Sculpting Unrealities: Using Machine Learning to Control Audiovisual Compositions in Virtual Reality

Bryan Dunphy

Goldsmiths, University of London
Computing
Thesis submitted for the degree of PhD in Arts and Computational Technology
April 2022

Declaration

Declaration of Authorship

I Bryan Dunphy……. (please insert name) hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed: _____ Date: 16/04/2022

Dedication

For Sarah and Liam.

Acknowledgements

I wish to thank my supervisors, Prof. Mick Grierson and Dr. Simon Katan for their invaluable expertise, insight and guidance throughout the PhD. Each supervision session would be a source of inspiration and motivation. Without their guidance I would not have progressed to where I am today. I would like to thank Prof. Atau Tanaka for providing opportunities to present my work to the public. These events were of seminal importance to the development of the project. I would like to thank Dr. Holly Rogers and Dr. Theo Papatheodorou for encouraging my writing and engaging with my work, providing invaluable feedback at a very important time. I would like to thank Prof. William Latham and Prof. Bret Battey for engaging with the work on a deep level and providing invaluable insight. Their expertise has helped enormously. I would also like to thank Pete, Nicky, Tom and everyone in the Computing Department tech team for being so helpful and for creating the amazing VR Labs where I spent so many hours. I would also like to thank my PhD colleagues and everyone in BPB2 who provided friendship and inspiration when I needed it most. I would like to thank my mother, Maria, and father, Billy, for nurturing my love of music and art from the very beginning and supporting me every step of the way. I would like to thank my brother, Ian, for being a constant inspiration, rock of support and for always being there with love and humour to lighten the load. I would also like to thank my son Liam for providing so much happiness and love which got me through some trying times. Thanks also to my amazing friends here in London who have always helped me so much by being there to make me laugh.

Finally, I would like to thank my wife Sarah for everything she has done for me throughout this long journey. She has been there to help me celebrate every high, and more importantly, has carried and supported me through every low. Simply and truthfully, I would not have made it without her.

Abstract

This thesis explores the use of interactive machine learning (IML) techniques to control audiovisual compositions within the emerging medium of virtual reality (VR). Accompanying the text is a portfolio of original compositions and open-source software. These research outputs represent the practical elements of the project that help to shed light on the core research question: how can IML techniques be used to control audiovisual compositions in VR? In order to find some answers to this question, it was broken down into its constituent elements. To situate the research, an exploration of the contemporary field of audiovisual art locates the practice between the areas of visual music and generative AV. This exploration of the field results in a new method of categorising the constituent practices. The practice of audiovisual composition is then explored, focusing on the concept of equality. It is found that, throughout the literature, audiovisual artists aim to treat audio and visual material equally. This is interpreted as a desire for balance between the audio and visual material. This concept is then examined in the context of VR. A feeling of presence is found to be central to this new medium and is identified as an important consideration for the audiovisual composer in addition to the senses of sight and sound. Several new terms are formulated which provide the means by which the compositions within the portfolio are analysed. A control system, based on IML techniques, is developed called the Neural AV Mapper. This is used to develop a compositional methodology through the creation of several studies. The outcomes from these studies are incorporated into two live performance pieces, *Ventriloquy I* and *Ventriloquy II*. These pieces showcase the use of IML techniques to control audiovisual compositions in a live performance context. The lessons learned from these pieces are incorporated into the development of the ImmersAV toolkit. This open-source software toolkit was built specifically to allow for the exploration of the IML control paradigm within VR. The toolkit provides the means by which the immersive audiovisual compositions, *Obj_#3* and *Ag Fás Ar Ais Arís* are created. *Obj_#3* takes the form of an immersive audiovisual sculpture that can be manipulated in real-time by the user. The title of the thesis references the physical act of sculpting audiovisual material. It also refers to the ability of VR to create alternate realities that are not bound to the physics of real-life. This exploration of unrealities emerges as an important aspect of the medium. The final piece in the portfolio, *Ag Fás Ar Ais Arís* takes the knowledge gained from the earlier work and pushes the boundaries to maximise the potential of the medium and the material.

Table of Contents

List of Figures

List of Code Examples

Chapter 1 Introduction

> It could be defined, of course, only with the existence of art works possessing the distinction of self-definition. (Whitney 1980: 33)

This thesis is an investigation into the use of interactive machine learning (IML) and virtual reality (VR) in the composition of audiovisual artworks. The aim of the research is to discover ways in which these emerging technologies can be used together to articulate compositional goals in the practice of audiovisual composition. By doing this, it is hoped that new avenues of creative expression can be identified within the context of audiovisual art which will, in turn, lead to new experiences for audiences. The core research question being addressed throughout the thesis is:

- How can interactive machine learning be used to control audiovisual compositions within the medium of virtual reality?

In order to arrive at the answer to this question, the thesis carefully expands upon three strands of enquiry before combining them in a final, coherent statement. The constituent strands of the core research question are as follows:

- The practice of audiovisual composition.
- The use of IML to control audiovisual material.
- The use of VR as an artistic medium, through which, audiovisual compositions can be presented.

1.1 Audiovisual Composition

The practice of audiovisual composition is a central component of the research undertaken here. To situate the work within a wider artistic context, Chapter 2 takes a look at the field of audiovisual art. The approximate location of the field itself is carefully identified before several constituent practices are discussed. Both shared, and unique, characteristics of each practice are identified through an examination of historical and contemporary work. The multitudinous ways in which audiovisual art is disseminated is then discussed in relation to the individual practices.

Once the wider artistic context is established, Chapter 3 drills down into the act of audiovisual composition itself. The aim of this discussion is to explore the questions; what artistic principles guide the audiovisual composer? Are there any common principles existent in, or suggested by,

contemporary literature and work, which can be identified to guide this practice? In order to answer these questions, some prominent literature is discussed and shown to suggest a recurring sentiment among audiovisual artists relating to a desire for equality between audio and visual material in their compositions.

This desire for equality is examined through the question; in what ways can audio and visual material be considered equal? This conceptualisation of the relationship between audio and visual material is expanded upon and developed into a set of compositional principles that provide the theoretical foundation for the practical work presented later in the thesis. This theoretical work provides the primary grounds for analysis and assessment of the effectiveness of the subsequent compositions. By developing a compositional foundation such as this, it is hoped to generate the means by which the artistic work can be analysed, according to principles specific to audiovisual composition.

Chapter 4 extends the scope of audiovisual composition theory into the emerging medium of VR. The motivation for exploring audiovisual compositions in VR is established before terminology, specific to the medium, is discussed. A significant element of this motivation is the potential for VR to act as 'an unreality simulator' (Slater and Sanchez-Vives 2016: 6). These immersive VR environments are the *unrealities* referenced in the title of the thesis. The chapter also asks the question; what freedoms or constraints would this new medium impose on the practice of audiovisual composition? The concept of presence is identified as a significant element that may affect the perception of the audiovisual material. The balance between representational and abstract elements is then explored as a way to explore the concept of presence within audiovisual compositions. This theoretical work merges with the earlier hypotheses as it asks how the sense of presence may affect the perception of equality between audio and visual material.

The opening quote is from John Whitney Sr. and refers to the idea of 'self-definition'. Although Whitney is talking about avant-garde cinema, this quote resonated with the audiovisual practice presented in this thesis. The analytical terms developed in Chapters 3 and 4 are based on observations of current and historical work in the field. These concepts, grounded in established audiovisual practice, are employed in an effort to lend the later artworks the 'distinction of self-definition' that will, in turn, contribute towards the continually evolving self-definition of the practice of audiovisual composition.

1.2 Interactive Machine Learning

The use of IML techniques is the central paradigm for control of audiovisual material within this thesis. Chapter 5 establishes the concept of parameter mapping as a significant technical element of audiovisual composition. It then asks whether an IML control paradigm would be suitable for controlling audiovisual compositions that are influenced by the concepts established in Chapters 3 and 4. The advantages of using an IML approach to parameter mapping is established. The use of a feed-forward multilayer-perceptron neural network is proposed to quickly map input control-data to a range of audio and visual parameters. These mappings are non-linear and represent a novel way of controlling audiovisual compositions in real-time. A software system for controlling audiovisual parameters using neural networks is presented. Four studies, composed using this system, are then presented, that aim to develop compositional and performance techniques informed by the use of this technology. The studies are analysed using the terminology established in Chapter 3 and several compositional outcomes are identified.

The outcomes from the studies are implemented in a suite of live performance pieces; *Ventriloquy I* and *II*. These pieces are presented in Chapter 6. The chapter opens with a discussion around live audiovisual performance, highlighting desirable characteristics of live performance systems. The knowledge gained from this discussion is then implemented in *Ventriloquy I*, where the IML control paradigm is extended from a 2D, to 3D, input parameter space. The development of the material and the compositional methodology, specifically informed by the IML control paradigm, is discussed, followed by an analysis of the piece through the lens of the theoretical concepts established in Chapter 3. *Ventriloquy II* is then presented, which utilises a shared immersive space as its performance context. This anticipates the extension of the practice into the fully immersive medium of VR. The development of the material, compositional methodology, method of control and an analysis of the performance is presented.

1.3 Virtual Reality

The final strand of the core research question is the implementation of the above work within the emerging artistic medium of VR. Chapter 7 presents an open-source toolkit, *ImmersAV*, that was built to facilitate the creation of immersive audiovisual compositions, harnessing the power of IML, within VR. The chapter opens with a survey of a range of contemporary tools used to create audio and visual material for VR. These tools are discussed in relation to the technical requirements needed to extend and explore the previous artistic work within a VR environment. Following this discussion, the

motivation for creating the toolkit is established. The constituent elements of the system are then described followed by a discussion on its functionality and structural design.

The *ImmersAV* toolkit is used to create two original immersive audiovisual compositions; *Obj_#3* and *Ag Fás Ar Ais Arís*. Chapter 8 describes the development of *Obj_#3*. This is the first realisation of the core research goal in its entirety; the real-time control, using IML, of audiovisual compositions within VR. The development of the audio and visual material is discussed in detail with a strong emphasis on creating a sense of presence. The piece is analysed according to the analytical terminology established in Chapters 3 and 4. Audience feedback is gathered following a public demonstration and used to inform the outcomes that, in turn, influence the development of the final piece in the portfolio.

Chapter 9 describes the development and composition of *Ag Fás Ar Ais Arís*. This piece takes the knowledge gained from the previous works, and builds on it, to maximise the expressive potential of this approach, to controlling material within the medium of VR. Central to this is the question; how can an immersive audiovisual composition move beyond a rigidly defined duality of environmental-versus-foreground material? What techniques can be employed to result in the blurring of boundaries between these elements? The development of the audio and visual material is discussed with this goal in mind. The compositional methodology is discussed, and an analysis of the piece is provided using the terminology established earlier in the thesis. In this way, *Ag Fás Ar Ais Arís* ties together the three strands of inquiry and presents a unified statement in answer to the core research question. Finally, Chapter 10 reflects on the outcomes of the research before suggesting some future avenues of exploration.

1.4 Original Contributions

The work presented in this thesis has led to several original contributions to the field of audiovisual art. These include:

- A new categorisation scheme for audiovisual art practices.
- New terms for analysis and conceptualisation of audiovisual relationships.
- Development of a software toolkit to facilitate the creation of IML-controlled audiovisual compositions within VR.
- A portfolio of original artwork demonstrating the dramatic and aesthetic potential of the IML control paradigm in live performance and within VR.

The field of audiovisual art contains a multitude of diverse practices, drawing influence from related fields of sonic and visual creative expression. In order to navigate this space, a new categorisation scheme is presented in Chapter 2 that highlights the difference between aesthetic practices and presentation contexts. In doing so, this new method of categorisation emphasises that it is through the combination of these elements, that the diverse flavours of audiovisual art arise.

Chapter 3 interprets the concept of equality between audio and visual material in terms of a perception of balance. The discussion develops this concept of balance under the term *audiovisual balance*. Constituent elements that contribute to the perception of audiovisual balance are identified that include *relative-temporal-motion*, *relative-expressive-range* and *isolated-structural-incoherence*. These new terms are used throughout the text to analyse the artwork in the portfolio.

The open-source software toolkit, *ImmersAV,* is an original contribution to the field of computational audiovisual composition. It is a minimal, code-based environment, intended to focus the audiovisual composer on the process of audiovisual composition by establishing clear working contexts where data can be mapped easily from any part of the environment to another. It is specifically tailored to emphasise the IML control paradigm and also to display the generated material in VR. It provides an alternative environment to commercial game engines for computational audiovisual composers to develop their work.

The original portfolio of compositions represents a contribution to the field, demonstrating the expressive potential of the technologies and compositional methodologies pursued through the research project. The titles of these compositions are shown in Fig. 1.1.

Figure 1.1 Titles of original artworks.

1.5 Conclusion

This chapter established the core research question that will be explored throughout the thesis. The strands of enquiry related to the core question were identified and an overview of the original contributions to knowledge was given. The next chapter will provide a survey of historical and contemporary audiovisual work, within which, this research sits.

Chapter 2 Audiovisual Art

This chapter aims to locate the practice-based elements of the thesis within a wider artistic context. To begin, the term *audiovisual art* will be discussed. As will be seen below, discussing it can be complicated due to inconsistent definitions and unclear terminology. This is followed by a review of work, identifying the *centres of gravity* around which contemporary practice is emerging. Several presentation methods will be discussed that demonstrate the possibility for each audiovisual practice to be presented within a range of contexts. The combination of practice and context provides the multitude of ways in which audiovisual art is manifested.

