the listener, and appropriately mixes the corresponding reflection and reverberation profiles. The method can produce realistic location-dependent acoustical effects without accessing the scene geometry directly, as the high expense of performing the geometrical calculations has already been done in pre-processing. This greatly saves the computing time from reaching high levels of perceptual realism.

In summary, the hybrid and precomputed methods provide a glimpse into the larger theme of spatial audio research, which is the use of multiple simulation techniques to achieve a compromise between the physical realism of the sound and the perceptual quality of the output, and also to accommodate computational efficiency. In interactive applications, realistic acoustics is especially important and must be delivered under tight real-time performance constraints, and such strategies have proven to be very useful.

4.2 Commercial Middleware Evaluation

With the development of more advanced spatial audio technologies, several commercial and open source middlewares have sprung up to enable state-of-the-art acoustic simulation to be used by game developers. The platforms come with pre-made functions for spatialisation, propagation and modelling of the acoustics of the environment, so the development team can add realistic sound behaviour without designing a complex acoustic solver from scratch. Of the available solutions, four platforms have seen the use of particularly widely spread throughout the gaming and immersive-media industry.

Steam Audio is a hybrid system, created by Valve Corporation and first released in 2016, based on Ambisonics, ray-traced sound propagation and HRTF-based binaural rendering. As an open source solution, the middleware can be directly integrated with the major game engines, including Unity and Unreal Engine. Steam Audio features several of the most important acoustic effects, such as sound occlusion, sound transmission through portals between connected spaces, indirect reflections and diffuse reverberation. Its propagation model, however, is mostly based on geometric acoustics and does not have the native wave-based simulation required to faithfully simulate some low frequency phenomena (diffraction and room resonances) [21].

Google introduced Resonance Audio in 2018, which is mainly used for virtual reality and augmented reality. It uses a hybrid Ambisonic rendering pipeline and HRTF-based spatialization and near-field correction techniques to enhance the realism of the nearby sound sources. The acoustic propagation system is based on the principles of geometry and acoustics, including support for the directional sound source and the frequency dependence of the surface absorption. Resonance Audio stopped

developing it in 2019, but the software development kit is still freely available and is employed in mobile VR and XR development projects because of its lightness and ease of integration [3].

Audiokinetic's Wwise is the leading audio middleware for AAA game development. In addition to the huge number of audio authoring and runtime features, Wwise features a dedicated module for Spatial Audio, added in version 2019.1 [40]. The module uses Geometric Sound Propagation (GSP) and Portal-based Room Modelling (PRM) to model the propagation of sound in a room by simulating the connections between the rooms. The HRTF-based techniques are used for spatial rendering, commonly done via a third party plug-in. Wwise does not include acoustic simulation based on waves, but offers the flexibility to add external wave-based propagation solvers and further, specialized spatial-audio technologies when needed. This flexibility has played a major role in the widespread use of this technology in large scale commercial productions [21].

FMOD Studio, created by Firelight Technologies, has features like Wwise, but with a focus on a streamlined workflow and intuitive content authoring environment. FMOD's Geometry API allows the developer to create acoustic geometry that can be passed to the FMOD Runtime library for real-time occlusion, obstruction and reflection calculations. FMOD does not include any engine for wave-based propagation but can be integrated with third party spatial-audio solutions like Steam Audio or other HRTF-based wave propagation rendering plug in solutions. The extensibility lets developers add to the core functionality of the platform while keeping to the simplicity and flexibility that game developers have found to be so well suited to FMOD.

To compare them quantitatively with empirical data from vendor documentation and academic tests with particular configurations: Wwise and FMOD allow lots of voices to be active, up to 64 or 128 channels with current consoles, and are based on simpler occlusion filters. In the case of Wwise Spatial Audio, for instance, under a representative setup, applying portal approximations [40] gives a reflection processing time between 0.2 and 0.5ms per active source. The propagation calculations of Steam Audio using CPU/GPU ray tracing require 1.5-3.0ms of CPU frame time (on a representative mid-range quad-core desktop CPU), which means that propagation calculations have a higher computational overhead, and that up to 32 active propagation sources can be simulated, with reflection and diffraction. The propagation of Resonance Audio is limited to simple boxes, and the first-order Ambisonics spatialisation engine takes less than 0.1ms per source on the desk, making it highly optimized. The values can vary widely based on the specific hardware, SDK version, scene polygon count and real voice load.

The sum total of these middleware platforms is what is currently done in game-audio production. Although they vary in architecture, user target and implementation details, they all have the same focus on efficient geometric-acoustic simulation and advanced spatialisation techniques. Remarkably, no one of the vast number of systems used today has a built-in comprehensive real-time wave-based propagation as standard, so the current problem is to achieve the acoustic realism and computational efficiency necessary for interactive applications.

Table -1: Comparative summary of acoustic modelling methods for real-time game audio, indicating frequency coverage, computational complexity, diffraction support, latency characteristics, and primary application context.

Method	Frequency Range	Complexity	Diffraction	Latency	Typical Use
Image Source Method (ISM)	Full (geometric)	O(N_w^r) (exponential, where N_w is room boundaries and r is reflection order)	No	Offline / Near-RT	Precise early reflections
Ray Tracing (RTM)	Full (geometric)	O(N_rays) (linear per ray)	No	Real-time	Complex scene reverberation
Beam Tracing	Full (geometric)	O(beams)	Limited	Real-time	High-order specular reflections
FDTD	Low (< 2 kHz)	O(M) per step (where M is total	Yes	Offline only	Research, validation

		grid cells) and O(M * T) overall			
ARD (GPU)	Low-to-mid frequency (< 1000 Hz)	O(K * M log M) (where K is sub-domains, M is modes per domain)	Yes	Near real-time	Low-freq diffraction in games
Precomputed Wave (7D)	Low-to-mid band (hybrid high-freq)	Fast spatial interpolation (scales with active source count)	Yes	Real-time (static scenes)	XR environments with fixed geometry
Hybrid Beam+Ray	Full band	Moderate	Supported only if hybridized with a wave or diffraction solver (e.g., UTD)	Real-time	AAA game environments

5. HEAD-RELATED TRANSFER FUNCTIONS AND BINAURAL RENDERING

The Head-Related Transfer Function (HRTF) refers to the transformation of an incoming sound wave caused by the physical structure of a listener's body that happens before they get to the eardrums. When sound travels around the head, some frequencies are amplified and some are muted, due to the interaction between the sound and the outer ears (pinnae) and the torso. These changes form unique spectrographic signatures which the human auditory system is able to interpret to locate a sound source and its direction.

Every person has an individual HRTF, as the shape of the head, ear and torso are different for each person. These differences in anatomy are especially critical in determining the pitch of a sound, and whether it is heard in front of or behind the listener. The spectral cues used to these dimensions are very personal and therefore strongly influenced by the anatomy of the listener to which the applied HRTF is compared [6, 7].

Spatial sound is usually created by convolving an audio signal with two HRTFs for the left and right ears in a binaural––audio system. That is, the signals x_L(t) and x_R(t) received at the listener's left and right ears are formed by convolving the source signal s(t) with the left and right Head-Related Impulse Response (HRIRs) at source direction (\theta, \phi) and distance, d: x_L(t) = s(t) * h_L(\theta, \phi, d, t) = \int_0^{\infty} s(t - \tau) h_L(\theta, \phi, d, \tau) d\tau and x_R(t) = s(t) * h_R(\theta, \phi, d, t) = \int_0^{\infty} s(t - \tau) h_R(\theta, \phi, d, \tau) d\tau. This is set up for the monophonic (mixed) source signal s(t), HRIRs h_{L,R} for the left and right speakers, and the integration delay \tau. This model assumes that the acoustic links between the sound source and the eardrums are linear, time-invariant (LTI) for each of the brief analysis frames. It is very useful for spatialized sound reproduction in headphones. The primary drawback of this method is the need to continuously interpolate between successive impulse responses (due to head rotation or source movement) which can result in processing delays or audible "clicks" when the interpolations are not sufficiently smoothed, and its computational complexity is linear with the number of sound sources being activated. If the HRTF/HRIR pair is similar to that of the listener, the resulting sound field can be extremely realistic, allowing for accurate localization and a high degree of perceived sense of sound presence, as opposed to being produced by headphones. This technique is generally considered the most perceptually accurate of all the techniques used for binaural recording presently in use.

While it is challenging to acquire individual HRTFs, there are many commercial systems that use HRTF datasets that are generic. Generalized profiles can be used to create a convincing spatial effect, but may cause a decrease in the localization accuracy. Some listeners might find it hard to determine how high or low a sound is; may mix up sounds from front and back; may hear sounds as if they were coming from inside their head and not in the outside world. This effect is called in-head localization, and is one of the most prevalent shortcomings of non-personalised binaural audio systems, and is a key driving force for further research on techniques for personalising the HRTF [6, 7].

This has led to accurate HRTF modelling being a focus of research in the fields of spatial audio, virtual reality and immersive media. Personalization of HRTFs has been perceived as an essential step towards more natural, realistic and universal 3D audio experiences.

