electronics

MDPI

Article

MAVIS: Multi-Stem Audio Visualisation in Immersive Spaces Framework

Jethro Shell [1,*,†] and Sophy Smith [2,†]

[1] Digital Futures Institute, De Montfort University, Gateway House, The Gateway, Leicester LE1 9BH, UK
[2] School of Creative and Cultural Industries, University of the Creative Arts, Falkner Road, Farnham, Surry GU9 7DS, UK; sophy.smith@uca.ac.uk
* Correspondence: jethros@dmu.ac.uk; Tel.: +44-0116-207-8520
† These authors contributed equally to this work.

Abstract

The visualisation of music has gained traction in both research and musical composition in recent years. The increased accessibility to immersive technologies, such as virtual reality (VR) and other forms of mixed reality (MR), lend themselves to the examination of how visualisation can impact the perception of audio virtual worlds. In this paper, we propose the MAVIS (Multi-stem Audio Visualisation in Immersive Spaces) design framework, an approach to generating a visualisation of multi-stem structured orchestral music in a virtual world. This research explores the impact on participants' interaction with an orchestral musical composition through the use of a two framework iterations informed by use cases. The resulting final design structure outlined in this article points towards constructing multi-stem virtual orchestral experiences through three pillars: semantic consistency, spatial agency, and complexity control. Whilst this research serves to propose a design intervention, future work requires a more extensive participant testing approach, coupled with an exploration of additional multimodal analysis.

Keywords: immersive visualisation; virtual reality; multi-stem audio

1. Introduction

Music visualisation gained particular traction over the last decade [1–5]. The increased accessibility to immersive technologies, such as virtual reality (VR) and other forms of mixed reality (MR), has opened new avenues for examining how visual data impacts the perception of audio-driven virtual worlds.

Whilst early work drilled into the visualisation of single stem, monophonic composition, the complexity of music and its capture has resulted in a need to examine multi-stem environments. This work is focussed on immersive interaction with orchestral forms, which by definition are composed of multiple instruments. Multi-stems are more complex due to their polyphonic nature making classification and visualisation challenging. A significant hurdle in this field is the arbitrary nature of mapping audio parameters such as frequency or amplitude to visual objects in a way that feels intuitive rather than confusing for the participant.

In this paper we propose the MAVIS (Multi-stem Audio Visualisation in Immersive Spaces) design framework to bridge the gap between passive observation and active creation. The purpose of the work is to facilitate design interventions for the development of immersive, multi-stem structured orchestral music visualisations to afford a new audience

Academic Editor: George Angelos Papadopoulos

Received: 15 January 2026
Revised: 30 March 2026
Accepted: 31 March 2026
Published: 8 April 2026

experience. Drawing on the work of Liang et al. [6], a hybrid Agile–Double Diamond [7] design methodology was adopted for the development process allowing for participant-centred, fast prototyping. The methodology utilised a four-stage structure:

1. Diamond One: Discover (diverge)—explore the context by researching the participant needs.
2. Diamond One: Define (converge)—synthesise the research into a single problem statement. By contextualising the application, this affords a greater focus, removing outlying requirements.
3. Diamond Two: Develop (diverge)—implement fast prototyping investigating multiple ways in which the questions can be resolved.
4. Diamond Two: Deliver (converge)—narrow to the final solution to create a functional prototype and subsequent design framework.

The second diamond is encapsulated within an Agile development method allowing for a test process and development to be refined iteratively. The final delivery phase is evaluated by a small participant group to inform the design intervention parameters [6]. The final element is subject to the future work of this research.

To formulate the framework, two use cases were used. The use cases took part in person at separate locations over a 12-month period. Participants were given VR based experiences with differing designs with separate orchestral compositions.

The immersive experience employed Meta Quest devices [8] VR technology. These allowed for participants to explore classical compositions by Gustav Holst and by Béla Bartók in an immersive virtual environment (this will be expanded upon in Section 4). The output from the engagement with the primary set of participants informed the design of the second framework, conforming to the Double Diamond approach.

The design framework is summarised as being constructed of three pillars:

1. Semantic Consistency: Ensuring the visual grammar; for example, using specific colours for instrument families that make sense to the participant.
2. Spatial Agency: Providing participants with the ability to move through space and manipulate audio stems to influence their own experience.
3. Complexity Control: Using cognitive scaffolding and visual constraints (such as light intensity) to prevent sensory overload.

The resultant contribution to knowledge can be distilled as being:

- The formation of a design framework for the creation of complex, multi-stem audio visualisation in immersive 3D virtual reality environments;
- A design approach to inform practitioners of ways in which to migrate away from wholly replicating real-world interactions, such as Philharmonia's immersive experience [9], towards the ability for participants to create their own narrative;
- A design framework structure that can help in informing the creation of agency in diversive musical compositions that are not solely orchestral in nature.

This research built upon the findings of the Audience of the Future project [10], which explored the use of new immersive technologies to both allow audiences to engage with performances in new ways and also to gain insight into ways in which mixed reality technologies can remove the restrictions of locality to a performance.

The rest of the paper is formed of eight sections:

1. Research Questions;
2. Related Work;
3. Materials and Methods;
4. Results;
5. Discussion;

6. Contribution To Knowledge;
7. Conclusions;
8. Future Work.

This paper opens with a presentation of the research questions and a deep dive into the related work. This discursive approach allows the authors to situate MAVIS within the existing landscape of immersive musical visualisation while identifying the specific research gaps that the framework seeks to address. In the Section 4, the text details the hybrid Agile and Double Diamond design methodology. This section discusses the choice of software used and the specific participant-centred prototyping stages used to refine the experience. The Section 5 discusses the findings from both quantitative surveys and qualitative participant feedback. This leads into a discussion of how these findings specifically shaped the final form of the MAVIS framework, emphasising the three pillars of semantic consistency, spatial agency, and complexity control. The final sections of the paper formalise the contribution to knowledge and offer a conclusion that synthesises the research insights. Finally, the Section 9 adopts a forward-looking perspective, discussing new thematic avenues and the potential for more extensive multimodal analysis.

2. Research Questions

This research focussed on two key research questions:

1. In the context of multi-stem audio visualization, what are the design principles for mapping complex sonic data to 3D virtual environments to move beyond mimetic replication toward abstract, participatory interaction?
2. To what extent can a cross-genre design framework facilitate musical agency in immersive environments, and how does this framework specifically enable non-orchestral compositional structures to support user-driven narrative autonomy?

3. Related Work

There are a number of areas of research within immersive musical visualisation stemming from a broad range of musical genres. The authors consider these applications and research practices; however, the focus of this work is initially on the application of visualisation for orchestral music and composition in a virtual, immersive space. There are many definitions of an orchestra, with a common consensus that there is no single one that is applicable to all ensembles. However, for the purposes of this research, the authors chose to adopt the definition in part proposed by Spitzer et al. [11] that there is "part doubling" in the piece, that being multiple instrument types playing a single section, the concept of sectioning an ensemble so that it contains areas such as strings, brass and percussion, and that the piece is written for a standardised structure.

Equally, discussion as to what the definition of immersion encompasses is a long standing one. For this research, a definition has been adopted whereby immersion is 'a state of deep mental involvement in which the subject may experience disassociation from the awareness of the physical world due to a shift in their attentional state' [12].

3.1. Visualisation

The visualisation of sound has a long history of being captured in the virtual world. Systems that use vocal timbre, instrument chord data and melodic information have been used to visualise responses in video outputs and avatars [13]. The structure of specific elements of sound have been used extensively to add depth to the virtual experience with animated graphical visualisation of shapes and colours that reflect changes in frequency and rhythm [14]. Aspects that can be extracted from music take the form of objects in the virtual world. Analyses of music such as the use of Fast Fourier Transforms (FFT) and amplitude are

mapped to particular elements that exist within a scene. Research that has been carried out has drawn attention to the fact that there are elements of difficulty in carrying out such processes, including how these parameters should be mapped to the objects in the virtual world [15] and what form the crossmodal correspondences should take [16]. Through exploring these gaps in the research and application, this work considers how these may be resolved and offers a framework for audio visual immersive designers to utilise.

3.2. Colour Representation of Instruments and Sounds

A relationship between pitch height and lightness is well-documented. Higher pitches are associated with lighter colours, while lower pitches are associated with darker colours [17]. Individuals listening to audio, even those without Chromesthesia (a perception of nonvisual stimuli, for example sounds, tastes, odours, or colour perception, when listening to audio [18]), tend to map sounds to colours using consistent logic. Studies into crossmodal effect have demonstrated that both adults and children paired brightness with loudness [19,20]. People were also shown to map high-pitched tones to brighter surfaces [21]. Taking these findings on board, the MAVIS framework considered the relationship between the orchestral stem output and virtual object structure.