The term *centres of gravity* conceptualises the practices in such a way that there are no hard borders between them. Their boundaries are porous. This aspect of audiovisual art means that whilst a piece might exist as a live cinema work, it can contain characteristics also found in generative audiovisual work. The different practices all have their own identities. However, they are all connected through these shared characteristics. Further, they can also share presentation contexts. This makes it difficult to identify where some practices end, and some begin. There is a space of overlap where work can exist between practices and could easily be discussed in terms of more than one discipline. Hence, it is appropriate to conceptualise the various practices as centres of gravity, wherein some work sits directly, whilst other work can be situated in-between. This approach guards against the conceptualisation of each practice as a self-contained, isolated entity and encourages the acceptance of a group of practices tied together under the umbrella of audiovisual art. See Fig. 2.1, which shows audiovisual practices in circles connected with lines of various types to some general characteristics that can be observed within those practices. The different types of lines are intended to aid the reader in following their path from the characteristic to the practice. It is clear that many of the practices share characteristics while also retaining their own identities. Identifying these practices provides a frame of reference for emerging practitioners approaching the field. This also provides clarity for those already working in the field which is essential to the effective communication of new practice and research. The intention here is not to isolate the practices or indeed audiovisual art as a whole from its many influences. Instead, it is to create discursive signposts that can be used to trace the development of ideas within a continually evolving area of creativity. Ultimately, the aim here is to

develop an understanding of the contemporary field within which the artistic work in this thesis can be located.

2.1 What do you mean ‘audiovisual art’?

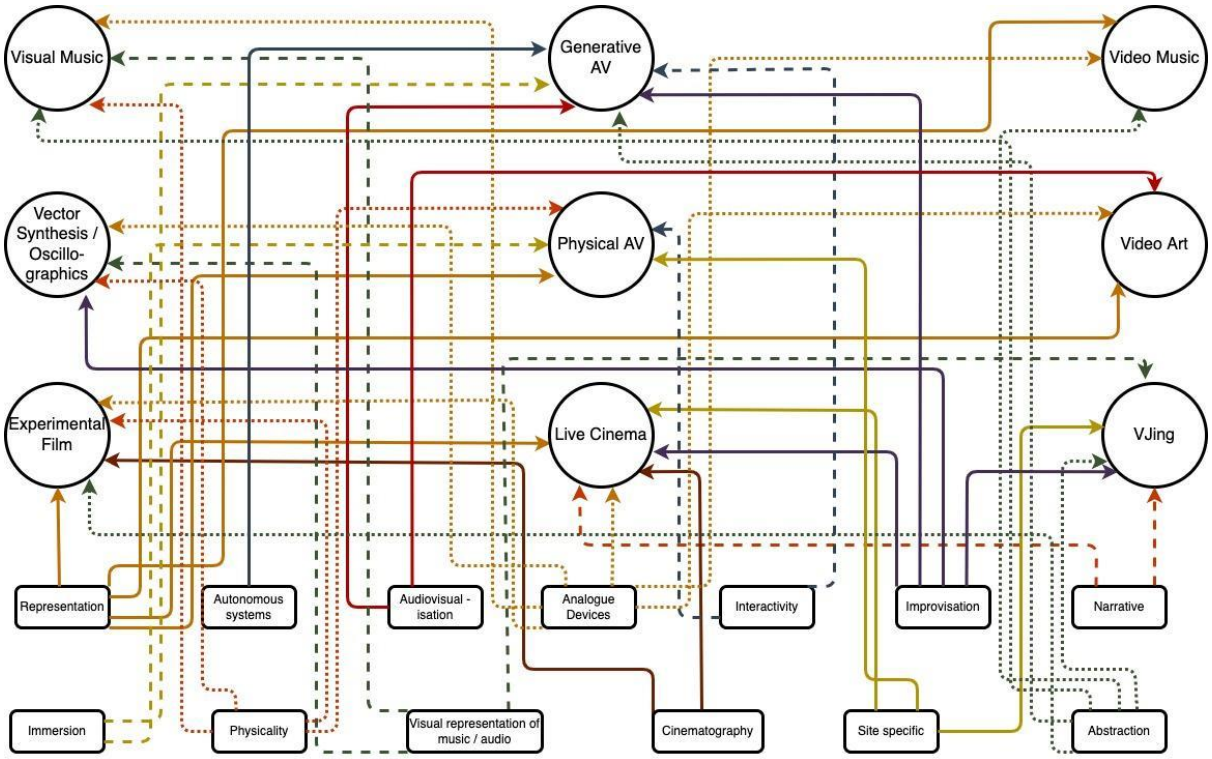

Figure 2.1 Audiovisual art practices (circles) and related characteristics (rounded rectangles). The lines joining the characteristics to the practices are rendered in different styles to allow the reader to follow their path.

Audiovisual art occupies an ambiguous space within the arts. This ambiguity is perhaps due to the fact that there are many types of audiovisual art, and the various practices each draw from elements of neighbouring artforms, making it difficult to discuss the field on its own terms. There are certainly many benefits to being so inclusive in terms of aesthetic influences. This imparts a vibrancy and sense of vastness to the field which is exciting. The problem with this tendency towards generality is that by being everywhere, it disappears. In *The Oxford Handbook of New Audiovisual Aesthetics*, Richardson, Gorbman and Vernallis (2013: 20) argue for a holistic approach to the study and analysis of audiovisual media. They present a collection of essays by artists, academics and commentators on current forms of practice and activity. The book offers a wide perspective on the type of multi-modal engagement currently present in many mass forms of media. Here, the authors identify audiovisual traits within already established fields. They state that the book ‘calls for a reassessment of the centre and boundaries of the audiovisual’ (Richardson, Gorbman and Vernallis 2013: 3), yet there is no mention of a distinct *audiovisual* field in and of itself and no attempt at defining such an area. It appears that there are two conceptualisations in the literature regarding audiovisual work. On the one hand there is the approach taken by Richardson, Gorbman and Vernallis amongst others. They talk

about the *audio-visual* as a ubiquitous manifestation of any human expression that engages the eyes and ears. In this conceptualisation, the effect of the combined sound and imagery is sometimes a by-product rather than the main aesthetic focus. In contrast, this thesis proceeds with the premise that there is a distinct field called *audiovisual art*, which encompasses a range of practices and styles, where the core intent of the work is to explore the interaction between sound and image. As will be seen below, many of these practices blur the lines between long established fields, including music, film, animation and computational art.

In *The Audiovisual Breakthrough*, Carvalho et al. (2015) acknowledge the current difficulty encountered by the public, by curators and by the artists themselves in discussing audiovisual art due to the disparity of terms and concepts in use. Here, they call it a 'confusing web of unclear, or even inconsistent, definitions' (Carvalho et al. 2015: 5). Due to this, they are attempting to clarify and find coherent ways of talking about the contemporary field which will enable greater access to the public and act as a base from which commentators and researchers can 'elaborate on philosophical, aesthetic, and theoretical implications related to contemporary practices' (ibid. 2015: 7). Although they claim to be working towards 'developing useful definitions for both the theoretical debate and the performance context' (ibid. 2015: 5-6), during the course of the book, the authors often fall short of committing to any concise explanation of the individual areas and tend to qualify each conversation at the end with the implication that these forms of practice defy description (ibid. 2015: 35, 102, 134). This seems to contradict the very purpose of the book and also highlights how difficult it is to achieve their aims. The presented essays and the survey that acts as their foundation set a precedent for the work being done in this chapter. It shows there is a need within the audiovisual community for a well-defined vocabulary despite the fact that some practitioners seem ambivalent towards its purpose.

> We might of course say – especially as performers – that we "really don't care" and that we are "more interested in doing than explaining," as one of the participants in our survey put it. (Ibid. 2015: 6)

This statement appears to disregard the benefits of being able to discuss artistic practice; one of which is helping people to understand where the work is coming from. Carvalho et al. include any practice that engages the senses of sight and sound in their conception of audiovisual art.

> Generally speaking, audiovisual works range across media such as TV, cinema and live shows to include all the possibilities that present a stimulus to both auditory and visual sensorial systems. (Ibid. 2015: 11)

Logically following from this they also state that practices such as 'puppetry and theatre can be audiovisual' (ibid. 2015: 11). This is a similar sentiment to Richardson, Gorbman and Vernallis (2013) who discuss audiovisual tendencies within already established fields of practice. They are in fact talking about the *audio-visual*. In doing this, they are avoiding the need to firmly define an audiovisual art field. In order to address this, the term *audiovisual,* without a hyphen, will be used here to describe practices included within the field of audiovisual art. Whilst practices such as puppetry or theatre technically engage the senses of both sight and sound, the interaction between audio and visual material is not the core focus of the work. These practices can be described as *audio-visual* in that they are seen and heard by the audience. However, they are not *audiovisual*. In terms of the vocabulary presented here, to describe a practice as audiovisual, is to explicitly state that the creator, or creators, of the work engaged in the composition of sounds and images such that the focus of the work is on the interaction *between* the two modalities. The focus here is the combination of sound and image itself and what emerges from it. Carvalho et al. (2015) undermine the stated purpose of their work by failing to identify audiovisual art as a distinct field. In their conception, because everything can be audiovisual, the term audiovisual art ceases to hold any meaning. Also, on further inspection of their definition of the term, *audiovisual,* they seem to contradict themselves. They state that anything that engages both the senses of sight and sound can be audiovisual, yet they go on to say that audiovisuality 'describes a generic group of practices' (Carvalho et al. 2015: 11). Considering any practice that is both seen and heard at the point of reception as audiovisual art is to describe nearly every known performative practice and also a multitude of fixed format practices that are too numerous to mention here. This can hardly describe a generic group of practices.

As evidenced above, there is ambiguity as to whether audiovisual art even constitutes a distinct artistic field. This thesis works on the premise that there is a distinct field, or 'metadiscipline' (Grierson 2005) that is called audiovisual art and the work produced here arises through the practice of audiovisual composition. This concept of *metadisciplinarity* is also present in Holly Roger's description of the early position of video art which she says, 'operated like a "meta-media", a multi-incorporative genre' (Rogers 2013: 39). This concept of metadisciplinarity is useful in the conceptualisation of how the different areas of audiovisual practice are influenced by related disciplines. Each of these areas takes what elements it needs from its related disciplines and moulds them into a new entity with its own independent identity. In the following sections the central characteristics of each practice are discussed, whilst acknowledging overlapping boundaries through analysis of the relevant literature and contemporary work.

2.2 Practice + Context

The rest of the chapter is divided into two sections. Firstly, identifiable practices within audiovisual art will be discussed. Some examples of work will be given and their relationship to other practices will be explored. Following this, several presentation contexts will be discussed. A distinction is made between artistic practice and presentation context. Each time an audiovisual work is created it is a manifestation of the particular practice *plus* the presentation context. Fig. 2.2 shows this combination with a double-headed arrow to indicate the reciprocal influence each element has on the other. Audiovisual art is often highly technological. This is part of why it can be very exciting. However, there is a danger of conflating artistic concerns with technological concerns. Rogers (2013: 79) points to Bill Viola as an example of an artist conscious of this danger.

> Viola has often complained that the term "video artist" defined him by the materials he used, and not the uses to which he put them. Rather, he considers his materials to be of secondary importance to his work, the means by which he expresses himself rather than the expression itself.

To describe an artistic practice in terms of the technology it uses is not very helpful when trying to probe the essence of that practice. However, it is also important to be mindful of the effect that technology and presentation context have on the art. Practicalities and limitations of the technologies used will inevitably affect the artwork. The conceptualisation of the audiovisual field, in terms of aesthetic practice and presentation context is intended to highlight how they can be combined and interchanged to create individual flavours of audiovisual art.

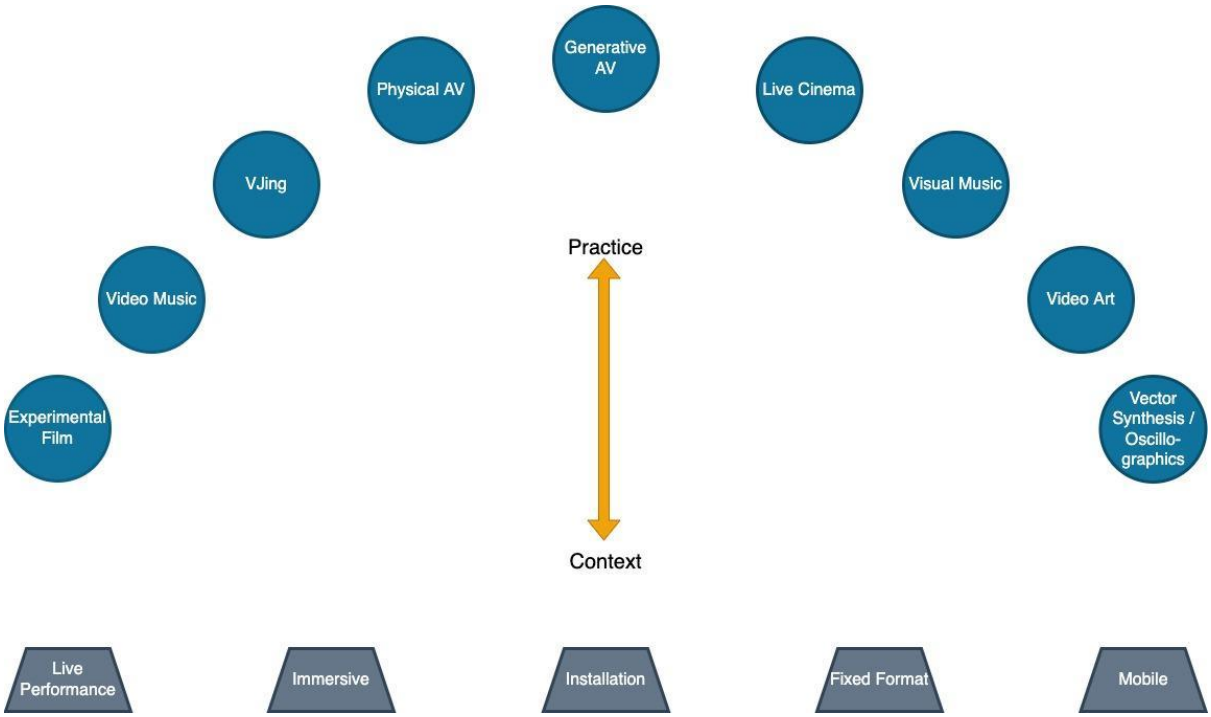

Figure 2.2 Practice plus context in audiovisual art.