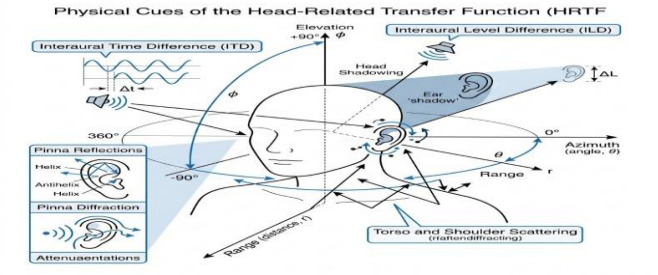

Fig -3: Coordinates (azimuth, elevation, range) and physical acoustic cues (ITD, ILD, pinna scattering, head shadowing, torso diffraction) that constitute the Head-Related Transfer Function (HRTF).

5.1 Measurement Approaches and Existing Databases

Traditionally, measuring an individual's Head-Related Transfer Function (HRTF) is a very controlled and resource intensive measurement procedure. The measurements are usually made in an anechoic chamber which is designed to reduce reflections and external noise. The participant is enveloped by a precisely designed set of loudspeakers that are located at hundreds of different azimuth and elevation angles during the procedure. Disks of tiny microphones are positioned at or close to the ear canal entrances to capture the effect of the listener's head, torso and outer ears on sound coming in from different directions. The resulting measurements are a dense set of directional impulse responses that are then later interpolated to approximate different locations of the sound sources [36].

In the past, there have been several public HRTF databases available to assist research and commercial spatial-audio applications. The most popular are the CIPIC database (45 subjects), the LISTEN database (51 subjects), the HUTUBS database (96 subjects), and the newer SONICOM database, which was particularly designed to increase the breadth of measurements and to have more subjects representative of the population [15, 36].

These databases have been vital for the advancement of spatial-audio research, but they also show an important limitation which has been gaining more and more interest in recent years. Most of the subjects of many popular HRTFs are young, male, and western European. Therefore, the measurements available do not represent all the variation in anatomy throughout the global population.

This dearth of diversity is not just a numerical issue, but a perceptual one as well. There are differences in the anatomical characteristics that affect HRTFs that exist between individuals and between demographic groups, such as head width, ear shape, torso size and even hair type. These differences change the listeners' perception and use of spectral cues, such as estimating elevation and discriminating between front and back sound sources. This makes listeners with physical characteristics significantly different from what has been measured in the databases more susceptible to localization errors when using generic HRTFs [14, 15].

An increased frequency of front-back confusion, that is, when listeners hear a sound coming from behind their head as if it were coming from the front or vice versa, is one common effect. In the broader context, the discrepancy in the HRTF chosen by the listener and their own anatomy can lead to inaccuracies in localization, a loss of the sense of an external source of sound and a decrease in realism for spatial audio experiences. The problem has been termed a "demographic bias" problem in the field of HRTF research, and it is being addressed by the need for more representative data sets and better personalization techniques.

Therefore, the aim of increasing the demographic coverage has gained significance in modern research on spatial-audio. New datasets and machine-learning-based personalization methods seek to eliminate the limitations of the current state of immersive audio technologies by providing users with a certain level of spatial perception across a wide range of ages, sex, ethnicities and anatomical variations, instead of optimizing it for only a small population.

5.2 Deep-Learning HRTF Personalization

Traditionally, HRTF measurements are expensive and time-consuming, and require special facilities, calibrated equipment and extensive data collection procedures. Researchers have been motivated to develop computational approaches that can estimate someone's HRTF from readily available physical data, such as measurements of their body. The aim is to enable personalized spatial sound without anechoic chamber recordings.

First approaches to HRTF personalization were mostly statistical-based. One of the most frequently adopted methods was to employ Principal Component Analysis (PCA) to find correlations between the HRTF characteristics and anthropometric parameters such as head width, ear size, and torso shape. These relationships were modelled, allowing researchers to estimate individual HRTFs with a reasonable accuracy without the complexity of direct measurement. These techniques were

a significant improvement but lacked adequate prediction accuracy in fine spectral details which are particularly important for accurate elevation estimation and front-back discrimination [6].

The development of machine learning techniques, especially deep learning, has greatly enhanced the prediction accuracy of HRTFs in recent years. In contrast, modern models are able to learn the complex relationship between anatomy and acoustic response, allowing a much more accurate personalisation than traditional statistical methods.

A distinguished one is GraphNF-SCA (Graph Neural Field with Spatial-Correlation Augmentation) presented by Hu et al. [11]. This framework consists of a representation of the outer ear geometry as a graph with a neural field that predicts the entire HRTF. The system explicitly models the spatial relationships in the pinna geometry, and provides state-of-the-art localization performance and directional accuracy with only non-invasive anatomical measurements [11].

In a similar manner, Hu et al. [12] proposed a transformer-based architecture, HRTFformer, which was specifically designed for the reconstruction and personalization of HRTFs. HRTFformer can reconstruct high-resolution directional HRTFs based on a small number of acoustic measurements. It is a spatially-aware attention based model, which has been shown to achieve reconstruction errors below 2 dB within mid frequency range (1–8 kHz), which is much more successful than previous interpolators [12].

One of the great benefits of these new methods is that they are all dependent on information that can be gathered with consumer equipment that is readily available. The ear photos, smartphone pictures or low-cost 3D scans are sufficient to describe the anatomy for the prediction process. This drastically decreases the barriers of conventional HRTF methods acquisition and allows for the provision of personalized spatial sound to a wide range of users without special equipment.

In addition, efforts have been made to decrease the computational complexity of using personalized HRTFs for real-time rendering. One such is MoD-ART (Modular Directional Auralization in Real-Time) by Scerbo et al. (2025). MoD-ART breaks up the HRTF into a series of modular directionally oriented components that are modifiable and processable separately. This design allows efficient multichannel binaural rendering with significantly decreased computational burden in the case of dynamic environment with continuously varying sound sources and listener positions [41].

In total, these developments clearly indicate a shift in HRTF research. There is a shift from expensive laboratory-based measures to scalable data-driven personalization measures. These techniques provide a realistic approach to creating personalized spatial audio systems that could be applied in consumer games, virtual reality, augmented reality, and other spatial applications, leveraging the advancements in computer vision, machine learning, and efficient audio rendering.

5.3 The Measurement Variation Problem

Although individual HRTFs can be successfully measured, the data obtained can still be prone to error. One of the difficulties is that the HRTF can be very sensitive to the position of the listener's head when recording it. Many of the spectral cues that aid in sound localization rely on subtle interactions between sound waves and the complex geometry of the outer ear, so even the smallest involuntary movements can make an appreciable difference in the recorded response.

It is demonstrated that small head movements during a measurement session can cause typical variations of 3–5 dB (up to 6 dB in extreme cases) in the magnitude of the recorded HRTF. These differences are more extreme at frequencies above 8 kHz, when the localization cues provided by the pinna are most important. At these frequencies, a small head movement can change the path that incoming sound waves follow through the ear sufficiently to result in a measurable change in the resulting HRTF, such as the case of a page 235 in [7] and [14].

The uncertainty in measurement does not just apply to the recording session. If a wrong HRTF is inserted into a database, then each subsequent binaural rendering which uses that data will inherit that error. Therefore, the performance of the localization system could be poor even if a seemingly individualized HRTF is being employed. These inaccuracies may be nested in the recorded measurements and may be hard to detect and even harder to correct once the data has been gathered.

There are some measures that have been proposed to mitigate this. A method is to reduce the amount of body motion when taking measurements with physical head-stabilization systems. The other uses motion tracking to capture head position during the measurement and then uses a technique of retrospective compensation to correct for any measured head motion. The third approach is to

apply repeated measurements and ensemble-averaging techniques to minimize the effect of random positional variations in the measurements, by averaging across multiple recordings to generate a representative HRTF [15, 36].

Therefore, this type of measurement variation is another significant inconsistency that is present in the available data sets and poses a major challenge in both HRTF measurement and personalization studies. This will hinder the reliability, repeatability and comparability across datasets for spatial-audio research until better standards and measurement methods are developed.

In broader terms, this issue is a reminder that HRTF personalization is not just a matter of having elaborate prediction models and large datasets, but also of having high-quality and consistent measurements that form the basis of the HRTF personalization systems. Therefore, reliable measurement is still a pivotal step towards next generation binaural audio technologies tailored to the individual.

Table -2: Comparison of major publicly available HRTF databases, indicating subject count, angular measurement resolution, demographic coverage, and data access policy.

Database	Subjects (N)	Measurement Points	Demographic Coverage	Availability
CIPIC	45	1250 directions (25 azimuths x 50 elevations)	Mostly Western adult	Public
LISTEN	51	187	Western European adult	Public
HUTUBS	96	440	German adult	Public
SONICOM	200+	792	Wider demographic scope	Public (free research-only license)
3D3A Lab (Princeton) [Choueiri, 2021]	109	2304	Mixed adult	Public
SADIE II [Armstrong et al., 2018]	20	2114	Limited	Public

6. SYSTEM WALKTHROUGH: THE CATHEDRAL SCENARIO

In order to relate the theoretical considerations covered in the preceding sections to their real-world use, the operation of a modern spatial audio system will be explored in a realistic game setting. Imagine that a player character is exploring a virtual Gothic cathedral, an approximate scale of 35 m × 15 m × 26 m, in the game. This is an extremely challenging acoustical space because of its large size, reflective stone walls and intricate architecture of the building.