Multi-Stem

Visualisation Development in multi-stem visualisation has progressed, building from a focus on single-stem. Early work such as Hiraga et al. [22] considered how an application could visualise multiple instruments. As opposed to other research, Hiraga et al. [22] focussed on live performance. Their system analysed MIDI (Musical Instrument Digital Interface) streams from multiple performers. A requirement of the system was to visualise the musical sense of the music perception [22]. This entailed the articulation (legato/staccato), tempo (ritard/accelerate), and dynamic change (crescendo/diminuendo) of each stem. Drawing on this, the iterative development of the MAVIS framework explored the use of differing analytical forms to increase the perceptual immersion of the multi-stem musical recording.

Olowe et al. [23] examined multi-stem audio that could, again, be incorporated into a live performance. As part of a larger system, the researchers produced an application similar in form to that of VJing (Video Jockeying). The structure of the software was to produce a visual accompaniment to the music being played while manipulating a 'curated combination of pictorial media using the performance style of Disc Jockeys (DJs)' [23]. The aim of the system was to enhance the listening experience for non-tonal, non-notated electronic music forms through this novel sound visualisation approach. The researchers used a mixed-methods approach to the study and drew on a number of themes. A significant focus of this work was that while FFT, MIDI, pitch and BPM (Beats Per Minute) were deemed to be important to the final performances, intuition was also valued. The research found that performers were interested in multi-stem visualisation as it was felt it could offer a more engaging experience. This work must be contextualised on the basis of its maturity. More current research has been able to use game engines and computing power to generate immersive visualisations more efficiently. Drawing on Olowe et al's [23] work, it can be summarised that visualisation can add an additional layer to performance with an increased level of immersion.

3.3. Musical Perception

Virtually visualising music has been applied to practical applications of musical understanding and perception. Taenzer et al. [24] used multiple analytical methods to allow the listener to draw a connection to individual instruments in a piece. It is their belief that to transcribe rhythm, harmony and melody, there is a requirement for a certain degree of musical knowledge [24]. The use of visualisation can add understanding of the music, thereby helping reduce the required knowledge. They implemented a system that

compared a number of different techniques based on non-negative matrix factorisation (NMF). The visualisation took the form of lights. The y-axis was used for frequency to move the lights with the centre point at the bottom of the screen. The frequencies that occur less often in music were cut off. The implementation was constructed in the games engine Unity. Lights would move across the screen and fade to make it appear as if they were flying. The system used two datasets, one of which was composed by the researchers. Issues with the experience of a group of participant testers arose when the piece was split down to individual instruments.

> 'The experience was less successful, as the miss rate of the pitch estimation did not allow for traceable light movement in accordance to the played melody, leaving many users confused.'

This research draws on the use of light as a visualisation tool, as other research has done. Whilst important to this study, other research shies away from this [13,25]. This research explores the use of light as one aspect of a possible design framework, acknowledging Taenzer et al.'s [24] work.

Malandrino et al. [26] introduces VisualHarmony, a software tool that uses visual clues, specifically colours, to represent similar tonalities and degrees; they are superimposed on music scores to aid in the harmonic analysis of chorale-style music compositions. The software focuses predominantly on chorale music which is inherently multi-stem based on its multiple voices. The application of the tool is to act as a learning system in education environments. VisualHarmony was given to 60 music experts and music students to test. It was reported that those that tested the visual approach to the learning process were able to compose chorales in less time when compared with participants performing harmonisation in the standard way, and, specifically, composing music with the aid of visualisation elements allowed them to save on average 65% of the time required to perform the same task without visualisation.

3.4. Music Classification

Although not primarily related to the research reported here, the classification of music has elements that relate to those within visualisation. The ability to identify an instrument in a composition is a challenging area. As an application domain for visualisation, musicology necessitates UCD (User-Centred Design) frameworks to mitigate technical barriers, facilitate intuitive data interpretation, and enhance engagement for both domain experts and lay audiences [27]. Han et al. [28] proposed the use of a convolutional neural network (CNN) framework. The use of the framework surpassed other algorithms in recognition of real-world, polyphonic instruments. The focus on the polyphonic nature of music and instruments, and the difficulties in classifying these, demonstrates that the formation of a framework could be a clear aid and an important research area. It has to be understood, however, the focus of this work is outside the scope of classification of instruments within a piece.

3.5. Visualisation Tools and Software

There are a number of tools and applications of software that relate to this area. They encompass software from large developers all the way to indie game and experience creators. Their tool use or visualisation focus demonstrates interest in this area. Max/MSP [29] is a software product that allows for the manipulation of audio-combining effects and multi elements. It additionally has the ability through its Jitter pipeline to create visuals driven by audio. There is embedded within the system the ability to alter meshes and manipulate forms. The visualisation process typically follows a three-step data pipeline:

1. Audio analysis: Objects like the volume and frequency spectrum analyse the incoming sound via Fast Fourier Transforms.

2. Data mapping: These values are scaled and mapped to visual parameters. For example, a heavy bass drum hit might be mapped to the scale attribute of a 3D sphere.
3. Jitter rendering: Using OpenGL, a visualisation scripting language, Max renders 3D objects.

Max/MSP has great strength in visualisation, however it lacks native support of VR and can struggle with large quantities of differing audio stems.

There are a number of tools that are used within industry, specifically performance including VJing that explore the application of visualisation. TouchDesigner [30] is a node-based visual programming environment used by the world's top VJs and installation artists. It has native audio analysis that can isolate and group frequency that is in turn converted to visuals. Additionally it supports many VR headsets. Synesthesia [31] is a dedicated audio visualizer that uses GLSL shaders. It is focussed on ease of use, giving participants the ability to drop audio files in and have visualisation applied to it from a large number of presets.

These visualisation tool kits have also embraced the use of generative AI. SYQEL [32] is an AI-driven browser visualization tool. It uses AI to generate visuals that react to audio frequencies. It works in a browser or as a desktop application and is geared toward ease of use.

In a similar fashion to this software, the use of game engines, as previously discussed, is very prominent in the visualisation of audio, especially when looking to afford immersive interaction and agency to the participants. Epic's Unreal Engine [33] can offer an open-world experience with a wide range of tools to manipulate the environment. It has the ability to create, in real time, complex interactions. Unreal supports a process to automatically extract audio metadata for the use in gameplay as either controlled audio or visualisation. This software allows for non-real- and real-time processing of sounds, allowing interaction into animations, effects and other elements linked to sounds. Unreal can so be seen to be ideal for the development of software for the visualisation of audio stems.

3.6. Research Gap

Based on the examination of the related work, a research gap emerged. This is summarised in Table 1.

Table 1. The elements extracted from the related work demonstrating the research gap.

Area	Focus	Audio Structure	Visualisation	Gap
Hiraga et al. [22]	Live Performance	Multi-instrument	2D/standard display	Focused on MIDI streams rather than complex digital audio stems.
Olowe et al. [23] (VJing)	Electronic (Non-Tonal)	Multi-stem audio	2D/ performance screen	Geared toward electronic music; lacks 3D spatial agency.
Taenzer et al. [24]	General	Split instruments	Unity (3D engine)	Did not implement a 3D virtual world using VR; users reported confusion.
Malandrino et al. [34]	Chorale/Education	Multi-voice (score-based)	Superimposed on scores	Limited to educational harmonic analysis of scores, not immersive spaces.
Max/MSP [35]	General/Creative	Multi-element	3D rendering	Lacks native VR support; struggles with large quantities of orchestral stems.
Synesthesia/ SYQEL [31]	Electronic/General	Single/mixed audio	2D/browser/ presets	Primarily passive observation; lacks spatial agency for complex orchestral structures.
Philharmonia Experience [9]	Orchestral	Multi-stem	Immersive	Replicates real-world interactions; lacks participant narrative agency.
MAVIS Framework (Proposed)	**Orchestral**	**Complex multi-stems**	**Immersive 3D VR**	**Resolves the gap by providing agency, complexity control, and semantic consistency for large-scale orchestral audio.**

4. Materials and Methods

4.1. Design Theory

The design of the methods used here are grounded in a Design Science Research (DSR) three-stage problem-solving paradigm [36]. This consists of:

1. The Relevance Cycle: Connects the research to the real-world environment.
2. The Rigour Cycle: Draws from existing scientific knowledge and adds new findings back into the knowledge base.
3. The Design Cycle: The artefact is iteratively built and reflected upon.

This design structure was applied as an umbrella methodology with the incorporation of the hybrid Agile–Double Diamond approach at the heart. As discussed, the application development method drew from the work of Liang et al. [6]. A hybrid Agile–Double Diamond [7] design methodology was adopted for the development process allowing for participant-centred, fast prototyping. The hybrid approach has benefits over the use of either Agile development or Double Diamond design alone as it combines the strategic, problem-solving rigour of a methodology developed by the British Design Council with the iterative execution of Agile. Often within a purely Agile approach the process can quickly focus on the how before significant exploration of requirements. The Double Diamond forces there to be a Discovery and Definition phase. This methodology adopts a dual-track Agile approach whereby designers and researchers work on an initial phase of sprints to validate ideas before a second stage that validates the outcomes of the first. This can eliminate the possibility of directionless development where design and implementation struggle to focus on the final goal.