2.2.1 Aesthetic Practices

This section will describe some of the practices that make up the audiovisual art landscape. It is not intended as an exhaustive account of all audiovisual practice, rather an overview of some of the work being done and an attempt to highlight interconnecting relationships.

Visual Music

Adriano Abbado (2017: 7) recognises the difficulty in strictly defining the practice when he says that there 'are different notions of what constitutes visual music'. This is a difficulty found throughout the audiovisual art field. The painter and art critic Roger Fry coined the term 'visual music' in his 1912 preface to the catalogue of the *Second Post-Impressionist Exhibition* at the Grafton Galleries in London (reprinted in Fry 1920: 156). The following year, in his 1913 essay *The Allied Artists* (reprinted in Fry 1996: 150), Fry used the term to describe three paintings by Wassily Kandinsky. It seems, from the 1912 text, that he was originally describing the general post-impressionist development of abstraction, using later works by Picasso as examples (reprinted in Fry 1920: 157). With this in mind, it was the writings of William Moritz in the 1970s and 1980s that 'documented the start of a history for the area of visual music' (McDonnell 2010) thereby establishing the term as a

descriptor for work stretching not only back to the absolute films of Viking Eggeling, Hans Richter, Walter Ruttmann and Oskar Fischinger, but as far back as the 18th century colour organs of Father Bertrand Castel, whom he describes as a pioneer.

> Pioneers of Visual Music (like Father Castel) struggled with unwieldy mechanisms…..that could produce only marginally satisfactory visual imagery. (Moritz 1986)

Guldemond, Bloemheuvel and Keefer (2012: 10) identify Fischinger as a more recent pioneer of the form stating that he 'paved the way for an art form that came to be known as Visual Music'. From the literature cited above, we can see the establishment of a long history of visual music activity. Fig. 2.3 shows some influential areas of practice that are related to some visual music work.

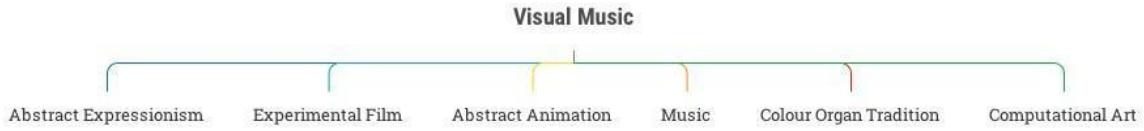

Figure 2.3Visual music with related fields and influential practices.

According to Moritz, the artistic aim of the practice is 'to create with moving lights a music for the eye comparable to the effects of sound for the ear' (Moritz 1986). This sentiment forms a central part of many contemporary definitions. With regard to his own differential dynamics, a method of visually representing harmonic ratios through motion, John Whitney (1980: 35) states that this 'new grammar must speak to us as eloquently as music or fail its very reason for existence'. Ox and Keefer (2008) state that a visual music piece can be a 'visualization of music which is the translation of a specific musical composition (or sound) into a visual language, with the original syntax being emulated in the new visual rendition'. They also state that a visual music piece can take the form of a 'time-based narrative visual structure that is similar to the structure of a kind or style of music' (Ox and Keefer 2008). McDonnell (2014) also stresses the structural role of music in these works when she states:

> A visual music piece uses a visual art medium in a way that is more analogous to that of musical composition or performance. Visual elements (via craft, artistic intention, mechanical means or software) are composed and presented with aesthetic strategies and procedures similar to those employed in the composing or performance of music.

Following from this, it is apparent that the structural basis of a visual music piece is fundamentally musical, presented in a visual form that 'can have sound, or exist silently' (Ox and Keefer 2008). *Rhythmus 21* (1921) by Hans Richter and *Symphonie Diagonale* (1921 - 1924) by Viking Eggeling are examples of silent films generally included in the visual music canon. In contrast to the silent presentation of those examples, Mary Ellen Bute's *Synchromy No.4: Escape* (1937) is presented with Bach's *Toccata and Fugue in D Minor* (BVW 565). Here the visuals are an expression of the musical composition. John and James Whitney take an alternative approach with their *Five Abstract Film Exercises* (1943 - 1944). Here, they use music composition techniques to arrange original audio and visual elements, rather than setting the visual material to pre-existing music. Regarding the fifth film of the series, they state that it opens 'with a short canonical statement of a theme upon which the entire film is constructed' (Whitney 1980: 150).

Visual music works don't even require motion in some cases. They can take the form of a 'visual composition that is not done in a linear, time-based manner, but rather something more static, like a 7' x 8' canvas' (Ox and Keefer 2008). This form of visual music is represented by such work as *Improvisation 35* (1914) by Wassily Kandinsky. Going by this definition, later 20th century graphic scores such as Cornelius Cardew's *Treatise* (1967) could be seen as visual music. Simon Katan (2012: 15) has noted this characteristic, stating that although he regards *Treatise* as a 'failure in terms of an improvisatory notation', he judges it 'successful as a pictorial representation of music'. Here we see visual music acting as a metadiscipline, using elements of painting and music theory to create something new. Ox and Keefer (2008) define a further form that visual music can take:

> A direct translation of image to sound or music, as images photographed, drawn or scratched onto a film's soundtrack are directly converted to sound when the film is projected. Often these films are simultaneously shown visually. Literally, what you see is also what you hear.

This form of visual music is represented by works such as Oskar Fischinger's *Ornament Sounds (1932)*, Norman McLaren's *Synchromy (1971)* and Lis Rhodes' *Dresden Dynamo* (1972). These works reveal a connecting thread between audiovisual practices. Rhodes' work in this vein creates a connection with the field of experimental film which will be discussed below. The film strip here is a common medium from which these artists were able to represent sound and image simultaneously. This ability to connect sound and image through a single medium was further refined with the development of video technology. Woody Vasulka's *No. 25* (1976) 'audibly and visually scans the raster field so the image can both be seen and heard' simultaneously (Rogers 2013: 33). Whitney's (1980) abstraction of periodicity and Pythagorean ratios from Western tonal harmony also acts as a

base from which sound and visuals can be simultaneously arranged. Just as Aimee Mollaghan (2015: 170) identifies Whitney as a 'transitional figure' between analogue and digital visual music, his visualisation of differential dynamics could also be seen as a precursor to the artistic audiovisual representation of dynamical systems and datasets that is a characteristic of contemporary generative AV work such as Ryo Ikeshiro's *Construction in Kneading* (2013) and Ryoichi Kurokawa's *unfold.alt* (2016).

Whilst the above pieces consist of abstract visual forms, representational visual material can also be arranged according to musical structure. Bridging the gap between experimental film and visual music are the *city symphony* films. These include Walter Ruttmann's *Berlin, Symphony of a Great City* (1927) and Dziga Vertov's *Man with a Movie Camera* (1929). These films were originally presented without sound, with the editing and footage arranged according to musical structures:

> The films of the genre offered symphonies in black and white, movies that sought to translate the power of music into the visual medium and thus appeal to the whole of humanity. (Levin 2018: 226)

This process of editing film according to musical techniques also featured in Sergei Eisenstein's works such as *The Battleship Potemkin* (1925) and *October* (1927). Here, Eisenstein sees rhythm as a fundamental part of montage. He further classifies types of montage as metric, rhythmic, tonal and overtonal (Eisenstein 2010: 228). The use of musical editing techniques can also be seen in Christian Marclay's *Video Quartet* (2002). This piece uses short musical clips from Hollywood films to create a 'fourteen-minute musical symphony - one with its own distinct rhythms and sections, including moments of calmness and dramatic counterpoints' (Martin 2014).

Ox and Keefer (2008) note that due to an increased interest in the field, the 'boundaries and definitions' are being tested 'by many who consider that any correlation of sound and image is Visual Music'. This can be seen throughout Lund and Lund (2009), which presents a collection of essays from within and without academia that deal with audiovisual practices such as live cinema, VJing, experimental film and early visual music films from Walter Ruttmann and Mary Ellen Bute. Their concept of visual music here appears similar to the concept of the ubiquitous *audio-visual* discussed in section 2.1 above. In their introduction to the book, they acknowledge the lack of a specialised vocabulary with which audiovisual practitioners can discuss and promote their work. They explicitly draw a link between 20th century visual music pioneers and contemporary work to contextualise

modern audiovisual practices (Lund and Lund 2009: 11). However, they tend to use visual music as a universal term for all contemporary and historical audiovisual practice rather than a participant in a wider audiovisual environment as is the understanding in this thesis. They highlight the diverse nature of the work contained within the book and acknowledge that some of the works cited 'directly contradict each other' (ibid.).

An example of this is Friedmann Dähn's essay *Visual Music – Forms and Possibilities*, where he states that today, it is 'presumed that proof of a universally accepted synesthesia of color and tone does not exist' (ibid. 2009: 149). This is a view that is nearly universally accepted within the visual music community. Yet Henry Keazor's essay, entitled *Visual Music in Mark Romanek/Coldplay, Speed of Sound* (ibid. 2009: 104 - 112), makes a link between the use of an LED light-wall in Mark Romanek's video and 'synesthetic colors and shapes' in order to contextualise it within a visual music heritage (ibid. 2009: 110). He remarks on the similarity between Romanek's choices of patterns and those produced by synesthetes in 'the late 1920s' (ibid. 2009: 108). It is unclear whether he is suggesting Romanek is himself a synesthete or if he had merely 'drawn some inspiration' (ibid. 2009: 110) from the published artwork. On one hand, Dahn is saying that synesthesia of tone and colour should not be counted on as a universal compositional guide due to its overtly subjective nature. On the other hand, Keazor is drawing parallels between Romanek's music video and the work of synesthetes. This type of contradiction on the nature of audiovisual forms only serves to confuse the field. The inclusion of Keazor's essay is in line with the Lunds' assertion that they see visual music tendencies in a diverse range of media forms, in this case, the music video. This is an interesting demonstration of the overlap between practices and disciplines. However, as discussed above regarding the *audio-visual*, perhaps it would be helpful to make a distinction between the ubiquitous proliferation of audio-visual media and specific audiovisual art practices such as visual music.

Some audiovisual practitioners avoid using the term visual music to describe their own work. For instance, Nuno Correia (2013: 48) states that while aspects of Ox and Keefer's definition of visual music are relevant to his work, he prefers 'the use of the related term "audiovisual" instead of "visual music"'. He gives several reasons for this rejection. Firstly, he associates the term 'visual music' with 'works from early to mid-twentieth century' (Correia 2013: 48). He also points out that his use of the term 'audiovisual' creates a link between his work and the work of contemporary influences such as Golan Levin who uses similar terminology, and Michel Chion's *Audio-Vision* (ibid. 2013: 48). Below, Correia's work will be discussed under the generative AV heading. However, some of his work such as *AVOL* (2008) and *AV Clash* (2010) could also be described as visual music. Correia himself

compares these pieces to the work of John Whitney, stating that the 'animations in AVOL and AV Clash resemble John Whitney's floral compositions' (ibid. 2013: 44). Here we see a contemporary audiovisual artist working in the overlapping space between generative AV and visual music.

Correia's stance on the historical connotations of visual music has some support as Lund (in Carvalho et al. 2015: 20) states that 'the term assumes two different functions: on the one hand, it is referred to as an ancestor that has engendered other, more recent audiovisual expressions, while, on the other hand, visual music is very much alive as a contemporary audiovisual expression in its own right'. This description of visual music, as an ancestor related to contemporary practices, is supported by the survey data in Carvalho et al. (2015: 70). In that survey, only 1.9% of the participants describe their work as visual music. Considering the survey data, it could be argued that visual music does tend to hold more weight as a historical term that calls to mind the colour organ tradition or the films of Fischinger, Whitney, McLaren and Bute rather than contemporary work. The reason for such a low percentage identifying as visual music practitioners could be related to the lack of terminological clarity surrounding the practice and the audiovisual art field as a whole.

Whilst this historical aspect of visual music has become a prominent characteristic, there are contemporary works being made. Events such as *Punto y Raya*,[1] *Seeing Sound*[2] and *Sound/Image*[3] are providing platforms for artists to present new visual music work such as Vibeke Sorensen's *Mayur* (2015) and *Duel Tones* (2016) by Maura and Bébhinn McDonnell. In the programme for the *Seeing Sound 2016* conference, *Duel Tones* is described as a 'fixed media work that explores through a visual music collaborative effort, the emergence of synthetic tones and timbres and synthetics forms and motion elements from a ground of blackness and silence' (Seeing Sound 2016). The audiovisual relationships between the sound and the image are quite tightly synchronised. As long vertical strips suddenly appear, synthesised tones are emitted which create momentary cross-modal bonds. Some of the visual textures dissolve in synchronisation with granular sonic textures creating further audiovisual bonds. The tight synchronisation of the sound to the visuals fools the audience into thinking that there is a cause-and-effect relationship between the two modalities. However, the visuals were created before adding the audio. In this way, the artists used the visuals to create 'the mood and structure of the music', thereby 'utilizing the visuals as a type of evolving, synthetic graphic score'

[1] https://www.puntoyrayafestival.com/ (accessed 07/09/2020).