These kinds of scenes are excellent stress-test for spatial audio technology as they encompass a number of acoustic phenomena that many advanced rendering systems are supposed to be able to render. The large enclosed area has long reverberation times which help to create a sense of scale and atmosphere. Multiple high order reflections are created by parallel stone walls and vaulted surfaces, giving a form to the reverberant sound field. The sound waves bend around the various pieces of architecture and massive pillars, causing diffraction effects, especially at low frequencies. Concurrently, localization is crucial, as players need to be able to pinpoint the location of sounds spread across the environment by identifying direction and distance.

On an engineering side of things, this is a tightly coupled scenario, involving the use of almost all elements of a modern spatial audio pipeline. The sources of the sounds have to be determined and all their direct propagation paths computed with respect to the listener. Reflection and reverberation systems are then used to define how the acoustic energy interacts with the geometry of the cathedral, creating early reflection and the late diffusion that is the characteristic of its acoustics. In addition to geometric methods, low-frequency propagation models can also include diffraction and room-resonance effects, which are not captured by geometric methods. Lastly, the finalized sound field needs to be rendered in a spatial manner using binaural rendering techniques, typically based on Head-Related Transfer Functions (HRTFs), which help to render a sound field as if it were coming from a particular location inside the virtual environment.

This allows to witness how the different modelling methods introduced in this paper can be applied in a comprehensive system in the context of the cathedral scenario. The example not only demonstrates the advantages of existing spatial sound technologies, but also the compromises that will need to be made between physical accuracy, computational efficiency and perceptual realism in creating an interactive experience.

6.1 Phase 1 - Geometric Initialization

When the player enters the nave, the acoustic rendering system starts by analysing the geometry of the environment in order to understand how sound is going to travel in the space. The cathedral is a space that primarily consists of large, smooth and regularly structured surfaces – a flat stone floor, walls made of ashlar masonry and a vaulted ceiling supported by the repetition of ribs – which makes it an ideal space for the application of geometric-acoustic techniques like the Image Source Method and beam tracing.

Beam tracing is implemented by the propagation engine to determine the first few important reflection paths between the active sources and listener. These can be ambient music played from the altar, sounds from the environment in the building, and the sound of the player's footsteps. When building such beam trees up to the fifth reflection order, the system is capable of determining the dominant early reflections that are most beneficial to spatial perception without losing computational efficiency.

These reflections make themselves heard nearly instantly acoustically. The first received sound is the direct sound and shortly after, a series of distinctive reflections from the surrounding stone surfaces. Reflections from the floor, ceiling and architectural details add extra layers of acoustic complexity, the characteristic flutter echoes, which rapidly repeat reflections bouncing back and forth across the nave, being produced by the parallel side walls. These reflections get added together and slowly coalesce into a thick field of reverberation, filling the inside space.

In combination these early reflections create the sound signature of the cathedral. Their time, direction and intensity give listeners much of the information about the size, shape and material texture of the environment. The direct sound, the reflections, and the reverberation without any visual input give the impression of a large stone structure and a large enclosed volume.

These cues, when correctly reproduced, play a very important role in the perception of presence. Listeners who have heard the acoustics of actual churches or cathedrals know the nature of this echo quite well – the echoes are strong, the decay time long, the reverberation massive, giving immediate auditory recognition of large, reverberant, Gothic interiors. In this manner, the early-reflection system does not only simulate sound propagation, but also conveys the architectural identity and scale of the environment, thereby immersing the player before he or she is able to perceive the visual information of the scene consciously.

6.2 Phase 2 - Wave-Based Refinement

One of the important limitations of purely geometric acoustic models is apparent as the player passes behind one of the cathedral's large load-bearing stone pillars. The vertical column from the perspective of a geometric propagation system completely obstructs the propagation line between the sound source on the altar and the listener. This implies that the model should predict an acoustic shadow region in which the sound from the altar should be greatly reduced or even be inaudible.

Actually, sound does not act as a bundle of rays. In the lower frequencies, where the wavelength is equal to or greater than the size of the obstacle, sound waves diffract around the obstacle. In the case of a stone pillar with radius a = 0.75 m (diameter 1.5 m), the significant wave bending (diffraction) is observed for the wave number k = 2 pi f / c, when it is less than 2 pi, that is, at frequencies less than about 450 Hz. At higher frequencies, sound goes into a "geometric shadow zone", causing a lot of attenuation. This means that a listener behind the pillar would still be able to hear the low frequencies of the cathedral organ (under 450 Hz), although the higher frequencies would be somewhat masked. This mixture of preserved bass energy and decreased amount of high frequency content is a well-known feature of the real world acoustic shadowing.

The behaviour is replicated by adding a low-frequency wave-based simulation to the geometric propagation model with Adaptive Rectangular Decomposition (ARD). The ARD solver does not model rays but rather models the propagation of sound waves directly in the architectural space. The interior of the cathedral is broken into a series of connected rectangular spaces, including the main nave, the side aisles divided by series of pillars and the raised triforium spaces above the arcade. For each region, the propagation of sound is calculated with efficient Discrete Cosine Transform (DCT)-based solutions of the acoustic wave equation.

These individual domains then connect acoustically via openings and shared boundaries to travel naturally

throughout the structure at low frequencies. This allows the diffraction effects, standing waves and other wave properties to be simulated that would not be possible purely geometrically.

At the time of simulation, the outputs of both methods are merged together during the run-time. The ARD solver gives the low-frequency transfer function that is representative of the diffraction around the pillar, whereas the geometric solver gives the high-frequency transfer function that represents direct sound, reflections and detailed spatial cues. These responses are mixed at a crossover frequency which is chosen based on the physical dimensions of the obstacle and the frequency range at which wave effects are meaningful.

The outcome is a more realistic sound effect. The player is not hearing an artificially deadened stretch behind the pillar, but rather the sound character changes gradually: Low frequencies remain and carry the weight of the organ while the higher frequencies are gradually lower. The behaviour is close to the actual acoustic perception and explains the reason why hybrid geometric-wave systems are frequently regarded as the most promising solution to physical plausible spatial audio rendering in complex architectural spaces.

6.3 Phase 3 - Binaural Rendering

After the complete Spatial Impulse Response (SIR) has been generated, it merges the influence of multiple stages of the acoustic modelling: the early reflections (geometric); the late reverberation (statistical); the low-frequency diffraction (wave). The binaural signals to be presented to the left and right ear are then created by convolving this unified impulse response with the listener's listening room Head-Related Transfer Function (HRTF) for each individual sound source.
This process needs to be continuous in virtual reality applications as the user rotates and moves its head. The rendering system updates the HRTF for every frame generated by head tracking (usually 90 frames per second for the modern VR platforms). For every update, the system calculates the direction of each sound source relative to the listener and interpolates between the closest measured and/or calculated HRTF directions. This way, the listener's orientation has a spatial cue to go along with it.

Imagine a player whose voice is coming from inside the cathedral, as an organist plays an organ piece from the altar. If the player is looking to a right side aisle, then the position of the organ in the auditory scene should be perceived towards the left side aisle. Moving away from the altar will, on the other hand, move the sound source forward. These changes are because the physical source doesn't move in the virtual environment, just the listener.

One of the most crucial things for convincing spatial audio is this seamless interlinkage between head motion and hearing. When humans move through an environment, they tend to expect the visual and auditory information to stay in sync with them as they go. As long as the eyes show the direction and ears show the direction, there is no problem with that. Any noticeable mismatch between the direction indicated by the eyes and the direction indicated by the ears can cause immersion to suffer and make the experience less realistic. Therefore, the correct head-tracked binaural rendering is of a crucial importance for preserving the spatial coherence in virtual environments [1, 8].

Correct binaural rendering isn't enough to create a strong sense of presence, but it serves as a starting point for adding other elements of the sense of presence. HRTF-based rendering can also be used in conjunction with a realistic propagation model, environmental acoustics and responsive head tracking, which will help to ensure that virtual sound sources are stable in the scene and that the listener feels located in a coherent and physically plausible acoustic space.

7. THE SAMOSA+ARD MULTIMODAL SIMULATION FRAMEWORK

For all geometry-dependent acoustic simulation techniques, whether using geometric acoustics or wave-based modelling, a fundamental challenge is the need for an accurate representation of the surrounding environment. This need has been met in traditional game development with manually written acoustic meshes: low polygon models with simplified geometry, designed by the level designer for the purpose of sound propagation calculations. These meshes are basically used to approximate the virtual environment and the basis for acoustic simulation, which are computational efficient.