4.2. Feedback to Experiences

There are a plethora of different techniques that can be used to gain feedback to multi-stem music experience. These can be categorised as:

1. Pulse;
2. Breath rhythm;
3. Eye movements.

There are both pros and cons to the implementation of these techniques and the technology that is required for them; see Table 2.

Table 2. Quantitative assessment of somatic effects in virtual reality visualisation.

Metric	Pros (Quantitative Value)	Cons (Noise and Interpretation)
Pulse/Heart Rate Variability	Validated proxy for cognitive load; measures physiological arousal.	Highly sensitive to physical motion; skewed by caffeine or baseline fitness.
Breath Rhythm	Linked to emotional regulation; indicates flow state and immersion.	Hardware (straps) can be intrusive; easily manipulated by conscious effort.
Eye Movements	Maps attentional focus; validates audio–visual spatial mapping.	Inattentional blindness can occur; eye-tracking hardware increases cost.

A number of factors in the initial use cases determined that pulse, breath rhythm and eye movement were not to be used. As outlined, heart rate variability, despite being a good indication of cognitive load in some experimental situations, is additionally sensitive to physical movement. The experiences that were developed were room-scale, affording the participant to physically navigate the space. This may induce changes in the heart rate that are not linked to the changes in the visualisation and musical interaction. The technologies

used for heart rate monitoring and breath rhythm can be complex and sensitive. The nature of the participant recruitment entailed that the technologies used were in constrained but public spaces. As such, these environments can interfere and cause issues with setting up sensitive technology. In terms of eye movement, the use was impacted by the nature of the VR headsets used. Neither of the Meta Quests utilised have the capacity to observe changes in eye movement or saccades. Although it is acknowledged that current VR headsets have incorporated this technology such as Varjo XR-4 [37] and the HTC Vive Focus Vision [38].

4.3. Implementation Approach

The overall approach and form of the experiences was created through iterative development. Initial structures derived from the related work reviewed formed an outline structure. The framework used audio analysis to output data that influenced an immersive experience. To develop the applications of the frameworks, Epic's game engine, Unreal, was used [33]. In a similar proposal to the research, Choy and Reich [39] discussed the use of a system to visualise music through the use of a games engine so indicating the validity of using this technology. They created a 'ring of cylinder bars indicating the audio frequency spectrum and visual effects indicating the pitch values' [39]. As a tool, Unreal allows the developer to interact with sound directly and to analyse its structure using built-in tools facilitating quick prototyping and the adoption of the hybrid Double Diamond approach. The process of development of VR experiences is optimised by the use of the Unreal VR camera and interaction tools. Other games engines have the capability to analyse a sound and visualise it in a VR environment; however, Unreal was chosen for its sophistication, ease of use and the underlining skills of the researchers.

4.4. Survey Design

The survey utilised a mixed-methods approach with a convergent–parallel approach [40]. Five quantitative questions were asked alongside eight qualitative. These were:

Quantitative questions:

1. Age?
2. How often do you listen to music?
3. How knowledgeable are you about orchestral music?
4. How experienced are you in gaming?
5. How experienced are you in VR?

Qualitative questions:

1. Were you able to investigate the environment well for example navigate?
2. Was the overall objective of the experience clear?
3. To what extent did you get a perception of the individual instruments?
4. Was the experience easy to engage with?
5. What elements of the experience did you feel were most successful?
6. Were you able to interact with the musical stems as you would expect?
7. What would you like to see more of in the experience?
8. Any other comments?

Breaking these down, the quantitative questions focussed on the participants interaction with both music and the technology that would be used in the experience. The questions tried to extract an understanding of how the participant would understand the visualisation of the stems, especially as they were orchestral. As VR experiences can instigate highly emotional reactions [41], an understanding of the level of previous use that the participants have had was gained. Whilst subjective, the researchers were exploring any correlation between VR experience and overall interaction.

The qualitative questions were designed to explore the participants interaction to a greater degree, in essence the "why". The initial questions were used to explore the utilisation of the participants interaction with the software itself. By integrating the participants sense of movement, exploration and clarity of stem sounds, feedback could be used to reevaluate the design frameworks structure to iterate into the next structure. The subsequent questions focussed more on the engagement and satisfaction level of the overall experience.

4.5. Participant Recruitment

The recruitment of the participants was defined by the locality and availability coupled with a focus on participants that are either interested or have some expertise in the area. The research undertook an open-call, purposive sampling approach [42]. In the context of immersive experiences (like VR, AR, or interactive exhibits), this allows for the provision of a naturalistic setting where the target population already exists. Media mail shots were used to draw participants that were either from academic cohorts or members of the public visiting an exhibition. While direct recruitment offers more control, the purposive sampling approach has unique strengths, especially for immersive technologies. Direct recruitment often happens in a sterile lab setting where participants feel observed. By advertising a public experience, it is possible for people to interact with the technology in a naturalistic environment. This provides a more accurate picture of real-world interaction.

Direct recruits often feel a subconscious pressure to perform or provide correct answers to positively engage with the researcher. In an open-call immersive experience, participants can arrive with a consumer mindset rather than a test-subject mindset, leading to more authentic reactions.

The resultant recruits predominantly came from either an academic or student background with interests in XR, music, gaming or creative technologies or public groups that had interest in XR experiences or applications of creative technology.

4.6. Structure: Neptune Recording

The MAVIS design framework structure was formed from findings through its application on two differing musical compositions. In this section the authors will discuss the methods and materials used to construct these.

The initial development was applied to the first of two pieces of work, Gustav Holst's Neptune from The Planets. The subsequent use case was a section of the opera, Bluebeard's Castle by Béla Barótk. Individual recordings of separate instruments were taken during the AOTF project. The recording process was two-fold with both being undertaken by Punchdrunk Theatre [43]. From these music stems, the virtual experiences was developed. The structure of the environments took two different approaches. Interaction with Holst's Neptune was abstracted into different coloured cubes representing different instruments. An experience was generated consisting of a virtual world where the standard orchestral positioning form was disrupted. Instead of the ordinary crescent structure, instruments were placed in groupings within a circular form. Each instrument of the audio experience was processed as a unique spatialised source. The participant in the experience was able to perceive changes in distance and location of the sound influencing how they explored the world. By allowing participants to perceive these changes as they moved, real-world acoustic behaviour was mirrored. This spatial consistency was targeted at reducing the cognitive effort needed to process the environment, thereby facilitating the shift in attentional state required for immersion. In Figure 1, the structure of the cubes within the virtual world can be seen. The world itself is simplistic in form to work alongside the

aesthetic of the whole interaction. Figure 1 shows the individual representation of sounds. Each cube is either:

- Green representing strings;
- Yellow representing brass;
- Red representing percussion;
- Dark blue representing vocals;
- Light blue representing woodwind;
- Pink representing a combination of harps, organ and celeste.

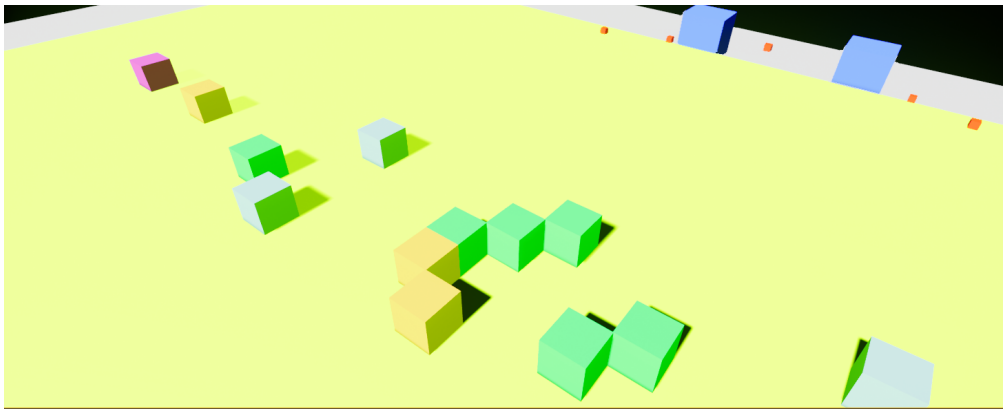

Figure 1. The audio stems are visualised as cubes within the Neptune Experience virtual environment, where each colour represents a separate instrument collection.

Figure 2 shows the overall layout of the sounds. Each section contains a grouping of cubes that relate to the single instrument or classification of instruments such as the strings.