[2] http://www.seeingsound.co.uk/ (accessed 07/09/2020).

[3] http://www.gre.ac.uk/ach/events/soundimage (accessed 07/09/2020).

(ibid.). *Duel Tones* is firmly based in the visual music tradition. In terms of its technical production, a direct line can be drawn between this piece and *Lichtspiel Opus I* (1921) by Walter Ruttmann. Ruttmann created the visuals using 'oil paints on glass plates beneath an animation camera, shooting a frame after each brush stroke' (Moritz 1997b). Ruttmann considered music essential to the experience of this film so he 'commissioned the composer Max Butting to compose a string quartet for it' (Valcke 2008: 173). Here we can see that the visuals were created first in both instances followed by a tightly synchronised score. This ordering of production has possible implications in the hierarchy of the media. The visuals dictate the form and structure of the audio. However, the temporal arrangement of the visuals themselves are inspired by musical form and movement. So there seems to be a cyclical process of transference between modalities.

As evidenced above, visual music is a substantial area within audiovisual art that has developed in many different directions. The boundaries of this practice are not easy to find and there are multiple theoretical definitions that appear contradictory. The above discussion attempted to clarify where the centre of the practice lies, taking into consideration these contradictory definitions and arguing for more accurate discourse surrounding the discipline.

Generative AV

The development of dedicated GPUs and faster CPUs has meant that a lot of the work done today tends to be computationally generated. McCormack and Dorin (2001: 3) define generative art as 'art that uses some form of generative process in its realization'. This definition is inclusive of generative forms of practice that don't necessarily involve a computer. Bret Battey (2016: 172, note 1) acknowledges that the 'term "generative art" is subject to numerous definitions'. He then specifically contextualises his own usage as 'another term for art (visual, music or other) that involves an artist coding and manipulating algorithms as part of his or her process'.

Generative AV is a term that is being used among a number of contemporary practitioners such as Nuno Correia (2015), Ryo Ikeshiro (2013) and Tom Betts (2013) to locate their practice. Generative AV works are related to the fields of computational art, algorithmic composition, audio synthesis, creative computing, visual music and procedural graphics (see Fig. 2.4). Here we see the audiovisual *metadiscipline* in action again, taking elements of related fields and creating something new. Whilst generative AV can be distinguished from visual music, the two areas of practice are intimately related.

Elements of generative AV practices can also be seen in earlier video art, or 'video art-music' (Rogers 2013: 7). Many practitioners transitioned to using digital media as the technology evolved, thereby continuing to explore the 'opportunities for synthesis' (ibid. 2013: 3) first made possible with magnetic videotape technology. *Violin Power* (1970 – 1978) by Steina Vasulka demonstrates a precursor to later generative AV work. In these live performances, Vasulka used the sound of her violin to affect video images in real-time. She extended this piece in 1991, utilising 'a MIDI instrument – the five stringed, electric ZETA Violin – and a PowerBook to interface her sounds' (ibid. 2013: 29-30).

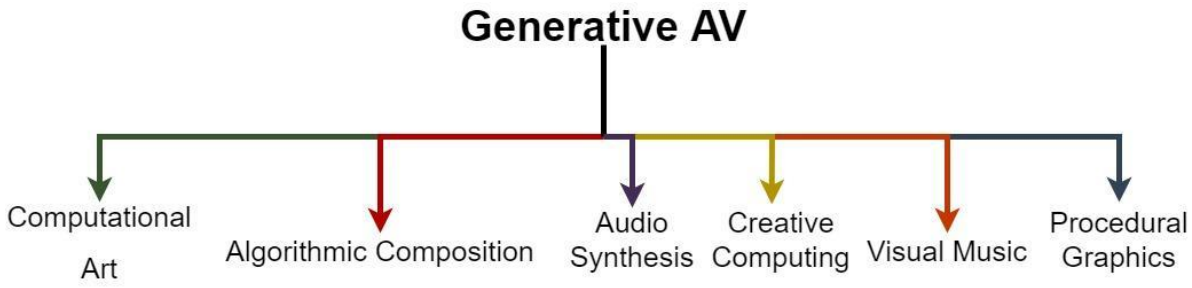

Figure 2.4 Generative AV and related fields.

Sometimes, a generative AV piece will generate both the audio and visual material simultaneously from an underlying source of data. Such an approach can be seen in Ryo Ikeshiro's *Construction in Zhuangzi* (2011). Here, Ikeshiro employs what he calls 'audiovisualisation' to generate the audio and visuals for the piece. He describes it as 'simultaneous sonification and visualisation' (Ikeshiro 2013: 58) where 'all audio and visuals are derived from the same system; hence, they are audiovisualisations of the same data source, a category within a wider range of current audiovisual practice' (ibid. 2013: 58). Here, Ikeshiro derives his material from 'a modified Lorenz dynamical system' (ibid. 2013: 83). The artistry here is in deciding how to represent the data in each modality so as to reveal the evolution of the underlying system. In doing so, Ikeshiro is showing the audience that there is a fundamental connection between the audio and the visuals.

> Thus in addition to metaphorical interpretations between the two media which are of course perfectly possible and valid, the system responsible for structuring the audiovisuals can be conveyed. (Ibid. 2013: 59)

This focus on structure and form can also be seen in *Cube with Magic Ribbons* (2012) by Simon Katan. Here, Katan develops the audio and visual elements sequentially, building up rhythmic and tonal sequences both visually and aurally. He utilises 2D computer graphics and synthesised sound.

The composition was written using Katan's custom *SoundCircuit* environment that 'allows multiple virtual tape-heads to travel through a two-dimensional wrapped space along tracks that can be freely inter-connected' (Katan 2012: 129). This is a live performance piece where the performer creates sequences of shapes, or 'blips' (ibid. 2012: 131) over which the 'tape-head', visualised as a box, passes. This causes the blip to deform and also emit sound. During the course of the piece, Katan creates multiple tracks and blips moving along their own trajectories. The result is a complex build-up of sonic textures and rhythms coupled with intricate visual movement.

Katan here is interested in preserving the audience's sense of structure and form, stating that he is 'exploring an intuition that formal relations between streams of events such as, [sic] symmetries, transpositions, retrogrades and inversions are considerably easier to parse visually than aurally' (ibid. 2012: 129). As discussed above, this motivation is similar to some of the core aims of visual music, namely the translation of formal structures across modalities. As the piece develops and the visual activity on the screen becomes more and more complex, the temporal nature of sound versus the spatial nature of sight becomes apparent. As multiple boxes travel around the environment activating the blips, it becomes impossible to visually take everything in whereas the audio evolves in the audiovisual space as a continuous stream. This causes the audience to let their eyes travel across the environment, resting here and there on a box as it traverses a blip and experiencing the resulting audiovisual interaction.

The blips in *Cube with Magic Ribbons* could be conceptualised as unified audiovisual objects that are triggered when the boxes pass over them. The exploration and development of virtual audiovisual objects is a core part of Nuno Correia's (2013) work, especially with *Gen.AV* (2015). Regarding his *Interactive AudioVisual Objects (IAVO)* approach, Correia explains that 'synchronization between audio and visuals is ensured via an algorithm that manipulates the visuals based on the analysis of the audio' (Correia 2013: 23). He further states that the choice of visualisations aims to be 'subjective interpretations of sounds or categories of sounds, where the image adds new meaning to the sound and vice-versa' (ibid.). This new meaning Correia speaks of is added-value, which will be discussed in the next chapter.

As outlined above, generative AV often involves the generation of audio and visual material using computer code and computational algorithms. It is a practice that is continually evolving, with strong roots in computational art. It can also be seen as a very close relation of visual music. Here we see an

example of the fluid boundaries between audiovisual practices. In many cases, both practices utilise abstract visuals. If representational visual material is used, it is generally used in an abstract manner. An example of a piece that utilises representational visuals and also straddles the practises of visual music and generative AV is Francesc Martí's *Speech 2* (2015). The base material for this piece is made up of short video clips from a US TV program. The audio from the clips is processed using granular synthesis techniques and tightly synchronised with the video, creating a highly engaging, and sometimes humorous audiovisual expression. The humour here arises from the repetitive visual motion and the short aural soundbites. The viewer sees and hears glitchy representations of talking heads but is often unable to catch what they are saying, thereby creating tension in the experience of the piece. These examples demonstrate that although many visual music and generative AV pieces utilise abstract visual material, representational material can be handled in an abstract way.

The practices of generative AV and visual music also share some compositional approaches, such as the simultaneous creation of sound and visuals from a common medium. This can be seen in both the generative AV work of Ryo Ikeshiro (*Construction in Zhuangzi* 2012*)* and Mick Grierson (*Light Speak* 2005*),* and the visual music work of Oskar Fischinger (*Ornament Sounds* 1932*)* and Norman McLaren (*Synchromy* 1971*)*. They also share many common pioneers such as John Whitney, Vibeke Sorenson and Larry Cuba. However, as is stated in many definitions of visual music, an emphasis has been historically placed on the visual arrangement of elements according to musical characteristics such as form and structure. The aim is to give visually centred art the abstract ethereality of music, which has led to an emphasis on the visual presentation of works over the musical. Consequently, this aspect of visual music has allowed for purely visual works that do not contain any audio. Instead, they contain some characteristics of audio or music translated into a visual form. This is not the case in generative AV pieces, which inherently contain both visuals and audio. Visual characteristics can be represented in the audio realm through sonification and audification techniques, just as audio characteristics can be represented in the visual realm through computational analysis of the audio signal. Some scholars include these characteristics in their definitions of visual music (Ox and Keefer, 2008). However, this may contribute to Lund's (Carvalho et al. 2015: 20) assertion that visual music 'has acquired an extremely broad meaning, to the point of becoming potentially meaningless'.

The VJ or 'video jockey' is usually seen as the counterpart of the club DJ (Carvalho et al. 2015: 106). The VJ scene has deep roots in club culture and emerged alongside electronic music and DJ mixing. As with the other areas within audiovisual art, VJing is influenced by several related areas (Fig. 2.5).

Figure 2.5 VJ practice and related fields.

For a detailed survey of the foundation of the scene and the development of VJ culture up to 2006 by artists such as The Light Surgeons, Coldcut, D-Fuse, Pfadfinderei, Raven Kwok, Yoshi Sodeoka and the label AntiVJ, see D-Fuse (2006). Here, Adrian Shaughnessy not only links the VJ to its original meaning on MTV in the early 1980s, but also traces the foundations of VJ culture through the 1960s in the USA and the emergence of light shows put on by Brotherhood of Light, The Light Sound Dimension and the Joshua Light Show. In the UK, similar activity was emerging through the work of Barry Miles at the UFO club in London (D-Fuse 2006: 10-11). Bram Crevits notes that although VJing has an aesthetic debt to pay to Dada, Fluxus, Expanded Cinema and the multimediality of Andy Warhol, it is culturally inextricable from the emergence of house and techno music (ibid. 2006: 14-15). Eva Fischer defines VJing in the following way:

> VJing as artistic practice stands for video mixing, visual jamming, or visual live coding, and defines itself via the act of selecting and intuitive jamming live as well as the processing of visual contents[sic] and real-time settings. (Carvalho et al. 2015: 106)

This definition of VJing emphasises the live and improvisational aspect of the practice. This is a central characteristic found at the heart of the practice. Another fundamental characteristic of VJing is the fact that they are always accompanying someone else, usually a DJ. They are always interpreting someone else's music in a visual way.

> VJing as an action addresses the visual side which, however, always occurs in combination with another level. Without implying that the visual part is secondary or less worthy, a VJ always visualises something else. (ibid. 2015: 112)

The imagery used can be representative or abstract. The representational content is often sampled from found footage or existing movies in much the same way as a DJ will take samples from existing recordings of songs. In this way 'VJing has developed into a visual format which defies traditional forms of narration' (ibid. 2015: 106-108).

Live Cinema

Live Cinema is an area of audiovisual art that seems to be closely related to experimental film and VJing. Although these are the main two areas of influence, it is also related to the fields of expanded cinema, electronica, sonic arts and documentary film (Fig. 2.6). As we saw above with visual music, there are problems of terminological clarity and boundary identification relating to this practice.