In Extended Reality (XR) applications aimed at augmenting or interacting with the real world, the situation is much more complicated. In these cases, there is no existing acoustic mesh network as the environment is not planned for in advance. Rather, the system has to constantly sense and model the physical environment with the onboard sensors. Thus, the acoustic simulation process is now tied to the real-time understanding of the environment and the understanding of the scene.

To overcome this difficulty, Xu et al. [17] proposed SAMOSA (Scene-Aware Multimodal Acoustic Rendering), which supports real-time environmental sensing and leverages the parallel processing power of GPUs to simulate audio properties.To cope with this challenge, Xu et al. have presented SAMOSA (Scene-Aware Multimodal Acoustic Rendering), a framework that fuses real-time environmental sensing with GPU-accelerated acoustic simulation [17]. In this paper we suggest a conceptual integration containing an Adaptive Rectangular Decomposition (ARD) wave solver stage that complements the SAMOSA pipeline. Please note that this integrated SAMOSA+ARD approach is not yet in real implementation, but is proposed as an architecture.

There are three key phases of the framework. The first stage is the sensing phase, which transforms data of depth and colour images into a three dimensional point cloud of the environment around the object. This data is then fed into a CNN producing dense voxel representation of the scene. The network also does semantic classification besides the geometric reconstruction of the environment: It assigns to different areas of the environment material labels such as concrete, wood, fabric, glass and metal. These labels are then used to look up the acoustic properties, such as the absorption coefficients as a function of frequency, from a look-up table that contains acoustical properties for various materials.

The second step is the acoustic propagation modelling. The reconstructed geometry with the environment annotated is suggested to be then passed on to an appropriate simulation engine. The system can use an Adaptive Rectangular Decomposition (ARD) solver when sufficient computational power is available and the environment is not too large, to provide an accurate calculation of the low-frequency wave, and can use a hybrid beam-tracing and ray-tracing framework when the frequency is not low, to provide an accurate calculation of the propagation of the high-frequency geometry. This adaptive selection aims to achieve a compromise between physical accuracy and computing resources, ensuring interactive performance.

The last step is the audio rendering. Combined with the listener's personal Head-Related Transfer Function (HRTF), the propagation engine-generated directional impulse responses are used to create binaural audio signals that accurately represent acoustical properties of the environment and listeners' spatial perception. This gives rise to a dynamically updated sound field which adapts to modifications of user location and environment setup.

The most noteworthy contribution of SAMOSA is acoustic adaptability as a result of the change in the physical environment. The acoustic model can be changed over time during the use of the space, like when a door is opened or closed, when furniture is moved, when objects are placed on reflective surfaces or when the room is moved from one to another, etc. This capability is a key towards truly adaptive acoustic XR systems where the virtual soundscape would adapt to the changing state of the physical world.

While significant progress has been made, there are still key technological issues to be addressed. This full sensing-simulation-rendering process adds latency since data about the environment has to be collected, processed, simulated and then rendered. This end-to-end latency was evaluated on standalone XR hardware (in this case an Oculus Quest 2 equivalent device) by Xu et al. (2025) and was found to be ~70-90 ms, with depth sensing taking ~33 ms, CNN material classification taking ~15-30 ms, voxelization taking ~10 ms, ARD wave propagation step taking ~8-15 ms, and binaural convolution taking ~1 ms. The timings are hardware dependent and would be different on other devices, or by the level of granularity of the scenes. The sensing and voxelization processes are proposed to be executed asynchronously with a key frame rate of 2-5 Hz (the geometry of the room is changing slowly) at the same time as the audio spatialization and head-tracked rendering are synchronously executed at a high frequency frame loop at 90 Hz (the geometry of the room is not changing fast). This optimization can be done with room propagation parameter optimization that is dynamically updated, to maintain high interactive performance [17, 42].

In general, SAMOSA demonstrates the growing integration of computer vision, machine learning, real-time simulation and spatial rendering of audio. It allows to generate the acoustic model directly from the sensed environment, rather than be manually written in geometry, and opens the door to new future audio systems that can dynamically adapt to the real world.

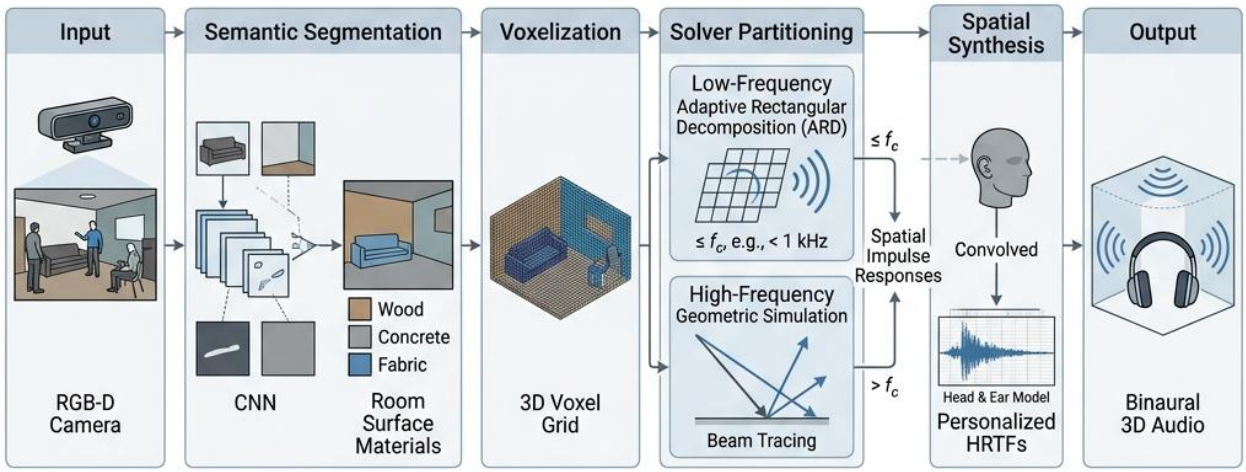

Fig -1: Conceptual architecture of the SAMOSA+ARD pipeline, integrating real-time environmental sensing, CNN-based semantic labeling, voxelization, and frequency-partitioned solvers with personalized HRTF rendering.

8. LANGUAGE-DRIVEN ACOUSTIC DESCRIPTOR SYSTEMS

The Sci-Phi model, developed by Jiang et al. (2026), is one example of a spatial audio large language model that is capable of describing entire spatial scenes of an acoustic environment. Instead of being a parameter generator, Sci-Phi is an audio analysis tool that receives First Order Ambisonics (FOA) spatial recordings and estimates the acoustic parameters and generates natural language descriptions of the scene, such as source directions, distance, loudness and room characteristics [43].

This work was inspired by a practical problem for sound designers and acoustic engineers. Physically based audio systems can create very realistic acoustic behaviour, but can be very time consuming and complex to set-up manually. Meaningful results require designers to first assign material properties in the design, to set up acoustic boundaries, to estimate absorption and scattering properties, and to set up a lot of simulation parameters. This can be a technically challenging and time consuming process in a large or complex environment.

With Sci-Phi, acoustic analysis is made easier with automated transcription of spatial characteristics. For example, if an audio scene is captured by a system, but parameters are not measured or tuned manually, the system could use Sci-Phi to identify the audio scene as a medium sized stone room with moderate reverberation. The model then decodes the speech and helps determine if the parameters are representative of the designer's intended speech or not, thus closing the qualitative/quantitative gap between user description and acoustic modeling.

This analysis may have a number of potential benefits. It can be used for automatic room characterization, speed up the verification of sound propagation models and increase the accessibility of the use of sophisticated audio analysis tools to developers who may not have an available specialist hardware for acoustic measurements. Systems like Sci-Phi could help designers be sure that their virtual spaces comply with desired acoustics, while also relieving them of some of the time-consuming aspects of scene auditing.

But the potential application of language models in this scenario also poses significant challenges. Ultimately, the acoustic simulation will be determined by physical laws, while the language models will produce outputs according to the patterns learned from data. Therefore, it is not guaranteed that the acoustic phenomena expressed by the concepts in an LLM are exactly the same as their acoustic behaviour in reality. Although the description makes sense in the mind of the programmer, it could lead to parameter values that do not make physical sense for a specific environment or are not optimum in that environment. It is, therefore, a current research area to keep the parameters generated in a physically plausible range.

While these are these worries, initial assessments of Sci-Phi are promising. Preliminary results indicate that the textual descriptions produced by the model are able to correctly classify the virtual room types and the sound sources in a wide range of different scenes, and that they are highly similar to human descriptions [43].

Overall, as this is a broader trend, there is an increasing overlap between systems in the field of artificial intelligence and sound systems. In addition to simulation

techniques, language models are being integrated into the systems for analyzing and indexing spatial audio outputs, and for validating the outputs. They have been successfully integrated into existing game engines and sound-middleware, and their vision is that future workflows will be possible, where the complicated acoustic environment can be automatically audited and described.

9. PROPOSED PERCEPTUAL EVALUATION PROTOCOL

As one of the typical challenges in spatial audio research, there is lacking an accepted and standardised evaluation framework. While many studies have documented an increase in localization accuracy, immersion, realism, and/or overall sound quality, it can be challenging to compare the results of various studies. Numerous experimental approaches, including different listening environments, HRTF configurations, acoustic scenes, rendering techniques and populations of participants. That is, if two studies give similar findings, it is not possible to determine whether these findings represent a real improvement or not, and whether the findings are really equivalent [16, 44].