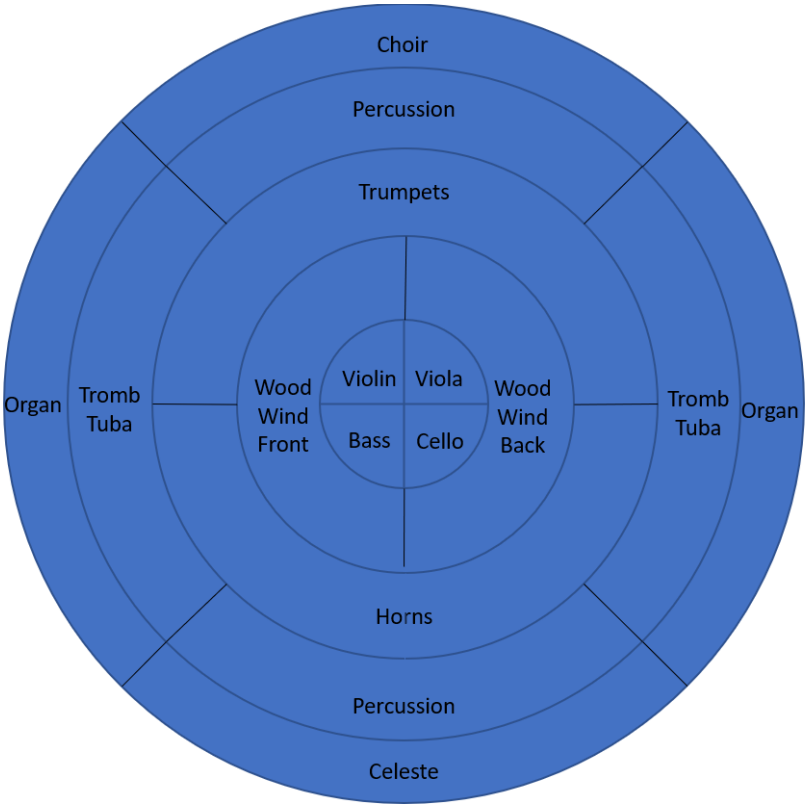

Figure 2. Layout of the instruments as part of the design structure for the Neptune experience.

The interaction with each cube differs based on its sound type. The majority of the sound cubes can be picked up using the standard VR hand interaction of Unreal's OpenXR implementation [44] and moved by the participant allowing for sound play. The participant is able to teleport to any location in the experience and pickup the cubes, moving them or throwing to any location within the space. They have the ability to construct their own experience of the piece with the placement of the individual instruments. As well as this, the vocals and percussion instrument audio cubes are static items and immovable by the participant; however, they are dynamically altered in appearance based on the processing of the semi-amplitude of the sound using Discrete Fourier Transform (DFT). Based on previous research and the use of DFT in the Unreal Engine, this process was chosen. DFT has the ability to capture increases and decreases in specific frequencies of a sound. In the context of visualising audio, a participant can view how a sound alters through real time, offering insight into its structure. To facilitate the visualisation, each audio stem is processed through a Fast Fourier Transform (FFT). The system calculates the average semi-amplitude (A_{avg}) across the defined frequency spectrum. This discrete amplitude value A_{avg}, derived from the summation of frequency bins k in the DFT, serves as the primary driver for the visual transformations. DFT can be formally defined as;

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi}{N}kn} \quad (1)$$

where n represents the discrete time-domain samples within the 0.1 s buffer. This data is then mapped to the physical wave properties of the virtual objects.

The form of the sound wave is captured over a set period of a 10th of a second giving the sense of real time. The output from the FFT is taken against three frequency ranges split evenly across the 20 Hz to 20,000 Hz human hearing range. An average of the semi-amplitude is calculated for three separate frequency buckets. These frequency buckets are used to alter the of dimensions of an object in the Unreal virtual environment. Figure 3 shows how the audio stem is mapped to a game object via the use of this analysis.

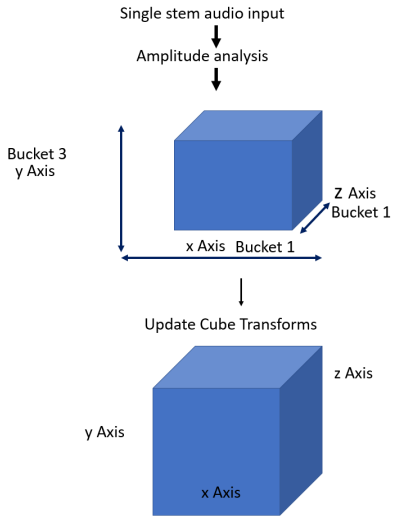

Figure 3. Process to transform the cubes axis based on the analysis of the audio stems using the DFT algorithm.

The semi-amplitude is defined as being half of the peak-to-peak amplitude, from either the crest or trough to equilibrium. Figure 4 is a formal definition of a wave and its amplitude, with the equation

$$x = A\sin(w[t - K]) + b,$$

where

- A is the amplitude;
- x is the oscillating variable;
- w is the angular frequency;
- t is time;
- K and b are arbitrary constants representing time and displacement offsets.

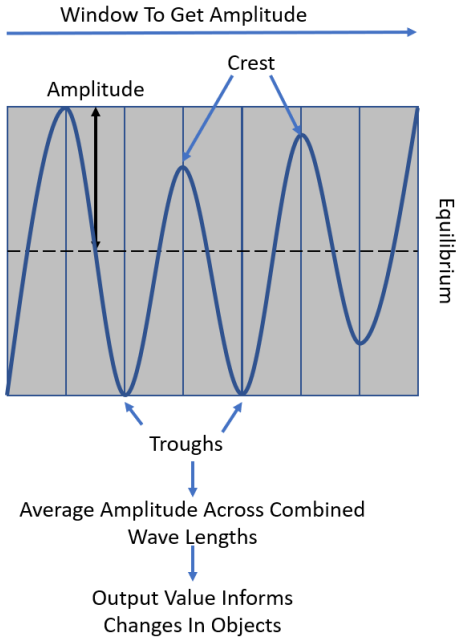

Figure 4. Output of the semi-amplitude process as it is processed through the DFT algorithm.

The process can be summarised as in Figure 5.

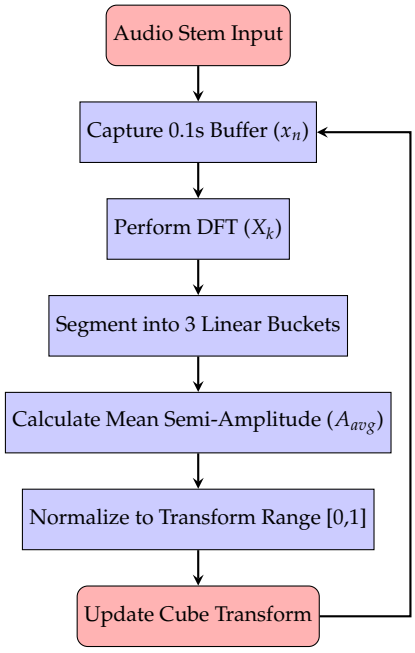

Figure 5. Logic flow for the real-time processing of audio stems.

The application of the DFT affords the participant to experience each instrument visually across defined discretised frequencies. Often used for the analysis of musical chords [45], the use of DFT have been implemented into systems to represent musical

knowledge in such a fashion that a listener with limited musical understanding can correlate what they have heard to what they see, increasing their understanding of the music [24]. Adding to the use of a DFT to analyse the musical stems, each of the audio stems was spatialised, allowing the participant to interact with them. The spatialisation form was an inner and outer hemisphere. The initial, inner hemisphere gave an unchanging audio stem to the participant. The distance to the second created a fall-off audio volume that reached zero. This fall off was logarithmic in nature. The values used were chosen based on the ability of the participant to be able to experience the whole composition from their initial starting position. The use of a single volume value for the central hemisphere afforded consistency in the experience for participants through a restricted barrier that allowed for participants attention to be kept on the audio. This mechanic allowed the participants to have agency over the experience and construct their own audio narrative by physically moving toward or away from a source. This design allows for complexity control, an aspect of cognitive scaffolding.

Within the structure of Unreal Engine, the output from the DFT was passed to the scaling of the dynamic objects. Each of the three frequency buckets drove the x, y and z transform of the object, giving a visual relationship to the audio stem.

4.7. Structure: Bluebeard's Castle Recording

A second iteration of the framework was created. A recording of Bluebeard's Castle was used for this case study. The form of the framework itself was altered, building on the previous implementation and feedback from the participant testing (see Section 5.1.1 for more details). To explore further the application of visualisation in an experience, a less abstract approach was taken. The opera, Blue Beard's Castle by Béla Bartók is made up of seven sections that each reference a door. They are composed of:

1. The torture chamber;
2. The armoury;
3. The treasury;
4. The garden;
5. The kingdom;
6. The pool of tears;
7. The wives.

The fourth door of the composition, The garden, was chosen as it is a relatively short section of the composition and would be acceptable to most audiences due to its form and style. The whole piece was split to leave only the required amount of audio. No other alteration to the audio stems was made. Following the same base framework structure as used for Holst's Neptune, each of the stems were analysed using the defined process. Additional layers were added to the basic framework in order to accommodate the greater complexity of the piece and the utilisation of a more descriptive environment. To explore the use of the visualisation framework across different forms of environment, rather than sound cubes, a forest of trees was chosen. As with the colour of the cubes, each tree type represented a separate musical section. Figure 6 shows a section of the Door 4 experience; here, the use of a more descriptive structure and the use of light to visualise the audio.

The sections were represented as:

- Medium yellow birch tree representing strings;
- Medium red birch tree representing brass;
- Large red birch representing flute;
- Large green birch representing woodwind;
- Medium brown bush representing harp.