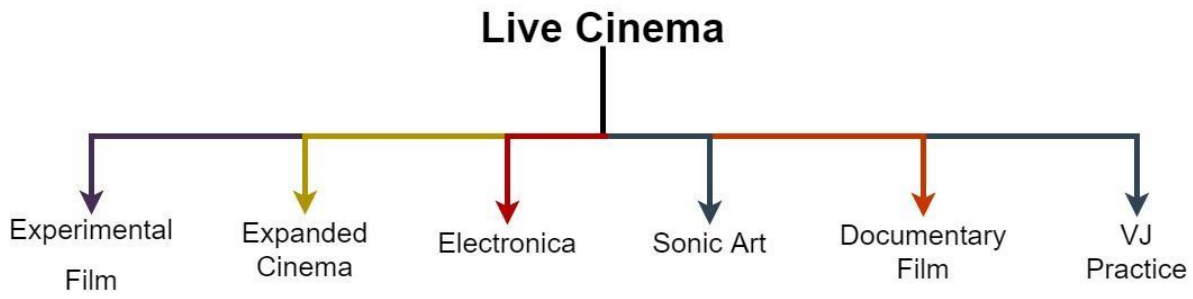

Figure 2.6 Live cinema and related fields

According to Gabriel Menotti, 'the practice encompasses forms of audio-visual performance that actively engage with traditional cinematographic conventions' (Carvalho et al. 2015: 81). These conventions can include the exploration of elements such as the presentation context and narrative. Mia Makela (2008: 1) notes that some live cinema performances are presented in a 'museum or theatre, and often to an audience similar to that of the cinema: sitting down and watching the performance attentively'. Amy Alexander (2009) also identifies this presentation context as a characteristic of live cinema, stating that 'the audience is typically seated and focused on the performance'.

Some practitioners from the experimental film tradition are bringing the techniques and ideas they developed in their earlier work across to live cinema. An example of this can be seen in some of Peter Greenaway's later work such as the *Tulse Luper Suitcases* (2003 - 2007), in which Greenaway presents the history of the 20th century through a loose narrative based on a central character called

Tulse Luper. Greenaway created a multitude of outputs for this project, including an online game, a full-length DVD feature film and also some live performances. Greenaway's work here demonstrates a very strong link between live cinema and experimental film as he transposes many structuralist characteristics from his earlier films such as *Vertical Features Remake* (1978) and *The Falls* (1980). Recent performances of Malcolm LeGrice's *After Leonardo* (1974), which consists of improvised live projection coupled with live improvised sound (Thomas 2017), can also be considered live cinema, but is most definitely rooted in the experimental films of the 60s and 70s (what some people call expanded cinema, especially in relation to LeGrice).

According to Menotti, some live cinema practice arose as a response to the marginalisation of the VJ in the club scene. He explains that the attraction of live cinema for VJs is that the practice places the VJ in control of all the aesthetic aspects of a performance as opposed to playing a reactionary role to the DJ (Carvalho et al. 2015: 85-91). Practitioners such as Toby Harris, aka *spark, have moved from VJing to live cinema and see the construction of narratives and cultural critique as central to their practice (ibid. 2015: 85). Alexander (2009) also allows for loose narrative structures in live cinema stating that 'performers often develop loose visual narratives over the course of the performance'.

These characteristics can be seen in *True Fictions* (2007) by The Light Surgeons, who originally made their name as VJs. *True Fictions* is a live performance combining interviews, found footage, motion graphics and music composed specifically for the project. The elements are manipulated live, creating a collage of non-linear narratives that explore 'the themes of truth and myth through a multitude of American and Native-American voices' (The Light Surgeons 2007). The visuals throughout the different sections are mainly figurative, sometimes showing passing scenery and blending it with video loops of musicians playing instruments tightly synchronised to the audio. This creates moments of synchresis that act as points of contact for the audience where they can bind the visuals and audio perceptually. Added to this are the three performers on stage reinforcing the liveness of the event.

Sometimes live footage of their actions on stage are also projected to show cause and effect between the performer's movements and the audio. Watching this performance, the influence of VJing and club culture is very much apparent. The structure of the music is firmly rooted in the traditions of dance music with a regular 4/4 beat and characteristic rhythmic drops. The Light Surgeons blend these familiar tropes with interviews, political speeches and old stock footage to impart an underlying

narrative related to American history. Here we again see audiovisual art acting as a metadiscipline, taking elements of documentary and club music to create something new, in this case live cinema.

Menotti (Carvalho et al. 2015: 91) claims that when former VJs call their performances live cinema, they are purposefully referencing the established history of the cinematic tradition and using that association to give their practice more cultural weight.

> Thus, live cinema is not simply a definition, but a proposition: a statement that certain works are not merely part of a technological fad - even when they might be - but exemplars of a late avant-garde.

He goes on to state that these VJs are carving out 'an exclusive segment in the wider territory of VJing' (ibid. 2015: 93). This seems to imply that all live cinema artists are also VJs. This is indicative of the problem of terminological clarity prevalent throughout the field as some live cinema artists may not identify with the VJ tradition.

Makela (2008: 1) sets the two practices firmly apart, stating that the goals of the live cinema practitioner 'appear to be more personal and artistic than those of VJs'. Here she is perhaps referring to the fact that VJs predominantly work in commercial club contexts. Menotti (Carvalho et al. 2015: 89) seems to agree with this sentiment saying that the latter represents a more commercial, industry focused practice whilst 'live cinema creators would occupy a place equivalent to that of *film auteurs'.* His agreement here with Makela makes his above statement about live cinema existing within the VJ scene even more confusing. Menotti seems to be using the term VJ to denote live audiovisual performance in general. A more useful conceptualisation, as presented in this thesis, utilises the term audiovisual art as the umbrella, under which areas such as VJing and live cinema can exist independently in some respects but related in others (Fig. 2.7). Here the green and yellow arrows point to the shared characteristics of live cinema and VJing whilst the characteristics specific to each practice are listed to their right and left sides respectively. Audiovisual art is in the centre showing that they are both part of a larger field.

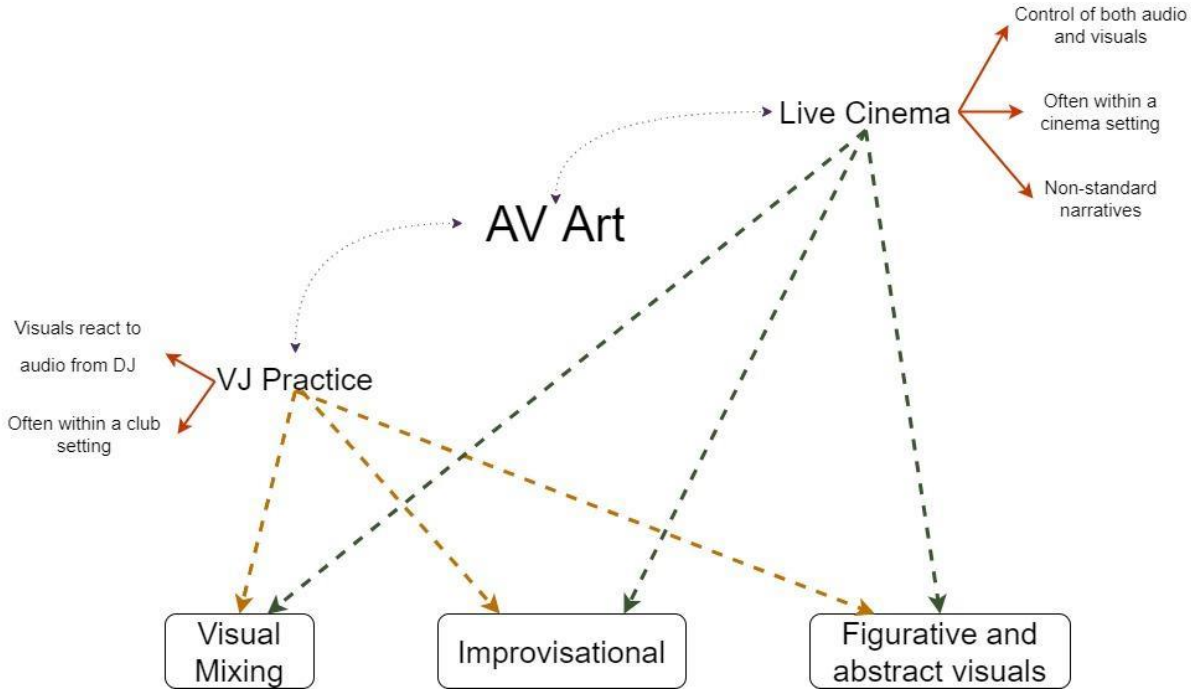

Figure 2.7 VJing and live cinema, related and distinct characteristics.

Makela (2008) quotes the program of the 2005 transmediale festival in Berlin stating that a defining characteristic of live cinema is the 'simultaneous creation of sound and image in real-time'. This definition more accurately describes generative AV and, in a more peripheral way, visual music. She further states that 'the term "Cinema" is now to be understood as embracing all forms of configuring moving images, beginning with the animation of painted or synthetic images' (Makela 2008). This statement vastly over-stretches the boundaries of what can be considered cinema. The term *cinema* has such a strong tradition and specific history attached to it, that it instantaneously sets up expectations in the mind of the public. As we saw above, this is the very reason some VJs have adopted the live cinema terminology. Trying to divorce the practice from this history would be nearly impossible. In saying this, Makela is diluting the term to such an extent that it becomes meaningless. As seen earlier with visual music, there seems to be a tendency within audiovisual art for practitioners in one area to try to claim all forms of audiovisual work for themselves.

Makela is attempting to establish a robust description of the practice of live cinema. She refers to the language of live cinema as a 'spoken language' that is lacking 'written grammar' (ibid. 2008: 1-2). This desire to solidify an audiovisual grammar, or vocabulary, is valid and can be seen as a proactive attempt to make discursive progress within the field. However, there seems to be an oversight of neighbouring practices as she further draws comparisons to visual music by proposing that 'the

principle of music composition could be helpful in constructing structure for non-figurative visuals' (ibid. 2008: 3). She also states that beyond the meaning of the imagery, musical characteristics such as 'rhythm, dynamics, movement, direction, speed, color, intensity and richness are important in a live performance' (ibid. 2008: 3). This description is quite close to the central characteristics of visual music. It could be argued that Makela is operating in the blurred space between the two practices. There is, of course, nothing wrong with this. However, an acknowledgement of the neighbouring practices would help to orientate the work within the wider artistic context.

In conclusion, it appears that live cinema, having emerged in some ways from VJ practice, is situated in close proximity to experimental film, utilising narrative and structural tropes in a creative and more abstracted way than commercial cinema. Here, the inclusion of the term *cinema* is fundamental in allying itself to that tradition. Makela's assertion that even abstract, synthetic moving images with no narrative can be described as live cinema, could also be interpreted as a blurring of the lines between live cinema, visual music and generative AV.

Experimental Film

Experimental film is a well-established art form stretching back to the beginnings of cinema. There is much crossover between this practice and visual music. Malcolm Le Grice (1982: 19) notes that the first group working in abstract film in Germany included Walter Ruttmann, Viking Eggeling, Hans Richter and Oskar Fischinger. This group of artists are also considered visual music pioneers. Whereas visual music films were concerned with transposing musical qualities into the movement of abstract visual forms, experimental films such as *Ballet Méchanique* (1924) by Fernand Léger began to introduce the camera itself as material in the film (Le Grice 1977: 38). The physicality of camera mechanics and the materiality of the film stock itself became central themes of the work of filmmakers such as Malcolm Le Grice (*Little Dog For Roger* 1967), Owen Land (*Film in Which There Appears Edge Lettering, Sprocket Holes, Dirt Particles Etc.* 1965) and Wilhelm and Birgit Hein (*Rohfilm* 1968). David Rimmer's *Surfacing on the Thames* (1970) re-films the screen to transform the image and to draw attention to the act of filming and its relationship to the act of projection (ibid. 1982: 136).

Some filmmakers were also concerned with perception. Tony Conrad's *The Flicker* (1966) consists of alternating black and white frames intended to produce an intense perceptual experience for the

audience. The goal of other filmmakers was to 'counteract the emotional manipulation and reactionary catharsis of popular cinematic form by the development of conscious, conceptual and reflexive modes of perception' (ibid. 1982: 152 - 153). The view of mainstream cinema as an emotionally manipulative medium stretches back at least to the early Soviet filmmaker Dziga Vertov, who argued that the financiers of commercial dramatic cinema were closing people's minds to the realities of life and thereby keeping them subdued in a dream world (ibid. 1982: 55).

The emergence of digital tools in the second half of the 20th century allowed experimental filmmakers to move into new media. However, during the 1990s there was a resurgence in Austria of artists working with film as a 'critique of the digital revolution' (Rees 2011: 134). Martin Arnold and Peter Tscherkassky used found footage and home-made devices to create films that revisited the loops and materialist focus of earlier experimental films (ibid.). A.L. Rees notes that 'the decade from 2000 - 10 saw a flowering of experimental and expanded film and projection art' (ibid. 2011: 139). He further elaborates on expanded cinema by tracing the 'promiscuous merging of technologies' (ibid. 2011: 140) back to the work of 'Stan Vanderbeek, Carolee Schneemann and others in the 1960s and 1970s who mixed film, video, live events, slides and sound in their multi-media practice' (ibid.).

Experimental film is wide ranging, steeped in history and is continually reinventing itself through combinations of various visual and sonic technologies. It has an influence on many of the audiovisual practices described in this chapter. It could also be perceived as existing outside of audiovisual art in some respects. The interaction between sounds and images is often not the main concern of experimental filmmakers. Instead, issues such as the materiality of film, socio-political critique of experimental versus commercial forms of media and critique of new media are sometimes explored by artists. This seems to highlight a core characteristic of audiovisual art, the perceptual interaction of purposefully composed sonic and visual material regardless of medium, where the *interaction* is the focus of the work. For me this is the essence of audiovisual art regardless of the method of creation or dissemination. Many experimental filmmakers can be described in this way. Lis Rhodes' 'generation of direct sound from the film strip in *Light Music*' (ibid. 2011: 142) situates that work in the same visual music arena as Fischinger and McLaren as noted above. However, Lis Rhodes would generally be considered an experimental filmmaker rather than a visual music artist. This highlights the fluidity of audiovisual boundaries yet again.