This is not methodological consistency, which is a big problem in science progress and deployment. If there is no common basis for the evaluation then researchers can't easily conclude whether the improvements are due to better acoustic modelling, better test conditions, different groups of participants or different experimental designs. This makes the field prone to scattergun reaction and makes it difficult to reproduce and prove without outside help.

The solution to this problem is proposed that a dual approach evaluation protocol is followed in which a combination of qualitative expert evaluation and quantitative statistical evaluation is practiced. The aim is to provide a balanced perspective that can span the subjective aspect of audibility and the objective aspect of output performance of spatial audio systems. The combined effect of these complementary viewpoints should lead to perceptually meaningful but scientifically sound results. It is presented here as an outline protocol for future studies and as a reference for this protocol, as it has not been used in an empirical study yet.

This qualitative part involves listening evaluations conducted by experts with experience in audio engineering, acoustics, psychoacoustics and/or critical listening. The focus of these evaluations is realism, space, externalization, environment and overall immersion. An expert listener is very valuable for picking up on subtle artefacts and perceptual inconsistencies that may not be apparent from objective measurements.

Objective performance indicators are included in the quantitative portion of the assessment. They may be the localization accuracy, front-back confusion rate, the error that is made in estimating the elevation of the source, consistency of responses among listeners, computational efficiency, response latency, and some other measure of the performance of the system, depending on the research purpose. Comparisons made by statistical analysis enable scientists to determine whether the differences they are seeing are not random, but are likely due to some other factor or influence.

Methodological independence is one of the key design features of the proposed design. The protocol can be applied to any kind of spatial audio system based on any type of simulation method. All these methods (geometric, wave, hybrid, propagation, or HRTF personalization, or new approaches based on machine learning) can be tested alike. This way, the comparison of the implementation details is avoided and the comparison is limited to perceptual and performance aspects.

This would improve the ability of other groups to reproduce experiments, and afford a useful comparison between various technologies. More important, it would give a benchmark to evaluate the progress of spatial audio in the future. Over the years, the field has developed, including the ability to simulate sounds acoustically, the adoption of XR technologies, and the ability to produce sound that is more realistic to the individual. This will necessitate standard practices to evaluate the developments and determine the best directions for future research.

9.1 Expert Qualitative Evaluation

This evaluation framework will be evaluated qualitatively by a panel of experts (N = 12) 6 of which are experts in Acoustic Engineering and 6 are interactive Media/domain experts for games. This equitable mix will help bring balance to future evaluations both on the technical side of how accurately the acoustic simulation was accomplished and on its practical use for immersion. All individuals would assess the different systems being tested on a 5 point Likert Scale: (5 = excellent, 1 = poor) [1, 16, 17]. As there were no human subjects involved in this study, IRB/Ethics Committee approval was not required. In subsequent trials, informed consent and approval will be obtained prior to conducting listening tests.
Based on the number of expert listeners (N = 12), we draw on known recommendations for subjective audio

measurement. For advanced audio systems, the ITU-R BS.1116-3 standard suggests using 10-15 experienced listeners, since their sensory sensitivity is at a higher level and they do not cause any cognitive fatigue that could otherwise result from a larger, untrained population of listeners. A priori power analysis suggests that for a repeated-measures design (within-subjects) involving four systems across four perceptual dimensions, a sample of N = 12 expert listeners would have a power > 0.80 to detect a large effect size (Cohen's f ≈ 0.40 or a rank-biserial r ≈ 0.50) at the alpha = 0.05 level of significance.

The perception takes place on four perceptual dimensions, which in total encapsulate the quality of the spatial audio experience.

The degree to which the acoustic environment (AE) that is rendered is in accordance with the listener's perception of how the space represented (SP) should sound is the measure of naturalness. The evaluators should decide if the acoustic response is realistic and if there are artefacts created by the simulation method which are perceptible. The higher the rating the more the acoustics overall feel like the perceived environment, and the lower the rating, the more there are acoustics artefacts like appear unnatural reflections, acoustics that seem to come from the wrong place or from nowhere, or other acoustics artefacts that make the acoustics not seem like the perceived environment.

Externalization is a way of measuring whether the sounds are heard as part of the environment or not. The goal of successful binaural rendering is to give the impression of the sounds being located in stable positions in 3-D space. Systems showing in-head localization are rated the lowest in this category [6, 7] since this is one of the most common characteristics of a mismatch between the HRTF and the system or lack of spatial rendering.

Also known as the acoustic clarity measure C50, Spatial Clarity is a measure of the balance between early reflections and late reverberation. A good acoustic design should maintain the source intelligibility and localization cues while providing the reverberant qualities of the surrounding area. Too much reverberation will smudge the sense of space while too little reverberation will lead to an unnatural feeling of space. Systems with better balance between competing factors and are able to accurately perceive the source position within the reverberant sound field are given higher ratings [4].

Presence is the degree of perceived “realness” of the auditory world, that is, the feeling of being physically present in the virtual space. This dimension was based on existing presence research, such as the Igroup Presence Questionnaire (IPQ) and other metrics of the immersive experience [45]. The evaluators are interested in how the audio contributes to their sense of 'situatedness' and how it may evoke a sense of acoustic 'coherence' and 'belieavability'.

All evaluations are carried out in standardized listening conditions, to make it consistent across participants and experimental sessions. Binaural presentation is via a pair of circumaural open-backed headphones with frequency-response equalization to reduce colouration resulting from the headphones. Different stimuli from various systems are presented in a random order, which minimizes order effects and evaluator's bias. Each participant goes through a familiarization session before formal assessment starts and this session familiarizes the participant with the rating criteria and the 5-point scale is calibrated. This training process can enhance interrater reliability and consistency and help eliminate rater interpretation differences in the scores of the evaluation criteria, so that the scores capture perceptual differences rather than simply individual differences in rater interpretation.

This qualitative evaluation framework is designed to be more structured and reproducible, and leverages expert judgment, controlled listening conditions and perceptual dimensions that are clearly defined in order to assess perceptual performance of spatial audio systems in a large variety of simulation approaches and application domains.

9.2 Inferential Statistical Verification

Objective acoustic measurements would be added to the subjective assessments given by expert listeners to complement the subjective measures. Perceptual evaluations describe the perception of the spatial audio system; objective measures would offer quantitative proof of the acoustic accuracy of a spatial audio system and allow it to be compared in a repeatable way across spatial audio system implementations. These two points of view are meant to be combined to help guarantee that the performance gain that is measured is the one that is perceptually valuable.

The two main objective measures are reverberation time (RT60) and the clarity index (C50), which are typically used in architecture acoustics and room-acoustic analysis. The following parameters are directly measured from the rendered impulse responses, according to the following standardised procedures as specified in ISO 3382. RT60 is the best measure of the reverberant nature of a space and is the time it takes for the sound energy to drop by 60 dB in the space. C50 is the ratio between energy reaching the

listener first and that reaching the listener later in the room, and is widely used as an indicator of the intelligibility and clarity of speech.

The measured RT60 values would be compared with the RT60 value of a reference simulation or reference acoustic model for consistency in acoustic accuracy. It is suggested that the target error value for this system is less than 0.18 seconds mean absolute error (MAE). For the sake of making this threshold "real" we will use the Just-Noticeable Difference (JND) for reverberation time. JND for RT60 is defined as 5% of the nominal value as per ISO 3382-1. This JND is roughly 0.05 - 0.15 seconds when the decay times are 1.0 to 3.0 seconds as is common in room acoustics. Therefore a reasonable, perceptually valid criterion, close to the human detection threshold [46, 47] is a MAE of 0.18 seconds.

The statistical analysis would then serve to decide whether any observed differences are statistically significant (making them more than just random fluctuation) or not. As Likert scale data is ordinal, and thus do not meet the normality assumptions of parametric statistics, non-parametric statistics would be used instead. If there was a repeated measures design with two systems, the Wilcoxon signed-ranked test would be used for each perceptual dimension to compare the two systems. As multiple comparisons are done simultaneously, Benjamini-Hochberg False Discovery Rate (FDR) correction would be used instead of the highly conservative Bonferroni correction to control the overall Type I error rate, and statistical significance evaluated at an alpha level of 0.05.

If three or more systems were involved in an experiment using a repeated measures design, a Friedman test (non-parametric version of one-way repeated measures ANOVA) would be used to determine if there are significant differences among the group medians. If Friedman test shows that there is a statistically significant effect, then post hoc pairwise comparisons would be conducted among the specific systems by Wilcoxon signed-rank test with Benjamini-Hochberg FDR correction so that the family-wise error rates are controlled.

Significance testing, as well as effect sizes (using the matched pairs rank biserial correlation r) would be reported for all pairwise comparisons. While p values tell us whether there is a statistically significant difference between the observed values, effect sizes tell us the size of the difference between observed values and it will provide us with a measure of practical importance scaled independently of sample size. Reporting of both statistical significance and non-parametric effect sizes provides a more comprehensive analysis of experimental results and ensures that differences that are only detected, but not perceptually meaningful, are not interpreted as statistically significant [16].