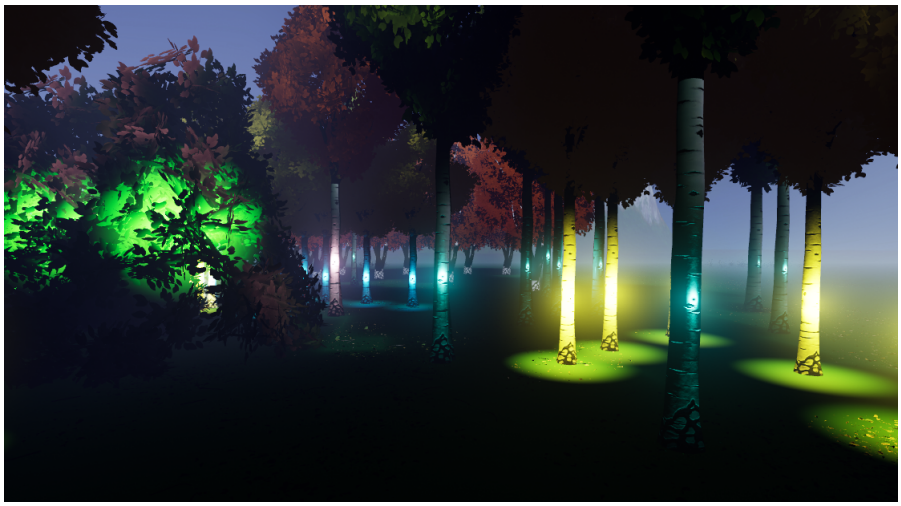

Figure 6. A section of the forest in the Door 4 experience as visualised within a VR headset.

In total, 51 instruments were recorded by the Philharmonia Orchestra [46], making up the forest. The recordings were made using a different method to Neptune. Greater isolation was achieved, and, as a result, greater clarity was also achieved. The positioning of the trees, as with Neptune, departed from the standard crescent shape. The forest, however, by its nature, had the trees placed in a more sporadic manner although clustered into sections. Figure 7 shows the layout of the forest.

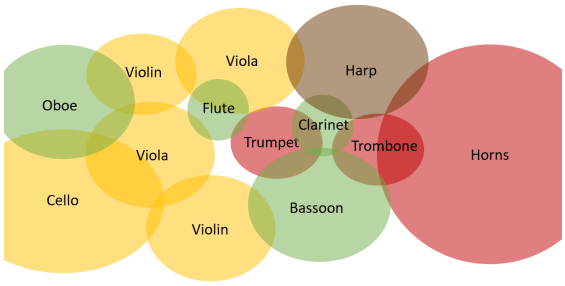

Figure 7. Layout of the Door 4 instruments in the virtual experience. This shows the diversity of distribution as opposed to the Neptune experience. Each colour relates to a different instrument collection mirroring the Neptune experience structure.

Each of the audio stems was analysed, splitting it into nine buckets. This was to allow the visualisation to be focused on the use of light. Again, in a similar way to the Planets implementation, Door 4 used amplitude to visualise the stems. The buckets influenced the attenuation radius, the source radius and the intensity of the light. Each tree was preassigned a colour. Initial experiments were carried out using a change in colour based on the amplitude. This proved to be confusing when trying to separate the sounds.

The buckets were processed so that:

- Bucket 1–3 generated the attenuation radius;
- Bucket 4–6 generated the source radius;
- Bucket 7–9 generated the intensity.

The attenuation radius clamps the light to a specific distance so as not to impact the participants visualisation should the amplitude be too large. The intensity is the total energy that the light emits. This was set to use lumen. The source of the radius had the most influence on the experience as it defines the distance from the individual tree of the light. Figure 8 shows the output flow of the analysis and the resultant output.

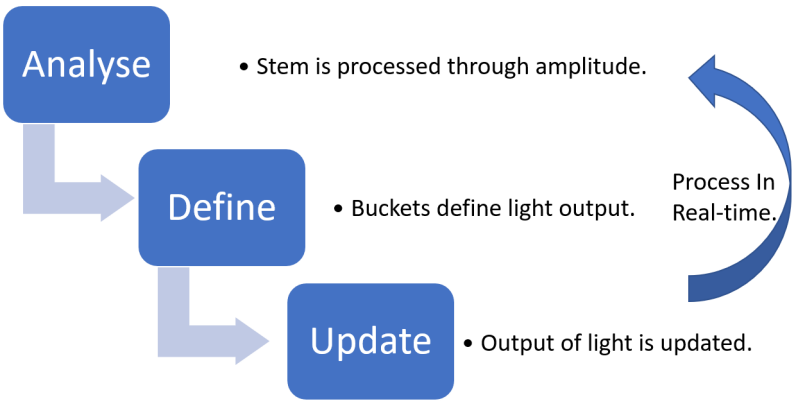

Figure 8. Flow of the door analysis process.

5. Results

5.1. Survey Feedback

The experiences created using the framework were given to two sets of individuals to record their interactions with the piece. This was split into a single group for each of the experiences. A combination of quantitative and qualitative questions were used, with a focus on qualitative.

5.1.1. Results from the Planets: Neptune Survey

For this survey, nine individuals engaged with the experience. Each person interacted through the use of an Oculus Quest headset head-mounted display (HMD). The system was setup as room-scale VR. The participant was able to walk around the space and also teleport to various areas. The experience had a restricted barrier in order to allow both exploration whilst maintaining a defined virtual area of the audio environment. As discussed in Section 4.4, five quantitative questions were asked in order to understand some of the background to the participant. Aside from the question regarding age, the other questions used a Likert scale of five points. The mean (M) age of the participants was 34 years with a standard deviation (SD) of 9.70, with the overall range between 21 and 52 years. Aside from a single individual, all participants recorded that they often listened to music (a M = 4.25 and a SD = 0.96). This infers that the participants have an understanding of musical composition, however, it is not known whether this relates to only non-western forms. As Neptune is a western musical form, orchestral musical compositions that encompass other regions may produce different results. There was a lower knowledge of orchestral music overall. This ranged from one to four. There was a correlation, however, between those that scored highly for listening frequency and those with higher knowledge of orchestral music.

Overall, there was a lower experience of gaming then other areas. A strong correlation was seen between a greater experience of gaming and that of VR. Although no solid conclusion can be drawn from this, it offers an area for further investigation as certain individuals did not see their interaction with VR as one that is gaming-related despite the possibility that their previous experiences were generated using a games engine for development. Figure 9 shows the correlation between participant gaming and VR experience.

In addition to the quantitative questions, eight qualitative questions were asked. These related to the ability to interact with the environment and the participants overall experience. To assist the participants, a short verbal tutorial was given to show how to use the VR HMD and general interaction with the environment. It did not cover, however, the objectives or what the experience may hold. It was clear from the comments that this

was a necessary addition. When asked was it clear what you were meant to do, it was commented by one participant that

> 'It wasn't particularly clear what the objective was. (It) could have done with a short tutorial.'

This is clearly an area that needs to be addressed. Despite there being a brief onboarding process to make people comfortable with a VR experience, an additional element is needed for either feedback within the experience or a more sophisticated onboarding structure.

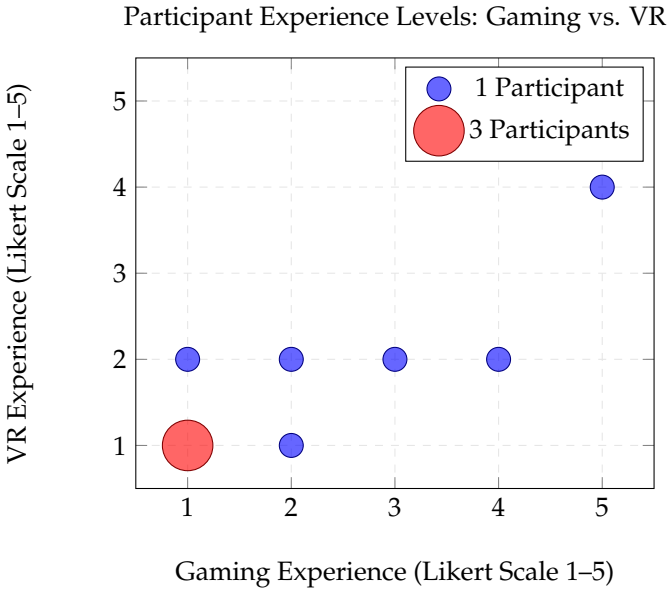

Figure 9. Weighted scatter plot showing participant expertise ($N = 9$). The red bubble highlights the highest frequency of participants ($n = 3$) reporting minimal experience (1,1) in both gaming and VR environments.

The participants perception of the navigation of the environment and its physical–virtual structure was of clarity. It can be inferred that the design afforded the participants this knowledge based on the interaction carried out in the movement, combined with the simplistic nature of the virtual environment itself.

Issues arose with the identification of the various instruments in the experience. This is a fundamental aspect of the experience. A goal of this first incarnation was to be able to isolate the different instruments in order to engage with an alternative orchestral structure. It was found that the recordings of the instruments contained a portion of bleed, disturbing the ability to isolate them.