Video *art-music* is a term used by Holly Rogers (2013) which accentuates the importance of sound in the development of early video art practice. By highlighting this she designates video artists as 'artist-composers' (Rogers 2013: 9). She argues that video is more closely related to sound than film or photography in its technical evolution. As an electromagnetic, rather than photosensitive, medium, its heritage 'is that of audiotape, rather than film technology' (ibid. 2013: 18).

Some of the earliest artistic practice with video is generally attributed to Nam June Paik and his 'experiments in New York with a new portable video camera in 1965' (Rees 2011: 96). Paik's work in the 60s can be seen as a reaction against commercial TV and the use of the TV as an art-object and creative medium in itself. Gene Youngblood identifies four simultaneous directions within Paik's work of this period. They include 'synaesthetic videotapes; videotronic distortions of the received signal; closed-circuit teledynamic environments; and sculptural pieces, usually of a satirical nature' (Youngblood 1970: 302). Rogers argues that throughout 'this early phase, then, video culture was in the process of becoming, acting more like an adhesive that pulled together hitherto disparate strands of art and music practices than as a protagonist in its own story; a facilitator of intermedial discourse rather than a genre' (Rogers 2013: 39). This could also be descriptive of the field of audiovisual art, which is a group of practices, not necessarily connected by use of a particular medium like video, but connected by the shared goal of investigating the interaction between sonic and visual material.

Paik's video work in the 60s inspired artists in the UK to begin working in the new medium (Rees 2011: 96). The influential UFO club run by 'video activist John Hopkins' (ibid.*)* began at this time. Hopkins later went on to form a community-based post-production centre with Sue Hall called the Fantasy Factory. This tendency towards community art was a characteristic of one of three branches of video work in the 70s. The remaining branches included those who called themselves 'video artists' concentrating on the 'conditions of video as a mode of perception and production', and those who were engaged in 'making "artists" video' as a 'rejection of traditional media rather than as an unexplored primary medium' (ibid. 2011: 97). Rees views this split within video culture as a possible reason for the 'absence of a developed theory of video - in contrast to film' (ibid. 2011: 98). He notes that the only group that were interested in formulating a cohesive theory was the 'video artists', whilst the other two groups rejected any form of theoretical approach. The results of this situation, according to Rees, 'continue to hold back critical debate and analysis of video and its digital descendants' (ibid.*)*.

Rogers states that 'the most innovative aspects of video come from the opportunities for audiovisual synthesis that it enabled' (Rogers 2013: 14). Regarding non-musical sound in video, she argues that 'if we consider them as sounds intentionally collected and meant to be "heard", then many video pieces come close to the aesthetics found in all three types of expanded music - noise music, sound art and sound by artists. Understood in this liberated context, the audio part of the video can be seen as an expansion of musical material into visuality: "video noise" can then be read as a form of audiovisual composition' (ibid. 2013: 38). This is a strong argument for the inherent audiovisuality of video work.

Video Music[4]

Video music is a practice strongly associated with artists in Montreal, Canada. Jean Piché, one of the leading practitioners of the discipline, has influenced and taught several contemporary artists such as Frieda Abtan, Myriam Boucher and Maxime Corbeil Perron. A central technique of the practice is the use of heavily processed video footage. The audio is influenced by the electroacoustic approaches of composers such as Dennis Smalley and Michel Chion. Chion's theories on the relationship between sound and image are also important. The practice is closely related to 'visual music, video art and experimental cinema', starting from the 'studio practices of electroacoustic music in the late 1980s' (Boucher and Piché 2020: 13). Boucher (2020) takes footage of natural events and processes them digitally to the point of semi-abstraction. Phenomenology and embodied experience are central philosophies of video music composers.

> Rather than approaching nature as a landscape, I am inspired by the physical experience of being present and immersed in the natural world, which consequently impacts on how I see and hear. I endeavour to perceive the world surrounding me in an active, participatory way and for this embodied knowledge to inform my work. (Boucher 2020: 226)

Video music practitioners such as Piché were attracted to video because it is an 'almost identical time-based technology' (Boucher and Piché 2020: 13) to the audio tape. This echoes the argument presented by Rogers (2013) that video as a technology is more closely related to the audio tape than film. The fusion of audio and visuals through the central medium of the tape is important in providing the technical cohesion on which more complicated perceptual and metaphorical correspondences can be composed (Boucher and Piché 2020: 14).

[4] This practice is also referred to as *Vidéomusique* in French. The English translation will be used here.

This approach to creating audiovisual art involves simultaneously sonifying and visualising electronic signals. The signal can be sonified using analog synthesisers and visualised using oscilloscopes, cathode ray tube (CRT) monitors or in the case of Andrew Duff (2014) and Derek Holzer (2017), the *Vectrex* video game console.

Holzer's practice investigates vector synthesis from a media archaeology perspective. He is interested in using obsolete technology that has outlived its intended purpose. By working in this way, he is recontextualising these machines as dedicated artistic tools regardless of their original functions. In particular, he sees the CRT monitor as 'a commercially buried format resurrected by willful misuse and creative experimentation' (Holzer 2017). Holzer uses his bespoke *Vector Synthesis Library*[5], *Pure Data*, a camera and various analogue audio effects units in live performances (ibid.).

Ted Davis has written a library for Processing called *XYscope*[6] that converts graphics to audio so that they can be rendered on a vector display such as an oscilloscope. Andrew Duff works with Vectrex consoles and a Eurorack modular system in his live performances.[7] Douglas Nunn (2018) notes that a vector synthesis approach to audiovisual art allows for several types of cross-modal interactions including independent accompaniment of visuals to audio, simultaneous generation of audio and visuals, and audio-reactive visuals.

There are several connections between the practice of vector synthesis and the wider audiovisual field. The simultaneous generation of audio and visual material from a central data source is a methodological practice similar to audiovisualisation in generative AV work. Further, Holzer's *Vector Synthesis Library* employs scan processing, enabling him to work with 3D models and images. This process is 'based on the same process as the Rutt-Etra video synthesizer from the 1970s' (Holzer 2017). This is the same video synthesiser used by Steina and Woody Vasulka in their video practice. These historical and methodological links bring a certain amount of clarification to the interconnected practices of audiovisual art and help to illuminate the close bonds many of the practices share.

[5] https://github.com/macumbista/vectorsynthesis (accessed 24/09/2020).

[6] https://teddavis.org/xyscope/ (accessed 24/09/2020).

[7] https://youtu.be/hc7eiFDeVNs (accessed 24/09/2020).

Physical AV is a term that will be used here to describe audiovisual works that incorporate some form of physicality into the creative process. To the best of the author's knowledge, the term Physical AV is not used anywhere else in the literature. Pieces of this type usually take place in artistic spaces such as galleries and are often site-specific. They often combine audio with physical objects instead of screen-based objects. If screens are used, they are usually employed in a non-standard, multi-screen format. Projection mapping techniques are sometimes used to project imagery onto objects or structural features of buildings. Physical immersion in the space is also an important factor in some of these works. Another feature sometimes encountered is the development of analogue electronics or mechanical devices as significant elements within the piece. There is also often a focus on sculptural artefacts. The artists working in this context can also draw on characteristics of the wider areas of installation art and sonic art (see Fig. 2.8).

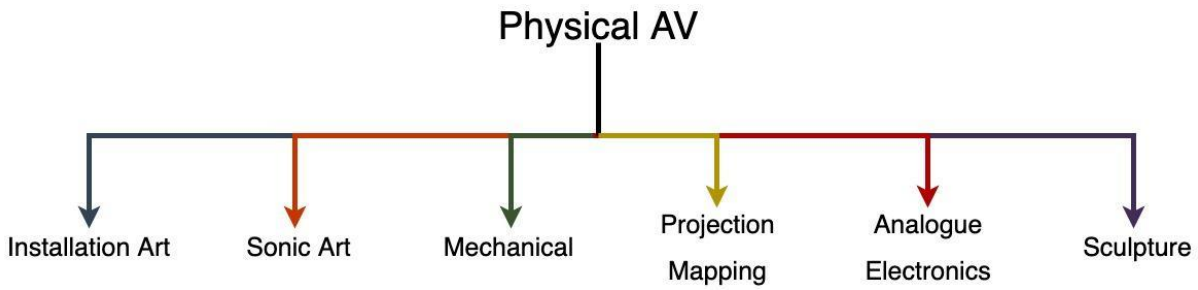

Figure 2.8 Physical AV and characteristics.

The use of physical objects is a central characteristic in the work of Trope (MacGillivray et al. 2013), the collective name of Carol McGillivray (animation) and Bruno Mathez (sound). Their work utilises audio and visuals but removes the visual elements from the computer or projection screen and brings them into the physical realm. They create installations where the visuals are made up of physical objects that are placed within a blacked-out space. Imagery is then projected onto the objects, resulting in the simultaneous triggering of sound and visual stimuli. This combination of physical objects, triggered sound and projection mapping creates audiovisual installations that exist in the physical world, or physical AV. They pull their techniques and methodology together under the term *Diasynchronoscope*, stating that it is:

> a prototypical, experimental medium that draws on tropes from animation and Gestalt grouping principles to create the perception of apparent motion using concrete objects. In short, it is a way of animating without a screen. (Ibid. 2013: 367)

The nature of the installations mean that the audio and visual elements are very tightly linked. As an object is illuminated it is synchronised with sound. For example, in their piece *One, Two, Three...* (2013), the illumination of each cube is accompanied by a vocal sample counting from one to three and back again depending on the direction of the animation. There is a minimalist aesthetic to this piece both in the visuals and the audio. The counting immediately calls to mind Philip Glass and his opera *Einstein on the Beach* (1975). The visuals are simple geometric cubes cascading down and splashing through puddles with a monochrome colour scheme. In other works, such as *Dandelion* (2014) and *Codex* (2014) there are elements of interactivity where the audience can trigger sounds by pressing on seashells (*Codex*) or blowing through a handheld device triggering visual stimuli (*Dandelion*).

The above works utilise projection mapping and tightly synchronised audio to create the illusion of motion and sound using inanimate sculptural objects. Another approach is to utilise objects that emit light and sound themselves. Artificiel's *condemned_bulbes*[8] (2003) is presented as an installation of 25, 36 or 49 suspended 1000W incandescent light bulbs. The intensity of the luminance of the bulbs is controlled using a bespoke generative system. This system modulates the electrical current that is being sent to the bulbs through a specially designed dimmer-switch. This allows the artists to make the filaments in the bulbs resonate at different frequencies, causing audible sound to disperse into the space. In this way they are creating dynamic audiovisual textures that resonate through the exhibition space. The installation can also be used as a performance tool with the artists controlling the bulbs in real-time.

Brian McKenna's *Continuously Variable Colour Field*[9] (2017) straddles the worlds of video art and physical AV. Here, McKenna converts five channels of audio into five channels of video by way of custom electronics. The audio consists of a composition for synthesiser and is presented in a surround-sound format whilst the video signals are routed to five CRT monitors. He maps the frequency of the audio to specific colours that appear on the screens of the monitors. Low frequencies are mapped to red, mid-range frequencies are mapped to green and high frequencies are mapped to blue. The piece is presented as an installation in a dark space, with the monitors placed on the floor and the speakers around the perimeter of the space. A significant element of this piece is the development of the bespoke analogue electronic circuits that convert the audio signals to video. This analogue

[8] https://www.artificiel.org/projet/bulbes (accessed 04/02/2022).
[9] https://vimeo.com/250821574 (accessed 03/02/2022).

physicality, a feature of McKenna's work, offers an alternative to the dominant digitalisation of audiovisual practices.

The physicality of nature can also be harnessed to create audiovisual expressions. *LITTORAL*[10] (2021), by Myriam Boucher and Kathy Hinde, takes scientific data related to the effects of climate change on coastal regions and uses it as material for an audiovisual performance incorporating sculptural objects, light and sound. The sculptural objects take the form of shallow transparent bowl-like objects that are stacked vertically. Each bowl contains water with a white light shining upwards through the liquid. The undulating shadows cast by the light are projected onto canvases placed over the objects. This results in continuously morphing abstract forms that appear and disappear according to the audio-reactive light sources. Natural phenomena are a feature of Hinde's work. The use of water as audiovisual material is also seen in her piece *Tipping Point*[11] (2014). This piece is presented as both an installation and a performance. The sculptural objects take the form of glass containers filled with water. They are arranged in pairs, with outlets at the bottom connected by tubes that allow the water to flow between the vessels. A mechanical device alternately raises and lowers each glass, causing one glass to fill whilst the other empties. The visual effect of this is that the water level between the glasses stays the same with the glasses moving around the water. There are microphones attached to each glass that amplify its resonance, with the pitch changing depending on the amount of water in the glass. There are also audio-reactive lights attached to the bottom of each glass that provide dynamic illumination that is tightly synchronised with the audio. The piece is a commentary on the importance of utilising the world's water resources in a balanced way. Further, the visual forms of the glass containers resemble hospital drips. This further accentuates the precarious position of the climate due to misuse of natural resources.

Drawing parallels to the static visual music work of Kandinsky and Cardew mentioned above, Rob Mullender's *Said Object*[12] (2010) is a video recording, documenting responses to his sculpture entitled *Daughter's Voice From Memory.* This sculpture is a physical manifestation of sound in a static 3D object. The artist recorded his daughter's voice and then made a lathe-cutting of the waveform using cloth and polyester resin. The piece is an investigation into the possibility of conveying properties of the recording through a physical incarnation of the sound. In this way the work is similar to static visual music pieces in that they too are attempting to visually communicate

[10] https://www.cryptic.org.uk/portfolio/littoral/ (accessed 03/02/2022).
[11] https://kathyhinde.co.uk/tipping-point/ (accessed 03/02/2022).
[12] https://vimeo.com/99818964 (accessed 04/02/2022).

structural and emotional characteristics of music. Although the object takes the form of a literal translation of the audio waveform, the diversity of responses by the public demonstrates that even a one-to-one translation such as this, is capable of eliciting many different feelings and imagery in the minds of the audience.