These objective measures and statistical analysis will give a sound quantitative basis to the suggested evaluation system. The methodology uses standardized acoustic metrics, controlled statistical testing, and reporting of effect size to allow for transparent reproducible comparisons between spatial audio systems and to ensure that a complementary perceptual–physical approach is used to assess perceptual and physical performance.

Table -3: Proposed perceptual evaluation metrics for spatial audio systems, indicating measurement procedure, target acceptance threshold, and source reference.

Metric	Measurement Method	Target Value	Reference
RT60 MAE	ISO 3382-1 RIR measurement	< 0.18 s	ISO 3382-1
C50 Clarity	Early-to-late energy ratio	> 0 dB (for intelligibility)	ISO 3382-1
HRTF Spectral Error	Log-spectral distance (LSD)	< 4.5 dB	Hu et al. (2025a)
Localisation MAE (azimuth)	Pointer task, anechoic	< 8°	Blauert (1997)
Localisation MAE (elevation)	Pointer task, anechoic	< 12°	Xie (2013)
Presence (Igroup Presence Questionnaire)	Self-report questionnaire	Score > 4.0/7.0	Schubert et al. (2001)
Naturalness (Likert)	Expert panel (N=12)	Median > 3.5/5.0	Proposed (this paper)
Externalisation (Likert)	Expert panel (N=12)	Median > 3.5/5.0	Proposed (this paper)

10. DISCUSSION AND LIMITATIONS

10.1 The Immersion Gap: Quantitative Evidence

The immersion gap between visual and audio realism of modern games is not just a subjective matter, it can be measured quantitatively. Although there have been significant strides in graphics technology in recent years that have allowed real-time rendering systems to produce stunning visual quality, the quality of spatial audio in many commercially available game engines has not kept pace. In interactive experiences, the auditory aspect is often not as accurate or perceptually realistic as the visual aspect, resulting in a delay between the visual and auditory information.

Fırat et al. (2022) tested the virtual acoustics performance of Unreal Engine 4 and Wwise with the commercial geometrical acoustics software Odeon and found that there was a difference. The study compared the physical acoustics parameters in different virtual environments against the physical standard, and found that the native game engine spatialization tools produced large differences in both the low frequency range (63 to 125 Hz) and the high frequency range (8 kHz). These differences were minimized by integrating a dedicated audio middleware, showing that native engine configurations are not enough to achieve physically accurate room acoustic simulation of audio. Spatial audio engines and personal HRTFs tend to give better localization, but often native implementations have large localization errors because they simplify the propagation model.

These results indicate a significant achievement deficit. In particular, deviations in the room acoustic parameters (such as Early Decay Time (EDT) and Reverberation Time (RT30) differences) in the low frequencies, directly affecting the localization cues and the spatial ambiguity, were found to be more than 2.0 s for native engine configurations without propagation middleware. A player's skill to accurately localize the sound source can be significantly impaired by these physical parameter deviations, and can increase the player's confusion between the front and back, and decrease their immersion in the virtual environment.

The same trend is observed in analyzing reverberation accuracy. There are many commercial game engines that use a pre-programmed convolution reverb or an algorithmic reverb system set up by the sound designer. These methods are computationally simple and relatively straightforward to write, but are not able to mimic the true acoustics of complex environments. There are considerable deviations from the room-acoustic parameters of native engines when compared to the physically simulated acoustic references as mentioned in the comparison of virtual engines and middleware by Fırat et al. (2022).

The geometry-based propagation systems are much more effective. These methods involve a significant improvement in acoustic accuracy by deriving the acoustic behaviour directly from the structure and materials of the environment. In comparison, physically-informed propagation models are more effective at reproducing the reverberant nature of real spaces, which gives a better approximation to the desired acoustic situation and increases the accuracy of the virtual acoustics to real measured data.

Overall, this study offers compelling empirical evidence that the sound quality that many commercial game engines currently deliver is still far from what can be accomplished with current state-of-the-art spatial audio technologies. Whereas modern engines have come a long way, achieving amazing realism in appearance, there has been much less work done in terms of realism when it comes to sound. The resulting difference would further confirm the presence of an immersion difference and the continued need for further research on more advanced propagation models, customized HRTF rendering, and physically based acoustic simulation. If there is a gap between the two, it will be crucial in developing virtual environments where both the auditory and visual realism can coexist at an equal pace.

10.2 Computational Budget Analysis

The perceptual advantages of enhanced spatial audio have been proven to be true, but the practical advantages are slight: computational cost. No matter how good an acoustic simulation can be, it has to meet the very demanding real-time interactive application performance requirements.

Most games nowadays work with a frame rate of 60 frames or more, which means that the calculations for rendering, physics, AI, network communications, and audio are supposed to be done in about 16.7 milliseconds per frame. In applications of virtual reality, there are even tighter constraints, with the typical requirement of 90 Hz, which cuts the frame time down to around 11.1 milliseconds. Audio processing is typically performed at a separate update rate (approx. 5-10 milliseconds), and therefore only a small portion of the available processing power can be used for acoustic simulation. In practice, this means that there are about 1-3 msec available (on a shared CPU thread!) or 5-8 msec available (on a dedicated thread for

audio processing!) for real-time spatial-audio computations.

In this range of constraints, the different spatial-audio techniques have very distinct computational properties. Wave technologies are still the most costly ones. For a moderately complex room, a full scene Adaptive Rectangular Decomposition (ARD) simulation takes 8 to 15 milliseconds using modern GPU hardware (e.g. mid-range NVIDIA RTX 3080/4070 desktop GPUs) [9]. This is a high cost per frame but is feasible to do asynchronously (in parallel) in the background and update at a lower rate (e.g., 10-30 Hz). Those update rates are typically adequate for the low frequency acoustic behavior since this behavior changes relatively slowly and does not strain the system resources to maintain perceptual realism.

The geometric-acoustic techniques are much milder. The typical processing time for beam tracing to compute early reflections is about 0.5 – 2 milliseconds for a moderate complexity scene (e.g., 10,000 polygons, 5th-order reflections) on a single CPU thread [30]. For the few sounds that are acoustically significant, like dialogue, the significant sounds (such as the environment) or the major sounds (such as music), this overhead is not out of reach of current technology.

The added computational cost of binaural rendering via HRTF convolution is a predictable additional cost. The required computation time for the HRTF processing mostly lies between 0.3 and 0.8 milliseconds per ear channel per source for typical 128-tap or 256-tap FIR filters on modern CPUs [7] depending on the complexity of the filters and their implementation details. All these active sources have to be treated separately, so the cost increases approximately in proportion to the number of sound sources that are audible simultaneously. For large and acoustically dense scenes, efficient source management is therefore an important factor to take into account.

These techniques can be used together in a well-designed hybrid pipeline and fit the performance constraints of today's games. Beam tracing can be used for early reflections, statistical reverberation for the late reverberant field, and HRTF convolution for binaural rendering can be computed within the standard 60 Hz game loop, and ARD simulation can be updated asynchronously. The evidence currently available indicates that a similar pipeline for an environment with around 10–20 concurrent audible sources is possible with the use of suitable level-of-detail techniques [21, 30].

These optimizations generally include lowering the fidelity of the simulated sound for sources that are far away, are very attenuated, or are not much of an acoustical interest to the player, but keeping the quality of the simulated sound at a higher level for sounds that are acoustically very interesting to the player. This dynamic resource allocation approach allows developers to strike a realistic balance between realism and real-time performance.

Thus, the main hurdle when implementing the use of sophisticated spatial sounds is no longer the sheer computational impracticality. Rather, it is a matter of effectively utilizing the resources available and making intelligent choices of the matching simulation techniques. Hybrid pipelines, which incorporate geometric, statistical and wave-based techniques, are increasingly proving that physically informed spatial audio can be integrated to real-time applications, without the performance burden of current gaming and XR platforms.

10.3 Demographic Bias and Inclusive Design

The demographic restrictions on current HRTF databases raise more than just technical issues, and they are fundamental questions of accessibility, inclusiveness, and fairness in spatial-audio technology. The participants in most commonly-used HRTF databases are typically small in number and skewed towards young adult males in Western European countries. Therefore, the anatomical features included in these databases are not representative of the population across the world.

This imbalance directly affects user experience. Listeners can have poor spatial-audio performance if they do not share the physical characteristics of the majority of the population that is represented in an HRTF set, including head size, ear shape, torso size and other physical attributes. This involves numerous women, children, older people and people from minority ethnic backgrounds. With the best available binaural rendering systems, these users can experience reduced localization accuracy, greater front-back confusion, less convincing sound externalization and a less convincing spatial experience [14, 15].

From this perspective, it's not just about which metrics are being captured, but about ensuring access to the best immersive technologies for everyone. Spatial-audio technologies, if they systematically deliver better experiences for certain groups of people than for others, have an unequal benefit. With the growing significance of spatial audio in the gaming, virtual reality and augmented reality worlds, education and communication, the

importance of solving these differences is both technical and inclusive.