The majority of the participants commented that it was most enjoyable to engage with the moving cubes. These comments showed that this was an important element. Moving forward, this approach was further examined and incorporated into the mechanics of the second iteration to produce the overall design framework structure.

To summarise, the participants were asked what they would like to see that was not offered. The answers to this were varied; however, they were used to redesign aspects of the framework structure. Participants requested to see further visualisation of the instruments with a focus on differing levels of colour intensity as the audio changes. As a mechanic, this was adopted in the second iteration. The greater number of instruments and the improved isolation of the second recorded stems lent itself to the adoption of this.

5.1.2. Results from Blue Beard's Castle: The Garden Survey

For Door 4, again nine participants, although at a different location, were asked the same questions. Each person interacted with the experience through the use of an Oculus Quest HMD. Room-scale was used, allowing the participant to be able to walk around the space and also teleport to various areas. The average age was 32 years, ranging from 20 to 51 years (SD = 9.65). As with the Neptune tests, there was a high value returned for the frequency of listening to music, an average of 4.33 (SD = 0.67). This was in contrast to orchestral music. There was a markedly lower average of 2.0 (SD = 0.82). Fewer people also had a large experience of interaction with VR. Four participants recorded a score of 1 with an average of 1.89 (SD = 0.99). Despite this, in the feedback from the question "Was the overall objective of the experience clear?", the majority commented that they understood what they were required to do, although taking analysis from the previous frameworks incarnation, a larger verbal explanation was given. All participants were positive when commenting on their ability to navigate the environment despite reporting an overall lack of knowledge of VR outside of the participation in this study.

When asked about interaction with the music stems, issues with recognition of the individual instruments came to light again. In this implementation of the framework a more complex visualisation was used. Despite the use of stems that had greater clarity, it is believed that the complexity of the environment caused issues.

The immersive aspect of the experience was seen as a great attribute to engaging with music in this way. The main thoughts focused on further interaction, immersion and agency. One participant commented that

> 'It would be interesting to be able to engage with the environment a bit more—go inside bushes and be surrounded by louder sounds'

These comments, combined with those from the previous experience, have informed some of the structure of the design framework that was able to be drawn out.

6. Discussion

6.1. Learning Approaches

Cognitive Scaffolding

Based on the findings of the initial development phase, we propose the MAVIS (Multi-stem Audio Visualisation in Immersive Spaces) design-orientated framework to bridge the gap between passive observation and active creation. This framework is based on three pillars: Semantic Consistency, Spatial Agency, and Bimodal Feedback.

6.2. Semantic Consistency and Visual Metaphors

Emerging from the findings of the Planets and Bluebeard's Castle framework development, one barrier to multi-stem visualisation is the arbitrary nature of mapping. This constitutes ensuring that the system's beliefs don't contradict one another. If the participant identifies an object as a tree, they must consistently expect the properties of treeness (height, texture, leaf type). It was found that to convey a consistent understanding of the instruments to be visualised, there needed to be a mapping of the sounds to specific sections of the orchestra, so migrating from an arbitrary structure. Extracting this from the Planets framework, a set of colours are suggested to be used:

- Green representing strings;
- Yellow representing brass;
- Red representing percussion;
- Dark blue representing vocals;
- Light blue representing woodwind;

- Pink representing a combination of harps, organ and celeste.

Although this mapping is not wholly ubiquitous as many orchestras are formed from differing types of instruments and amounts of participants, it is in line with the definition used in this research, that being an orchestra is an ensemble that contains areas such as strings, brass and percussion, and that the piece is written for a standardised structure [11]. Beyond the colour representation, the framework proposes a mapping of the stems to instrument visualisation through alterations of the visualised form. Again drawing from the Planets framework, abstract geometric shapes are used.

Figure 10 shows the analysis of a sound stem and the resultant visualisation input for a cube shape. This is the foundation of the 1st pillar of the framework.

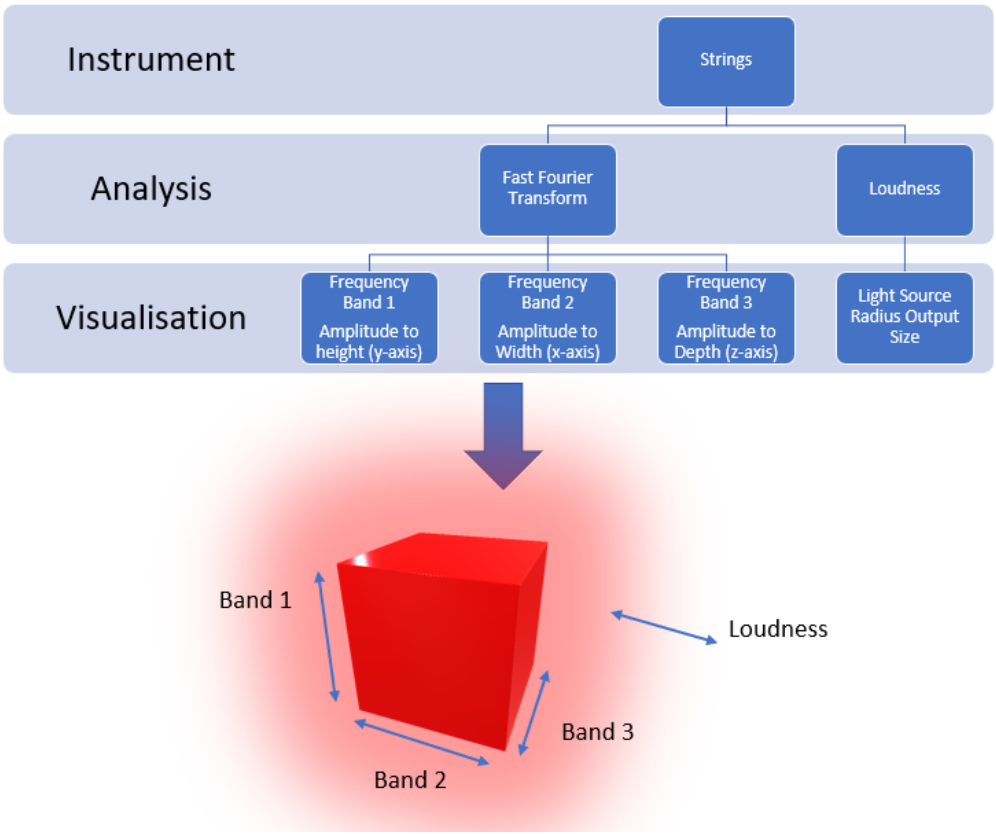

Figure 10. MAVIS framework approach to shape alteration and light changing.

6.3. Spatial Agency and Proxemic Interaction

In mixed reality (MR) environments, interaction should be governed by the participant's physical proximity to the stem object, allowing for greater spatial agency based on the proxemic interaction. Based on the second development phase, MAVIS incorporates the application of attenuation of the sounds to allow the participant to gain agency over the experience. Exploration and manipulation of the stems within the virtual space affords the participant the ability to construct their own musical composition. Overall the framework emphasises playing in the space allowing participants to move and manipulate specific music stems. The Neptune case study suggested that participants find picking up and moving audio cubes highly interesting and engaging. This design aligns with the use of proxemic interaction to afford the participant greater agency. Proxemic interaction has four zones:

1. Intimate: Defined as being for high sensory input. In this context the participant grabs or physically alters the audio stem.

2. Personal: Defined as being for friends and family. In this zone the stem reacts to the participant's presence (the audio and/or visual aspects of the stem alter).
3. Social: Defined as formal or professional interaction. Within the framework this interaction alters the stem as the participant moves away or towards it.
4. Public: This is beyond the circle of direct involvement. As such the audio is heard as a collective mix with no focus.

Through the application of these zones, there is a shift from the observation to participation. As the participant moves from the public zone to the intimate zone, they transition from an audience member to a conductor. As this process is based on proxemic zones, there is less of a need for on boarding or tutorial. The intrinsic nature of the knowledge required means they instinctively know that moving closer to an object in 3D space should make that object more important or interactive

6.4. Complexity Control and Immersion

Based on the Bluebeard study, using light intensity and radius rather than shifting colours reduces confusion when separating multiple audio sources. As such, complexity must be strictly controlled to prevent sensory overload. Developers should allow participants to explore environments to simulate an open-world feel, but space must be restricted to maintain focus on the intended audio–visual interactions.

When considering the virtual space that orchestral, multi-stem virtual environments encompass, participants can find objectives unclear in these abstract VR spaces, even with a strong on boarding process. To counter this, the MAVIS framework incorporates the use of cognitive scaffolding, incorporating just-in-time [47] information cues to afford swift engagement with the sounds. In any virtual experience, there is a tension between representing accuracy and being overwhelming. The focus is to minimise surprise whilst affording the participant the ability to engage.