The works described above demonstrate a desire to move away from screen-based media and explore approaches to creating audiovisual work utilising physical objects that are sometimes sculpted by the artist and sometimes repurposed from everyday use. At times, the physicality of analogue electronics and mechanical devices play a central role in the dissemination of the pieces. The physicality of nature can also be used as inspiration for some artists.

2.2.2 Presentation Contexts

This section describes some of the contexts within which the above practices can be presented. As above, this is not intended as an exhaustive account of all modes of presentation. The tendency for many of the above practices to share presentation contexts will be highlighted.

Installation

Audiovisual installations can be presented within the gallery or in a public space. They are sometimes site specific and can also aim to immerse the audience. Ryoji Ikeda uses non-standard projection and multi-channel audio techniques to physically immerse the audience in his installations. His *datamatics* (2008) project sonifies and visualises data from 'hard drive errors and studies of software code' (Ikeda 2008). His aesthetic is stripped back and minimal, consisting of black and white visuals, dotted here and there with flashes of primary colour and sparse electronic sonic textures; at times extremely pointillistic and at times dense and noisy. One incarnation of this project, called *data.tron[8K enhanced version]* (2009), utilised eight projectors and a 9.2 channel sound system to physically immerse the audience in visuals and sound. The visuals were projected onto the floor and walls of a room measuring W16m x H9m x D9m. The speakers were also arranged throughout the room. In creating this immersive space, Ikeda is creating an optimal environment within which the audience's senses of sight and sound can be completely dominated by the artwork. In projecting not only on the walls, but also on the floor, Ikeda is using as much of the space as he can to create a more immersive audiovisual environment. Ikeda would be considered a generative AV artist and presents his work in various contexts including installations and live performance.

Jane Cassidy's work, *They Upped Their Game After The Oranges* (2012), is an audiovisual artwork for stereo sound and projection-mapped visuals. The visuals are projected into the upper corner of a room, utilising the physical characteristics of the space in which it is shown. The visuals consist of sharp animations of squares and triangles that grow in size alongside sustained electronic audio tones. The longer the tone is sustained, the larger the square or triangle grows. When the tone is abruptly cut off, the shape recedes leaving the outline of its final size. The corner of the room is illuminated to give the impression of a 3D box. This creates a hypnotising visual illusion, giving the undulating shapes a 3D sensibility. Aside from the previously mentioned sustained tones, the audio is sometimes glitchy, synchronised with a slight flicker in the visuals. These are the most apparent and direct audiovisual mappings, with the rest of the audio continuing in an ambient textural manner. The abstract visuals are reminiscent of John Whitney's *Matrix III* (1972). On her website, Cassidy describes this piece as a 'mapped visual music piece for a corner' (Cassidy 2012). This is an example of visual music being presented in an installation context.

Immersive

Virtual reality (VR) is a rapidly growing area of practice within which there is a lot of potential to create fully immersive audiovisual environments. Whilst most of the high-profile activity taking place here is related to gaming, audiovisual artists are also creating work using VR technologies. Ox and Britton (2000) describe their virtual immersive experience as a '21st century virtual color organ' (Ox and Britton 2000: 1), revealing a link to the visual music tradition. More recently, *Morphogenesis* (2016) by Can Buyukberber and Yagmur Uyanik is an example of a generative AV piece that uses VR technologies to completely immerse the participant in a world of impressive visuals and unfolding electronic audio textures. *Mutator VR* (2016) by William Latham, Lance Putnam and Sam Devlin places the participant in a psychedelic world populated by interactive creatures. These pieces are generative AV works presented in an immersive context. According to Lombard and Ditton (1997), immersive technologies such as VR provide a 'mediated experience that seems very much like it is not mediated; a mediated experience that creates for the user a strong sense of presence' (Lombard and Ditton 1997). *Mutator VR* and the significance of presence within virtual environments will be further discussed in Chapter 4.

Jon Weinel's *Cyberdream VR* (2019) is a short experience that 'aims to realize a VJ-experience of 1990s rave music and vaporwave in VR' (Weinel 2019: 278). Weinel explores the symbolism of

‘techno-utopias and dystopias’ (ibid. 2019: 281) that was prevalent in 80s and 90s digital technology and electronic music. He suggests that deep connections between the music and visuals can be made through ‘symbolic representation of the imaginative worlds suggested by music’ (ibid. 2019: 280) rather than through audio-reactive analysis. This idea of different levels of mapping complexity is central to the practice of audiovisual composition . *Cyberdream VR* is an example of VJ practice being presented in an immersive context. The core practice of interpreting music visually is taken from the traditional club environment and re-contextualised in VR.

Live Performance

Ana Carvalho (Carvalho et al. 2015: 124) identifies live audiovisual performance as a distinct area within audiovisual art. She describes it as a practice that consists of ‘contemporary artistic expressions of live manipulated sound and image, defined as time based, media-based, and performative’. She further states that live audiovisual performance is difficult to theorise as there is no ‘specific style, technique, or medium’ (ibid.). However, it could be argued that the style, technique and medium of a live audiovisual performance piece is predominantly dictated by the audiovisual art practice that is being expressed. She notes that most of the artists surveyed identified themselves as practising live audiovisual performance. They then tend to narrow down their description to either live cinema, visual music or VJing (ibid. 2015: 128). This indicates that live audiovisual performance is a general umbrella term that introduces a live aspect to the other audiovisual areas. This begs the question as to whether the classification of live audiovisual performance, as a specific audiovisual art practice, is necessary. VJing and live cinema are inherently live. Contemporary visual music has also been described as including the possibility of live performance (McDonnell 2014). Generative AV pieces are often performed live. Further to this, there is an element of liveness in interactive physical AV pieces. It seems that the term *live audiovisual performance* can be applied to all forms of audiovisual art practice. Therefore, due to this universality there may be no need to designate it as a separate area of practice, rather a presentation context within which various aesthetic practices can be disseminated.

Fixed-Format

Fixed-format presentation is a common method for disseminating audiovisual art. Included under this specification are single-screen, multi-screen and surround-sound formats. This would be the more traditional way to experience audiovisual art with screenings going back to the absolute films of the early 20th century. One of the most common ways that audiovisual art is now consumed is online, through artists websites or video-hosting sites including YouTube and Vimeo. There are advantages

and disadvantages to this. The advantages are that there is more access to audiovisual art now than ever before. The disadvantages are that videos uploaded to these websites have to be compressed. This degrades the quality of the audio and visuals. Live screenings are commonplace at festivals and showcases. Organisations such as the Center for Visual Music (CVM) regularly lease films out to galleries and also host their own screenings of original 16mm and 35mm visual music films.

Mobile

Some high-profile musicians have developed AV apps as outputs for their work. Bjork's *Biophilia* (2011) was augmented by a mobile release which contained ten separate AV apps tied to each of the songs on the album. This allows the user to interact with a touchscreen device triggering visuals and sound. Radiohead released an interactive app for mobile devices entitled *Polyfauna* (2014) which brings the user through various surreal 3D environments populated by generative landscapes, floating objects and strange creatures. The environment is completed by processed and distorted samples of Radiohead's music. The user explores these environments, using their device as a window into the world, able to turn 360 degrees, giving the illusion of immersion. The user travels through the environment following a red dot. When reaching this red dot, the scenery suddenly changes to a new landscape with a new accompanying soundscape.

Simon Katan's *Conditional Love* (2016) combines generative AV practice with live performance utilising audience participation through mobile devices. It is a live performance piece in which the performer controls a *Supercollider* patch from a laptop with visuals projected on a screen. At the beginning of the piece, Katan sets up a private network and invites the audience to connect via their mobile phones. He then directs the audience to a webpage where they receive instructions typed in real-time directly to their phones. An interactive audiovisual object then appears on each phone, which the audience is then directed to interact with through 'caresses' that 'cause their avatars to grow in tamagotchi-like fashion' (Katan 2016). This interaction visually vibrates the avatar which also emits various 'purring' sounds. This creates a very personal audiovisual experience, so much so that at a recent performance, the combination of visual movement, sound and tactile interaction enhanced the author's sense of touch creating a brief illusion of the whole phone physically vibrating.

As the piece continues, sound is sporadically emitted from the audience's devices throughout the performance space. This dispersed sound augments the ambient audio textures, controlled by Katan, coming from a stereo speaker setup at the top of the room. The projected visuals then show each avatar on a grid moving around slowly. This arouses further interaction from the audience as they try to figure out which avatar is theirs. While all of this is going on, Katan is communicating directly

with members of the audience who were unable to access the mobile avatar but were able to connect to the webpage. In this communication, these audience members are instructed to look around them and consider the silliness of those enthralled by their mobile phones. By engaging in this performance, Katan and the audience are exploring 'the theme of narcissism and its digital manifestations' (ibid.). This piece brings many areas together and exists between generative AV, theatre and performance art. It uses web technologies and mobile apps to create an interactive and thought-provoking piece of art that is also fun to engage with.

2.3 Conclusion

A review of work within audiovisual art revealed a severe problem of definitional clarity permeating the field. Even the designation of audiovisual art *as* a field was ambiguous. After arguing for its recognition as such, centres of gravity around which contemporary work is emerging were explored. Concise descriptions of each area were attempted where possible, and characteristics central to each were discussed. In order to provide clarity when surveying the field, the various presentation contexts were separated from the core aesthetic practices. As part of this, the use of live audiovisual performance was re-conceptualised as a presentation context by arguing that it is a universal term that can be applied to any audiovisual aesthetic practice. The pieces *Ventriloquy I* and *II*, that are discussed in Chapter 6, are generative AV compositions that happen to be performed live. This is not to diminish the importance of live performance practices. When performing live, the real time unfolding of the piece in that particular space at that particular time is an incredibly powerful and important element of the piece. The argument here is that this liveness is the context within which the particular aesthetic practice manifests itself.

Now that a review of the audiovisual landscape has been undertaken, it is possible to locate the art practice that will be presented throughout the thesis (see Fig. 2.9). The aesthetic practice is located within the overlapping areas of generative AV and visual music. The medium through which this practice will ultimately be presented is the emerging medium of VR. However, to get to the intended destination, dissemination of the practice through live performance and shared immersive spaces will also be explored.

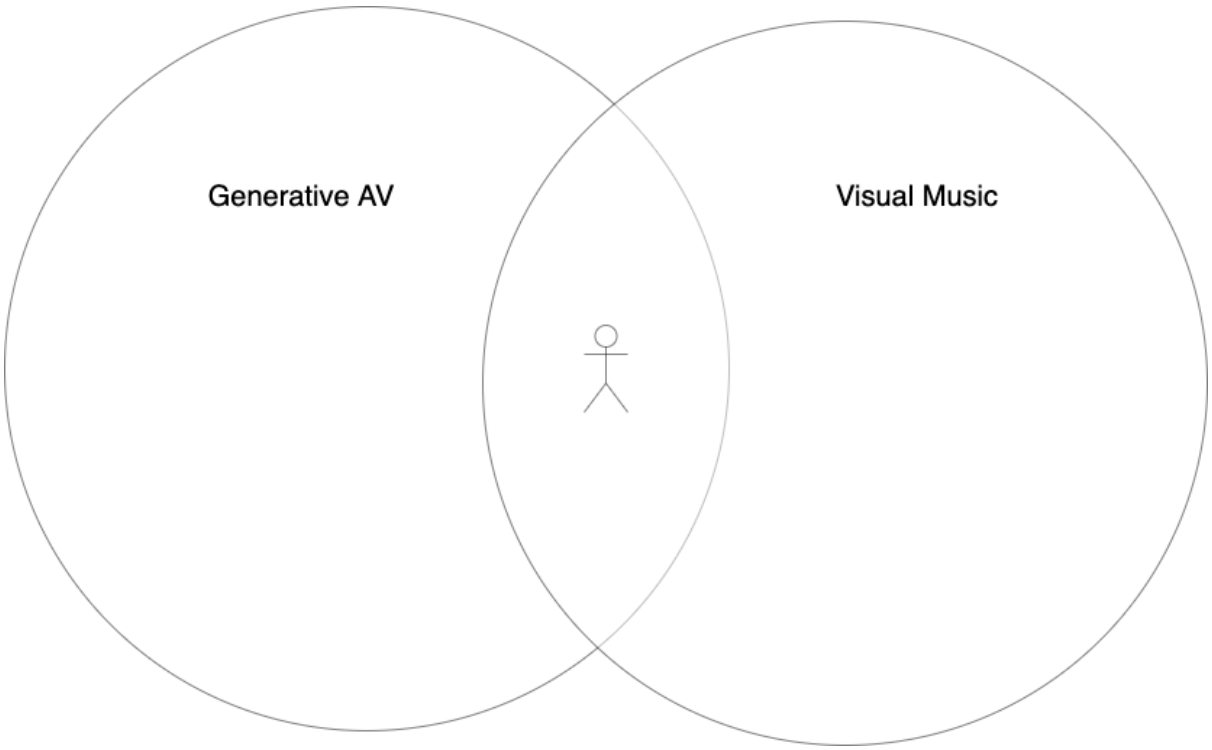

Figure 2.9 Location of my audiovisual practice.