In theory, the answer is simple, develop larger, more representative HRTF databases, including participants of a wide age range and sex, ethnic group and anatomical variation. These would offer a richer basis for direct HRTF selection and the future development of personalization systems. But it's not an easy solution to implement. Historically, HRTFs have been acquired in specialized laboratories, with specialized equipment, and during a lengthy measurement session, which makes large-scale measurement time consuming and expensive. The building of a demographically balanced database with hundreds of participants from various population groups, with long planning and huge resources from worldwide research institutions would be very costly [15].

This challenge has started to be tackled by efforts like SONICOM, which have worked on increasing the demographic coverage of publicly available data. While these are great strides towards minimizing bias in spatial-audio technologies, there are still many efforts to go before a truly worldwide coverage can be accomplished.

Machine-learning-based personalization is a complementary approach that has the potential to be effective. Deep-learning systems measure each person indirectly instead of directly, using a set of easily obtained inputs, like photographs, images from smart phones or three-dimensional scans, to estimate the HRTFs of the individual. Several approaches have shown that a realistic prediction of the HRTF might be possible without the costly anechoic-chamber measurements, thus cutting by several orders the cost of personalization [11, 12] as in the approaches hereof.

These approaches do not, however, remove the use of representative datasets. Deep-learning models are like any data-driven system: They learn from examples they've been trained with. When the training data is small and narrow, the models generated may not accurately apply to people with anatomy that is not well represented in the data. That is, even with state-of-the-art machine-learning models, measurement bias can result in biased prediction models.
Therefore, expanding databases and using machine-learning for personalization shouldn't be seen as mutually exclusive solutions, but rather as part of a larger approach. The large and varied HRTF databases are essential not only to enhance direct HRTF selection but also to increase training data for reliable personalization algorithms. These efforts can all contribute to a more accurate and immersive spatial-audio experience across a broad spectrum of users, instead of limiting the experience to a few.

There is still much work to be done to achieve truly inclusive spatial audio, and this will require the continued efforts of many groups, including acousticians, audio engineers, machine-learning researchers, hardware manufacturers, and funding bodies. These synchronous works are needed to achieve a technically advanced spatial-audio system that is widely representative of human listeners' diversity.

10.4 Open Problems and Research Directions

Even with all the advancements in spatial audio, there are still a number of key issues that have not been addressed and are still on the research agenda for spatial audio.

Firstly, the issue is how to integrate real-time wave based acoustic simulation into a dynamic environment. The first one is about real-time wave-based acoustic simulation in a dynamic environment. Today's games and XR experiences are growing in complexity by incorporating a constantly evolving world that responds to a player's presence and actions, from doors opening and closing, the movement of furniture, the addition or removal of objects, to the destruction or creation of environments. Ideally, the acoustic response of the environment should change instantly to these changes. But it is hard for this level of responsiveness to be realized. Pre-computed acoustic methods rely on a good knowledge of the geometry of the scene and its variation to a smaller degree, while fully dynamic wave based methods like FDTD and ARD require much more computational power and frequently claim a lot of GPU memory when the graphics are rendered. This means that it is currently impractical to have real-time updates at such a rate for today's games. Another potentially fruitful area of future work is incremental update algorithms that take advantage of the locality of environmental changes in many cases. These would adjust only the parts of the acoustic field that are modified, rather than re-computing the entire acoustic field, and could increase the efficiency of computation to a reasonable level.

Language model-based acoustic design tools are another region that needs to be studied further, which involves validation. Sci-Phi shows that the use of large language models for analysing spatial audio recordings and producing text descriptions and sometimes perceptually convincing results can be achieved. But the correlation between these parameters and the true acoustical response of the system is not fully understood. Prior evaluations have been subjective, rather than physical.

Systematic comparisons against measured acoustic environments and physically validated simulations would be needed to ensure the confidence in the use of such tools in professional audio productions or any other perceptually sensitive application. A key open research question is whether or not a language-generated parameter set always represents a realistic acoustic scenario [43].

Another one of the challenges is comprehending the interaction between spatial audio and other sensory modalities that allow to create a sense of presence. Immersive experiences are multimodal not only are they using sound, they are using pictures, perceiving depth, using motion and more and more the haptics. While most researchers consider audio as an important factor of immersion, the influence of specific acoustic cues is still not well understood. For instance, it remains to be determined whether the difference in presence can be attributed to the modelling of the diffraction or to the high level of personalization of the HRTFs, or whether the difference in perceptual benefit can be attributed to the presence or to the high level of localization accuracy under certain circumstances. These questions are especially crucial when developers have to allocate computational power to various resources and run the computation on a limited set of resources.

To tackle such considerations, there needs to be a better understanding of human perception. In addition to the improvements of the acoustic simulation algorithms, more close cooperation is needed between the acoustic engineers, psychoacousticians, cognitive scientists and XR researchers to make future progress. Knowing which parts of spatial audio are the most important for immersion will help the researchers create systems that capture perceptually relevant properties, while reducing or eliminating those elements that have the least impact on the user's experience, but are more computationally intensive.

All of these issues point to a greater fact, that the remaining issues in spatial audio are no longer just those of simulation precision. They are often increasingly complex, requiring balancing of physical realism, computational efficiency, perceptual effectiveness and accessibility. This will take the collaboration of interdisciplinary research that leverages advances in acoustics, artificial intelligence, graphics hardware, human perception, and interactive media design. The answers that will be created in these areas will be at the heart of the creation of the next generation of immersive digital environments.

11. CONCLUSIONS

This paper has reviewed the evolution of spatial audio technologies from the geometric acoustics approach to the current wave-based, hybrid and machine-learning approaches and included a comprehensive taxonomy. We have systematically compared propagation solvers, HRTF database biases and middleware solutions, and identified the important physical and perceptual properties that determine the acoustic realism in interactive game environments.

The key issues in the future for the next generation of spatial audio are not so much finding specific solvers, but integrating systems and making them accessible to the workflow. For real time dynamic environments, mixed pipelines are needed that can execute wave accurate low frequency simulations asynchronously with the low latency high frequency geometric solver. At the same time, moving away from demographically limited HRTF databases to representative, world-wide diverse corpora is becoming increasingly important, and is being driven by low cost machine-learning personalization models in order to achieve auditory equity.

Finally, spatial audio should not be viewed as a mere add-on to immersive computing but as an integral part of the experience. Digital environments are becoming more prevalent in healthcare, collaborative work and remote education, and the fidelity of what is being heard is becoming as important as what is being seen. It is hoped that standardisation of evaluation protocols, including the non-parametric statistical framework presented in this work, will be important in facilitating the creation of reproducible benchmarks and in the systematic progress achieved industry-wide.

Data and Code Availability

The conceptual framework proposed in this paper does not have a public software release. The datasets referenced in this study are publicly available at their respective repositories: the CIPIC HRTF database (https://www.cipic.ucdavis.edu), the SONICOM HRTF dataset (https://www.sonicom.eu), the LISTEN HRTF database (https://www.ircam.fr), and the HUTUBS HRTF dataset (https://depositonce.tu-berlin.de).

Author Contributions

Anubhav A. Patil and Prof. Sunny Nahar are the sole authors of this manuscript, responsible for the conceptualization, literature review, taxonomy

development, proposed SAMOSA+ARD integration, evaluation protocol design, and writing of the paper.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the editorial board and the anonymous reviewers for their valuable comments, insightful suggestions, and constructive feedback throughout the review process. Their observations helped improve both the clarity and quality of this work.
We also acknowledge the broader spatial audio research community for its continued contributions to the field. The availability of open-access datasets, publicly released source code, benchmark frameworks, and research tools has played a crucial role in advancing spatial audio research and made this comparative study possible. The collaborative efforts of researchers, developers, and institutions worldwide continue to accelerate innovation in acoustic simulation, binaural rendering, and immersive audio technologies. Finally, we recognize the contributions of the many researchers whose foundational work over the past several decades has established the theoretical and practical foundations upon which contemporary spatial audio systems are built. Their efforts have enabled the development of increasingly realistic and accessible immersive audio experiences across gaming, virtual reality, augmented reality, and other interactive applications.

REFERENCES

[1] Hong, J.Y., He, J., Lam, B., Gupta, R., Gan, W.-S.: "Spatial Audio for Soundscape Design: Recording and Reproduction"; Applied Sciences, 7(6), 627 (2017).

[2] Witmer, B.G., Singer, M.J.: "Measuring presence in virtual environments: A presence questionnaire"; Presence: Teleoperators and Virtual Environments, 7, 3 (1998), 225–240.

[3] Melchior, F., et al.: "Spatial Audio in Immersive Environments: From VR Production to Consumer Delivery"; Journal of the Audio Engineering Society, 67, 9 (2019), 669–685.

[4] Fırat, H.B., Maffei, L., Masullo, M.: "3D sound spatialization with game engines: the virtual acoustics performance of a game engine and a middleware for interactive audio design"; Virtual Reality, 26, 4 (2022), 1481–1499.