6.5. Design Structure

A number of design suggestions emerged from the development of the first and second design framework iterations. The structure of the framework can be summarised as follows:

- To implement multi-stem visualisation in VR, complexity needs to be controlled. The visualisation should not detract from the musical form, but rather enhance it. The participant should be able to experience the virtual world without distraction from the exploration and understanding of the audio stems.
- Those with a small quantity of VR interaction would benefit from an initial tutorial to help with a swift engagement with the sounds. This can take the form of a defined tutorial or, to try to aid immersion, the use of practices such as cognitive scaffolding.
- Allow the participant to explore the environment to infer an open world, but restrict the open space so that a focus can occur on the desired outcomes.
- To assist the participant in understanding and forming a unique interaction with multi-stem music, mechanics that allow the participant to 'play' in the space should be explored. The ability to move sound, manipulate sound and remove sound all emerged from the participant testing as areas enhance.
- To assist in the understanding of orchestral composition form via multi-stem visualisation, there is a need for a defined mapping of sounds to colours.

It was seen that a failure of the design framework was the understanding of participants hearing range as this may influence the immersion and understanding of the virtual environment. As such, as part of an audio onboarding process a hearing test would be administered.

Figure 11 shows the interaction structure of the final design framework iteration.

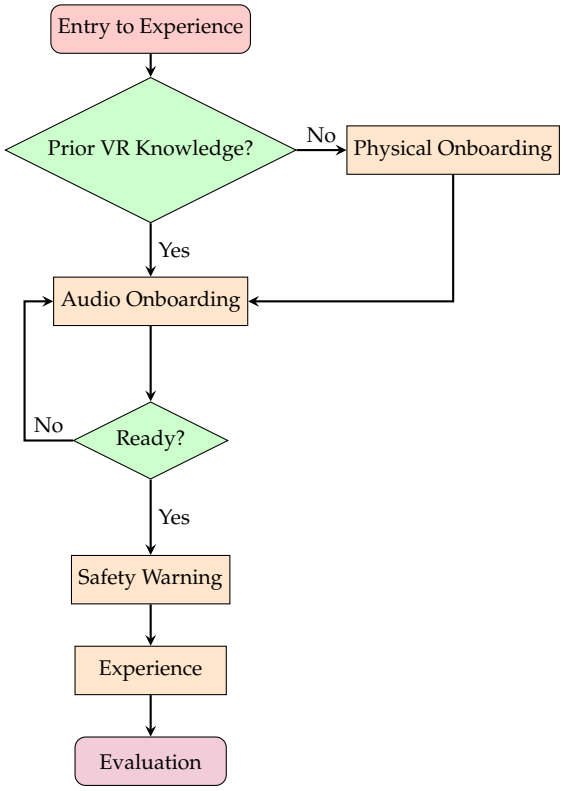

Figure 11. A flow diagram of the proposed participant interaction with the summative technological artefact based on the MAVIS framework.

7. Contribution to Knowledge

In answering the research questions posed, the contributions to knowledge can be extracted.

Q1: Through an exploratory evaluation, how can we design new ways for people to interact across multimodal structures with sound in VR that go beyond real-world actions, allowing them to shape their own unique experience? The research explores interaction methods that allow users to move beyond real-world actions like passive listening to forge their own experience. A key design rule extracted from the two use cases was that providing spatial agency, where participants can physically move through the virtual space and manipulate audio stems, aided immersion and afforded a way for participants to forge their own narrative. In the provided prototype, participants can pick up, move, or throw instrument cubes, effectively reconstructing the musical arrangement to create their own unique version of the piece. The framework employs a data pipeline where audio analysis is mapped directly to visual parameters. By migrating away from wholly replicating real-world interactions, the framework specifically aims to afford narrative autonomy, allowing even non-experts to interact with complex musical structures in a way that feels like active creation rather than passive observation.

Q2: Through an exploratory evaluation, what design rules help organise complex, multi-layered audio in VR so that it makes sense to a participant without simply replicating a real-world structure? The MAVIS design framework offers three areas that can assist designers in forming new ways to interact with multimodal interactions. These can be summarised as:

1. Semantic Consistency (Visual Grammar): Rather than mimicking physical instruments, the framework uses a consistent visual grammar. For example, specific in-

strument families are mapped to distinct colours—strings are green, brass is yellow, and percussion is red. This helps participants make sense of the complexity without needing to see a literal violin or trumpet.

2. Abstract Spatial Layout: In the Neptune use case, the standard crescent-shaped orchestral layout is disrupted. Instruments are instead placed in abstract groupings within a circular form, allowing the participant to explore the composition as a spatialised environment rather than a static performance.
3. Complexity Control: To prevent sensory overload from multiple simultaneous audio stems, the framework uses cognitive scaffolding and visual constraints, such as adjusting light intensity, to guide the participant's focus.

There are areas, however, which the research did not satisfy to answer the questions. The final delivery phase was only evaluated by a small participant size. This limits the generalisability of the design rules, as a broader demographic might interpret the abstract visual grammar differently. Limitations of the hardware did not allow for fully validating how attentional focus maps to audio–visual spatial arrangements. As discussed in Section 9, this is an area for additional research. A goal of the framework is user-driven narrative autonomy whilst maintaining complexity control. It must be acknowledged that there is a risk that these constraints might limit the very agency the framework seeks to provide, potentially steering the participant back toward a more guided experience rather than a truly unique one. While the framework proposes semantic consistency, it can be deemed that what feels intuitive to one participant may still be confusing to another.

Overall, the implementational contributions of this work are: (1) the development of the MAVIS (Music Audio–Visual Immersive System) framework; (2) a novel mapping logic for orchestral stems in Unreal Engine; and (3) an empirical evaluation of participant presence and accuracy in immersive music environments.

8. Conclusions

This research established a framework to determine a design-orientated framework for the virtual reality (VR) visualisation of multi-stem audio. Two distinct VR experiences were developed using different orchestral compositions. Participant feedback and interaction data were analysed to refine the initial framework into its final structure. Qualitative analysis of the first iteration revealed that the framework successfully facilitated spatial exploration and participant interaction. Furthermore, participants reported a high degree of virtual agency, particularly when reconstructing the orchestral layout, which allowed them to reimagine traditional modes of musical engagement.

Despite the positive feedback, several participants indicated that the experience would have benefited from a formal onboarding process. Such an introduction would assist participants in mastering the hardware interface and clarifying the experimental objectives, thereby allowing for better integration of multimodal interactions.

To evaluate the framework's robustness and transferability, a second musical composition was processed. This iteration involved nine participants and utilised a more descriptive visualisation, a forest environment, contrasting with the previous Neptune experience. In this version, dynamic lighting was employed to represent audio stems. Results suggest that this visual feedback marginally improved participant comprehension of the relationship between the environment and the audio structure.

Overall, the framework suggests potential for cross-composition transferability and effective visualisation strategies. While some participants lacked clear instrument identification, the Bluebeard experience demonstrated a possible ability to enhance, at a base level, immersion during musical interaction.

Elements from the Neptune and Bluebeard frameworks were compiled into a cohesive framework that acts as a design intervention when developing visualisation of multi-stem audio orchestral recordings. This design-oriented framework was implemented in a final version as a proof-of-concept again using Unreal Engine by applying the three pillar structure.

9. Future Work

In the next phase, the MAVIS (Multimodal Audio–Visual Interaction System) design framework will be validated through the development and application of a final multi-stem orchestral visualisation experience. These findings will feed into the current design approach. Larger participant cohorts will be employed across a number of varying musical compositions to facilitate a deeper longitudinal understanding of the design framework's transferability and scalability across diverse orchestral contexts. To further examine the generalisation ability of the framework, compositions that sit outside of the standard western orchestral form will be explored. A short coming of the research was the equipment available to draw a quantitative analysis of the participant interactions with the experiences. Further development will look to incorporate technologies that offer the recording of somatic effects including the capturing of eye movement.

Author Contributions: Conceptualization, J.S. and S.S.; methodology, J.S. and S.S.; software, Unreal 5.2 , J.S.; validation, J.S.; investigation, J.S. and S.S.; writing—original draft preparation, J.S. and S.S.; writing—review and editing, J.S. and S.S.; funding acquisition, J.S. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the AHRC (Arts and Humanities Research Council) Audience of the Future fund.

Data Availability Statement: The datasets presented in this article are not readily available as the data is part of performers rights. Requests to access the datasets should be directed to Jethro Shell at jethros@dmu.ac.uk.

Acknowledgments: The authors would like to acknowledge the support of the Philharmonia Orchestra in the production of the stems for this research. Although involved in the funded project, their input in this aspect was extremely helpful.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CLT	Cognitive Load Theory
DSR	Design Science Research
FFT	Fast Fourier Transform
HMD	Head-Mounted Display
MAVIS	Multi-stem Audio Visualisation in Immersive Spaces Framework
MIDI	Musical Instrument Digital Interface
MR	Mixed Reality
M	Mean
SD	Standard Deviation
VR	Virtual Reality
ZPD	Zone of Proximal Development

References

1. Kaper, H.G. *Manifold Compositions, Music Visualization, and Scientific Sonification in an Immersive Virtual-Reality Environment*; Technical Report; Argonne National Lab.: Lemont, IL, USA, 1998.
2. Lima, H.B.; Santos, C.G.R.D.; Meiguins, B.S. A survey of music visualization techniques. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–29 [CrossRef]
3. Chen, C.H.; Weng, M.F.; Jeng, S.K.; Chuang, Y.Y. Emotion-based music visualization using photos. In *Proceedings of the International Conference on Multimedia Modeling, Kyoto, Japan, 9–11 January 2008* ; Springer: Berlin/Heidelberg, Germany, 2008; pp. 358–368.
4. Smith, S.M.; Williams, G.N. A visualization of music. In Proceedings of the IEEE Proceedings. Visualization'97 (Cat. No. 97CB36155), Phoenix, AZ, USA, 24 October 1997; pp. 499–503.
5. Fonteles, J.H.; Rodrigues, M.A.F.; Basso, V.E.D. Creating and evaluating a particle system for music visualization. *J. Vis. Lang. Comput.* **2013**, *24*, 472–482. [CrossRef]
6. Liang, R.; Chu, S.; Lawton, D.; Pan, G. Human-centered design based on the double diamond model for optimizing hybrid game design. In *Human Factors in Design, Engineering, and Computing*; AHFE Open Access: New York, NY, USA, 2024; Volume 1.
7. Design Council. The Double Diamond. Design Council—designcouncil.org.uk. 2026. Available online: https://www.designcouncil.org.uk/our-resources/the-double-diamond/ (accessed on 9 February 2026).
8. www.wevolver.com. Meta Quest 2. 2026. wevolver.com. Available online: https://www.wevolver.com/specs/meta-quest-2 (accessed on 9 February 2026).
9. Philharmonia. Immersive. 2026. Available online: https://philharmonia.co.uk/what-we-do/digital-immersive/immersive/ (accessed on 20 March 2026).
10. De Montfort University. DMU to Help Shape How Audiences Experience Live Performance. 2019. Available online: https://www.dmu.ac.uk/about-dmu/news/2019/january/dmu-to-help-shape-how-audiences-experience-live-performance.aspx (accessed on 27 February 2026).
11. Spitzer, J.; Zaslaw, N. *The Birth of the Orchestra: History of an Institution, 1650–1815*; Oxford University Press: Oxford, UK , 2004.
12. Agrawal, S.; Simon, A.; Bech, S.; Bærentsen, K.; Forchhammer, S. Defining immersion: Literature review and implications for research on immersive audiovisual experiences. *J. Audio Eng. Soc.* **2019**, *68*, 404–417. [CrossRef]
13. Taylor, R.; Boulanger, P.; Torres, D. Real-time music visualization using responsive imagery. In Proceedings of the 8th International Conference on Virtual Reality, Alexandria, VA, USA, 26–29 June 2006; pp. 26–30.
14. Li, Y.; Zhuo, J.; Fan, L.; Wang, Z.; Wang, H.J. Semantically Enriched Music Visualization via Multimodal Color Generation. In Proceedings of the NIME 2021, Shanghai, China, 14–18 June 2021 .
15. Kubelka, O. Interactive music visualization. In Proceedings of the Central European Seminar on Computer Graphics, Budmerice, Slovakia, 1–3 May 2000; Volume 4.
16. Erdmann, M.; von Berg, M.; Steffens, J. Development and evaluation of a mixed reality music visualization for a live performance based on music information retrieval. *Front. Virtual Real.* **2025**, *6*, 1552321. [CrossRef]
17. Reymore, E.C.; Lindsey, D.T. Color and tone color: Audiovisual crossmodal correspondences with musical instrument timbre. *Front. Psychol.* **2025**, *15*, 1520131. [CrossRef] [PubMed]
18. De Kegel, B.; Haahr, M. Procedural puzzle generation: A survey. *IEEE Trans. Games* **2020**, *12*, 21–40. [CrossRef]
19. Stevens, J.C.; Marks, L.E. Cross-modality matching of brightness and loudness. *Proc. Natl. Acad. Sci. USA* **1965**, *54*, 407–411. [CrossRef] [PubMed]
20. Bond, B.; Stevens, S.S. Cross-modality matching of brightness to loudness by 5-year-olds. *Percept. Psychophys.* **1969**, *6*, 337–339. [CrossRef]
21. Marks, L.E. On associations of light and sound: The mediation of brightness, pitch, and loudness. *Am. J. Psychol.* **1974**, *87* , 173–188. [CrossRef] [PubMed]
22. Hiraga, R.; Mizaki, R.; Fujishiro, I. Performance visualization: A new challenge to music through visualization. In Proceedings of the tenth ACM international conference on Multimedia, Juan les Pins, France, 1–6 December 2002; pp. 239–242.
23. Olowe, I.; Grierson, M.; Barthet, M. User requirements for live sound visualization system using multitrack audio. In Proceedings of the 12th International Audio Mostly Conference, London, UK, 23–26 August 2017; pp. 1–8.
24. Taenzer, M.; Wünsche, B.C.; Müller, S. Analysis and visualisation of music. In Proceedings of the 2019 IEEE International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019; pp. 1–6.
25. Bain, M.N. Real Time Music Visualization: A Study in the Visual Extension of Music. Ph.D. Thesis, The Ohio State University, Columbus, OH, USA, 2008.
26. Malandrino, D.; Pirozzi, D.; Zaccagnino, R. Learning the harmonic analysis: Is visualization an effective approach? *Multimed. Tools Appl.* **2019**, *78*, 32967–32998. [CrossRef]
27. Khulusi, R.; Kusnick, J.; Meinecke, C.; Gillmann, C.; Focht, J.; Jänicke, S. A survey on visualizations for musical data. *Comput. Graph. Forum* **2020**, *39*, 461–488. [CrossRef]

28. Han, Y.; Kim, J.; Lee, K. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *25*, 208–221. [CrossRef]
29. Cycling '74. Cycling '74. 2026. Available online: https://cycling74.com/ (accessed on 9 January 2026).
30. Derivative. TouchDesigner. 2026. Available online: https://derivative.ca/ (accessed on 9 February 2026).
31. Synesthesia. Synesthesia.live. 2026. Available online: https://synesthesia.live/ (accessed on 9 February 2026).
32. SYQEL. SYQEL. 2026. Available online: https://syqel.com/ (accessed on 9 February 2026).
33. Games, E. Unreal Engine. 2024. Available online: https://www.unrealengine.com/en-US (accessed on 9 January 2026).
34. De Prisco, R.; Malandrino, D.; Pirozzi, D.; Zaccagnino, G. Understanding the structure of musical compositions: Is visualization an effective approach? *Inf. Vis.* **2017**, *16*, 139–152. [CrossRef]
35. Cycling '74. *Max: A Visual Programming Language for Media.* 2026. Version 8.6. Available online: https://cycling74.com/products/max (accessed on 24 March 2026).
36. Hevner, A.R. A three cycle view of design science research. *Scand. J. Inf. Syst.* **2007**, *19*, 4.
37. Varjo. VR/XR Headset for Training and Simulation. 2026. Available online: https://varjo.com/products/xr-4 (accessed on 20 March 2026).
38. HTC Corporation. *VIVE Pro 2 Full Kit: Specifications and Overview.* 2026. Available online: https://www.vive.com/uk/product/vive-pro2-full-kit/ (accessed on 20 March 2026).
39. Choy, C.; Reich, M. Music Visualizer and Synesthesia Simulation in Virtual Reality. Pomona College. 2016. Available online: https://github.com/celia96/music-visualizer (accessed on 20 March 2026).
40. Edmonds, W.A.; Kennedy, T.D. Convergent-Parallel Approach. In *An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods*, 2nd ed.; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2017; pp. 181–188. [CrossRef]
41. Linares-Vargas, B.G.; Cieza-Mostacero, S.E. Interactive virtual reality environments and emotions: A systematic review. *Virtual Real.* **2024**, *29*, 3. [CrossRef]
42. Galloway, A. Non-Probability Sampling. In *Encyclopedia of Social Measurement*; Kempf-Leonard, K., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; pp. 859–864. [CrossRef]
43. Punchdrunk. Punchdrunk Official Website. 2025. Available online: https://www.punchdrunk.com/ (accessed on 24 March 2026).
44. Games, E. Developing for Head-Mounted Experiences with OpenXR in Unreal Engine. 2022. Available online: https://dev.epicgames.com/documentation/en-us/unreal-engine/developing-for-head-mounted-experiences-with-openxr-in-unreal-engine. (accessed on 9 January 2026).
45. Lenssen, N.; Needell, D. An introduction to fourier analysis with applications to music. *J. Humanist. Math.* **2014**, *4*, 72–91. [CrossRef]
46. Philharmonia Orchestra. *Duke Bluebeard's Castle: An Immersive Digital Experience.* Directed by Esa-Pekka Salonen. 2026. Available online: https://philharmonia.co.uk/products/duke-bluebeards-castle/ (accessed on 24 March 2026).
47. Kester, L.; Kirschner, P.A.; Van Merriënboer, J.J.; Baumer, A. Just-in-time information presentation and the acquisition of complex cognitive skills. *Comput. Hum. Behav.* **2001**, *17*, 373–391. [CrossRef]