[5] Kleiner, M., Dalenback, B., Svensson, P.: "Auralization — An Overview"; Journal of the Audio Engineering Society, 41, 11 (1993), 861–875.

[6] Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition); MIT Press, Cambridge, MA (1997).

[7] Xie, B.-S.: Head-Related Transfer Function and Virtual Auditory Display (2nd ed.); J. Ross Publishing, Fort Lauderdale, FL (2013).

[8] Noisternig, M., et al.: "A 3D Ambisonic Based Binaural Sound Reproduction System"; Proc. AES 24th International Conference, Banff, Canada (2003).

[9] Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M., Govindaraju, N.: "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition"; IEEE Transactions on Visualization and Computer Graphics, 15, 5 (2009), 789–801.

[10] Southern, A., Murphy, D., Savioja, L.: "Spatial Encoding of Finite Difference Time Domain Acoustic Models for Auralization"; IEEE Transactions on Audio, Speech, and Language Processing, 20, 9 (2013).

[11] Hu, D., et al.: "Graph Neural Field with Spatial-Correlation Augmentation for HRTF Personalization"; arXiv preprint arXiv:2511.10697 (2025), to be published in Proc. AAAI Conference on Artificial Intelligence (AAAI 2026).

[12] Hu, X., Li, J., Zhang, S., Goetz, S., Hogg, A.O.T.: "HRTFformer: A Spatially-Aware Transformer for Personalized HRTF Upsampling in Immersive Audio Rendering"; arXiv preprint arXiv:2510.01891 (2025).

[13] Funkhouser, T., Carlbom, I., Elko, G., Pingali, G., Sondhi, M., West, J.: "A Beam Tracing Approach to Acoustic Modeling for Interactive Virtual Environments"; Proc. ACM SIGGRAPH 1998, ACM, New York (1998), 21–32.

[14] Bruschi, V., et al.: "A Review on Head-Related Transfer Function Generation for Spatial Audio"; Applied Sciences, 14(23), 11242 (2024).

[15] Thiemann, J., et al.: "The SONICOM HRTF dataset: Methodology, validation, and practical applications"; Journal of the Audio Engineering Society, 70, 4 (2022), 256–268.

[16] Bech, S., Zacharov, N.: Perceptual Audio Evaluation: Theory, Method and Application; John Wiley & Sons, Chichester (2006).

[17] Xu, T., et al.: "Enhancing XR Auditory Realism via Multimodal Scene-Aware Acoustic Rendering (SAMOSA)"; Proc. ACM Symposium on User Interface Software and Technology (UIST '25) (2025).

[18] Schroeder, M.R., Atal, B.S.: "Computer simulation of sound transmission in rooms"; Proc. IEEE International Convention Record, 7 (1963), 150–155.

[19] Tsingos, N., Funkhouser, T., Ngan, A., Carlbom, I.: "Modeling acoustics in virtual environments using the uniform theory of diffraction"; Proc. ACM SIGGRAPH 2001, ACM, New York (2001).

[20] Creative Labs: "Environmental Audio Extensions (EAX) Software Development Kit, Version 1.0"; Creative Technology Ltd. (1998).

[21] Han, Y., et al.: "Perspectives of Sound Designers on Real-Time Sound Propagation in Games"; Proc. IEEE Conference on Games (CoG 2025) (2025).

[22] Allen, J.B., Berkley, D.A.: "Image method for efficiently simulating small-room acoustics"; Journal of the Acoustical Society of America, 65, 4 (1979), 943–950.

[23] Borish, J.: "Extension of the image model to arbitrary polyhedra"; Journal of the Acoustical Society of America, 75, 6 (1984), 1827–1836.

[24] Kuttruff, H.: Room Acoustics (4th ed.); Spon Press, London (2000).

[25] Sikora, M., et al.: "Hybrid Beam and Ray Tracing Simulation Method for Architectural Acoustics"; Proc. Euronoise 2018, Crete, Greece (2018).

[26] Kulowski, A.: "Algorithmic representation of the ray tracing technique"; Applied Acoustics, 18, 6 (1984), 449–469.

[27] Krokstad, A., Strom, S., Sorsdal, S.: "Calculating the acoustical room response by the use of a ray tracing technique"; Journal of Sound and Vibration, 8, 1 (1968), 118–125.

[28] Savioja, L., Svensson, U.P.: "Overview of geometrical room acoustic modeling techniques"; Journal of the Acoustical Society of America, 138, 2 (2015), 708–730.

[29] Laine, S., et al.: "Accelerated 3D sound propagation in a virtual environment using a hierarchy of portals"; Proc. IEEE Symposium on Interactive 3D Graphics and Games (2009).

[30] Chandak, A., et al.: "AD-Frustum: Adaptive Frustum Tracing for Bidirectional Sound Propagation"; IEEE Transactions on Visualization and Computer Graphics, 14, 6 (2008), 1475–1482.

[31] Taylor, M., et al.: "Guided Multiview Ray Tracing for Fast Auralization"; IEEE Transactions on Visualization and Computer Graphics, 15, 1 (2009), 114–124.

[32] Botteldooren, D.: "Numerical simulation of the effect of obstacles on sound propagation"; Journal of the Acoustical Society of America, 95, 6 (1994), 3497–3500.

[33] Kowalczyk, K., Van Walstijn, M.: "Room acoustics simulation using 3-D compact explicit FDTD schemes"; IEEE Transactions on Audio, Speech, and Language Processing, 19, 1 (2011), 34–46.

[34] Bilbao, S.: Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics; John Wiley & Sons, Chichester, UK (2009).

[35] Hamilton, B., Webb, C.J.: "Room acoustics modelling using GPU-accelerated finite difference and finite volume methods on a face-centred cubic grid"; Proc. Digital Audio Effects (DAFx-13), Maynooth, Ireland (2013).

[36] Algazi, V.R., Duda, R.O., Thompson, D.M., Avendano, C.: "The CIPIC HRTF Database"; Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2001), 99–102.

[37] Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M., Govindaraju, N.: "Precomputed Wave Simulation for Real-Time Sound Propagation of Dynamic Sources in Complex Scenes"; ACM Transactions on Graphics, 29, 4 (2010), Article 68.

[38] Mehra, R., et al.: "Wave-based sound propagation in large open scenes using an equivalent source formulation"; ACM Transactions on Graphics, 33, 2 (2014), Article 19.

[39] Tsingos, N.: "Pre-computing geometry-based reverberation effects for games"; Proc. AES 35th International Conference, London (2009).

[40] Audiokinetic: "Wwise Spatial Audio Integration Guide, Version 2019.1"; Audiokinetic Inc. (2019).

[41] Scerbo, M., et al.: "Efficient Multichannel Auralization Based on MoD-ART"; IEEE Transactions on Audio, Speech, and Language Processing (2025).

[42] Zang, Y., Kong, J.: "GSound-SIR: A Decoupled Ray-Tracing Toolkit for Spatial Impulse Response Research";

Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025).

[43] Jiang, X., et al.: "Sci-Phi: A Large Language Model Spatial Audio Descriptor"; IEEE Open Journal of Signal Processing (2026).

[44] Sakamoto, S., et al.: "Computational simulation of sound field in a room using the finite-difference time-domain method"; Journal of Environmental Engineering, 10, 2 (2015).

[45] Schubert, T., Friedmann, F., Regenbrecht, H.: "The experience of presence: Factor analytic insights"; Presence: Teleoperators and Virtual Environments, 10, 3 (2001), 266–281.

[46] ISO 3382-1: Acoustics Measurement of room acoustic parameters Part 1: Performance spaces (2009).

[47] Martellotta, F.: "Just noticeable differences of spatial room acoustic parameters in a small theater"; Journal of the Acoustical Society of America, 128, 2 (2010), 654–663.

[48] Funkhouser, T., Jot, J.-M., Tsingos, N.: "Sounds Good to Me! Computational Sound for Graphics, Virtual Reality, and Interactive Systems"; SIGGRAPH 1998 Course Notes (1998).

[49] Thiemann, J., et al.: "The SONICOM HRTF Dataset: An Open Database for Research on Personalized Binaural Audio"; Frontiers in Signal Processing, 2 (2022), Article 894567.

[50] Vorlander, M.: Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality (2nd ed.); Springer, Berlin (2020).

[51] Choueiri, E.Y.: "The 3D3A HRTF Database"; Princeton University 3D3A Lab Report (2021).

[52] Armstrong, M., et al.: "SADIE II: A Database of Head-Related Transfer Functions and Associated 3D Models"; Proc. Audio Engineering Society Conference on Spatial Audio (2018).

[53] Siltanen, S., Lokki, T., Savioja, L.: "A boundary element method for room acoustics using image sources"; Journal of the Acoustical Society of America, 126, 4 (2009), 1828–1840.

[54] Rindel, J.H.: "The use of computer modeling in room acoustics"; Proc. MISA 2000, 1–17 (2000).

[55] Xiang, L., Schissler, C.: "Real-time sound propagation with GPU-accelerated wave-based solvers"; Proc. IEEE (2014).