Regarding the generative aspects of the practice, the pieces discussed in the following chapters will be primarily controlled using machine learning techniques. Machine learning algorithms will be used to create complex, non-linear mappings between audio and visual material. Regarding the visual music aspects of the practice, the method of arranging and developing the audiovisual material in time is influenced by musical structures and concepts such as exposition, development and repetition. This is especially apparent in the live performances of *Ventriloquy I* and *II*. Further, there is an emphasis on the idea of a unified expression similar to the aims of John and James Whitney when they were creating their *Five Film Exercises*:

> We are attracted by the prospects of an idiom as unified, bi-sensorially, as the sound film can be. Naturally, we have wanted to avoid weakening that unity, which would be the very essence of an abstract film medium. (Whitney 1980: 145)

It is clear from the range of practices and media discussed in this chapter that audiovisual art is fundamentally free from media specificity. It can be practised in any form of media we now possess. The core concern for audiovisual composers is the arrangement of audio and visual material such that the audience can perceive the results of their interaction. Vision is often considered the dominant sense (Sinnett, Spence and Soto-Faraco 2007). Therefore, in order to clearly perceive these interactions the job of the audiovisual composer is to try to subvert the generally accepted state of

visual hegemony. In doing this the audiovisual composer attempts to instil equal importance into their audio and visual source material. This concept of equality will be explored further in the next chapter.

Chapter 3 Equality and Balance in Audiovisual Composition

The previous chapter established the wider artistic context for the work in this thesis. The primary goal of this research is to utilise the emerging technologies of machine learning and virtual reality in the context of audiovisual art. In order to provide a compositional foundation for the later practical work, this chapter drills down into the concept of equality between audio and visual material. It will be shown below how this 'artistic credo' (Garro 2012: 106) permeates the audiovisual literature before proposing an approach to achieving equality through the pursuit of balance in certain aspects of the material. Three characteristics of an audiovisual composition will be explored for their potential to provide a sense of balance within a finished piece.

3.1 Audiovisual Composition

The main focus of an audiovisual work is the interaction between audio and visuals. This is what makes the art form unique. It is where meaning is found within the practice. Here, the term *audiovisual composition*, is used to refer to the practical application of compositional techniques to create audiovisual art pieces. The work in this thesis employs the use of abstraction. However, there are many examples of audiovisual work using representational material such as Freida Abtan's *The Hands of the Dancer*[13] (2011). With abstraction, the material aims to be non-representational (Fig. 3.1). The use of abstraction has been posited (Grierson 2005: 19) as a way to better perceive the structural interaction between audio and visual material.

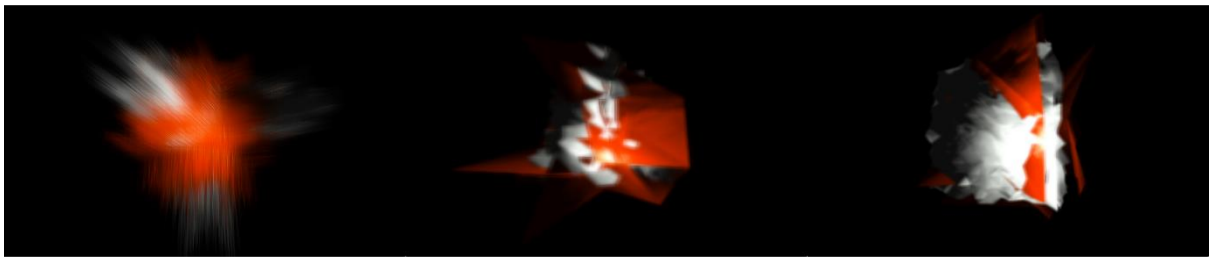

Figure 3.1 Stills of abstract visuals from my piece Ventriloquy I (2018).

As stated in the previous chapter, Grierson (2005: 5) describes his audiovisual art practice as a metadiscipline that combines several artistic fields of practice. This implies the idea of a meta-framework for composition and analysis. To discuss an audiovisual work only in terms of its visual

[13] https://vimeo.com/11692672 (accessed 01/12/21)

content, or indeed its musical content, would limit our potential understanding, as it would then only be confined within that specific frame of reference.

Treating audiovisual composition as a distinct discipline establishes a need for a vocabulary and framework of its own. This is not to deny the influence of musical theory and structure, or to sever any relationship with methods common to fields such as film or the visual arts. Rather, this approach attempts to provide meaning associated with relationships between audio and visuals. In this way it can complement the well-established musical and visual art frameworks that artists are already familiar with, whilst at the same time contributing to the emergence of an artistic vocabulary specific to audiovisual composition. This would be a step towards achieving what Carvalho and Lund aim for in their work, namely to 'move forward beyond definitions, to elaborate on philosophical, aesthetic, and theoretical implications, related to contemporary practices' (Carvalho et al. 2015: 7).

Audiovisual composition approaches throughout the centuries have taken several forms. A particularly popular approach, synonymous with the colour organ tradition, was to map the notes of musical scales to particular hues of colour. Examples of this approach include the instruments created by Louis-Bertrand Castel and Alexander Wallace Rimington in 1734 and 1893 respectively (McDonnell 2014). A more recent example is the virtual colour organ created by Jack Ox that utilises complex colour and textural mapping strategies to visualise musical compositions (Ox and Britton 2000). This approach can produce rich and complex artistic expressions.
Sometimes the mapping of particular tones to hues of colour has been attributed to artists' experience of synesthesia. Richard Cytowic (1995: 1) defines synesthesia as 'the involuntary physical experience of cross-modal association. That is, the stimulation of one sensory modality reliably causes a perception in one or more different senses'. As noted by Mitchell Whitelaw (2008: 260), whilst

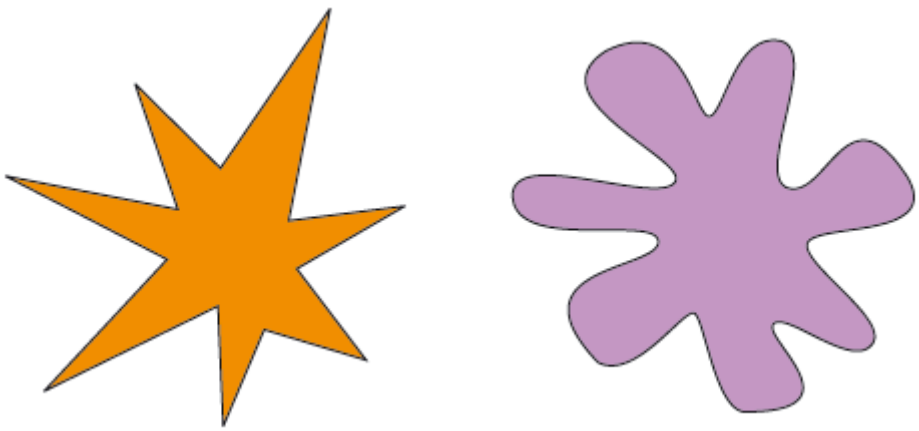

Figure 3.2 Kiki and Bouba.

neurological synesthesia is rare, 'auditory-to-visual synesthesia, "colored hearing" is much rarer still'. Whitelaw further notes that Wassily 'Kandinsky and composer Alexander Scriabin seem to have experienced it, while many other artists have been inspired by, or in some cases literally borrowed, synesthetic correspondences' (Whitelaw 2008: 265). Whilst the concept of synesthesia has inspired many artists to combine audio and visual material in rich and interesting ways, each mapping strategy tends to be subjective and can only really be utilised effectively by the individual artist. Whitelaw acknowledges that 'when it comes to practically manifesting that sensory relation it founders on the problem of the map, the pattern of correspondences' (ibid.). Whilst synesthetic associations between tone and colour are extremely rare and vary from person to person, studies have been conducted that show there are certain universal, non-arbitrary connections between sight and sound. These are grouped together under the term cross-modal correspondences. Charles Spence (2011: 973) defines the term as 'a compatibility effect between attributes or dimensions of a stimulus (i.e. an object or event) in different sensory modalities'. A famous example of an audiovisual cross-modal correspondence is the 'bouba/kiki' effect (Fig. 3.2) where the vast majority of people will match the word kiki with an angular figure and bouba with a rounded figure (Ramachandran and Hubbard 2001: 19). Cross-modal correspondences will be discussed further in the next chapter.

John Whitney Sr. is responsible for developing an approach to the composition of audiovisual material using his theory of differential dynamics (Whitney 1980). He states that 'the relationships of sight and sound would be served best if it were possible to compose both components of an aural / visual work within some common aesthetic such as harmony would offer' (ibid. 1980: 30). Whitney's 'differential dynamics' (ibid. 1980: 65) provides an important framework with which to approach the problem of how to combine audio and visual elements at a fundamental level. His compositions *Matrix III* (1972) and *Arabesque* (1975) were composed according to this approach. Whitney was motivated by a 'search for a coherent idea of abstract composition inspired by the rules of Pythagorean harmonics' (Abbado 2017: 55). Utilising these ratios, his intention was to create visual animations imbued with musical movement. His ultimate goal was to build a digital instrument where he could simultaneously generate audio and visuals based on his theory. During the 1980s he collaborated with Jerry Reed to create this instrument, which he called RDTD (Whitney 1997). It was his belief that 'sound and image composed on the same digital instrument will have totally revolutionary consequences' (Whitney 1980: 95).

Bill Alves (2005: 45) recalls that he was 'privileged to work with the computer animation pioneer John Whitney Sr. and was profoundly influenced by his ideas on how to apply musical concepts of harmony to visual arts of motion'. Adriano Abbado (2017: 55) identifies Whitney's work as being

responsible for leaving ‘indelible marks on the history of visual music’. One of the reasons his approach has proven so popular is because it is not dependent on his own subjectivity. Instead, it is grounded in mathematical ratios that provide universal meaning due to our lived experience of tonal harmony. He also states that his intention is to document his ‘own approach and to propose the seminal idea of making an approach’ (Whitney 1980: 7), thereby indicating that there are, of course, more than one.

Abbado (1988: 4) himself proposed an approach to composing audiovisual works utilising ‘fundamental correspondences’ between “audio and video events: timbre - shape, perceived location and perceived intensity’. Similar to Whitney’s approach, Abbado relies on universal correspondences between modalities to create his mappings. He states that he associates ‘low-energy spectra with smooth shapes and high-energy spectra with edged shapes’ (Abbado 1998: 4). This correspondence is also known as the bouba/kiki effect as described above.

Audiovisual practice is inherently influenced by multiple fields. This has led to approaches originating from both the visual and musical fields. Bret Battey (2015) proposes his idea of ‘fluid audiovisual counterpoint’ that is developed through an understanding of species counterpoint. Whereas traditional species counterpoint pedagogy relies on stringent rules, Battey (2015: 28) states that the main learning outcome ‘ultimately isn’t about the rules’. Instead, he locates value in the intuition developed by the student for identifying how ‘vertical (harmonic) and horizontal (melodic) relationships can be managed as a conceptually coherent ebb and flow of hierarchically ordered tensions and releases’ (ibid.). However, the rules themselves still hold relevance as they are the result of ‘perceptual criteria’ (ibid.). This formalisation of perceptual criteria into a system could point the way for a similar contrapuntal framework between audio and visuals. To this end, he proposes a distancing of inquiry from ‘simplistic note to image-object correspondence’ to instead ‘consider relationships among sonic and visual gestures, more qualitatively than quantitatively informed’ (ibid.). Ultimately, Battey reaches the conclusion that attempting to find a comprehensive system of counterpoint between audio and visual media ‘is not possible, nor even desirable’ (ibid.).

Hyde (2012) draws on concepts from *musique concréte* to propose an approach to the practice of video-based visual music composition. He notes Pierre Schaeffer’s late declaration of the failure of *musique concréte* due to the perceived impossibility of separating sound from source whilst still retaining meaning (Hyde 2012: 172). Hyde focuses on two areas of sound that inherently lean towards non-representation. Sound in a state of ‘tending-to-silence’ and ‘tending-to-noise’ (ibid. 2012: 170). He proposes that material in these perceptual zones resist our tendency to try to find representation

in sound or image. Visual correlations of sonic silence and noise are proposed based on an abstraction of these states to periods of low difference (silence) and high difference (noise). Hyde proposes that a visual silence could be represented as 'darkness, or black' (ibid. 2012: 174) and visual noise could be represented as 'video snow' (ibid. 2012: 175). The concept of avoiding representation through the use of noise is compelling and there may be a link between this idea and the concept of isolated-structural-incoherence discussed later in the chapter.

Unsurprisingly, film theory has proven a fertile ground from which audiovisual practitioners draw inspiration. Brian Evans (2005) relates the musical concepts of consonance and dissonance to tension and release. He then builds a syntax of visual music composition using the elements of two dimensional visual design and the film montage theories of Serge Eisenstein. His approach is based on 'the simple premise that the resolution of tension moves us through time' (Evans 2005: 11). He equates visual consonance with qualities such as 'dynamic balance or symmetry' (ibid. 2005: 13), proportions such as the golden ratio and resolution of temporal camera movement 'to cadences of visually balanced, well-composed moments' (ibid. 2005: 15). The logic follows that if 'rightness is codified and understood, wrongness is easily defined by not being right. We might call this wrongness visual dissonance, that is, visually active moments of tension in a temporal design' (ibid. 2005: 13). The concepts of tension, release and balance form important elements of the ideas that will be explored in the next section.

Michel Chion's (1994) writing on sound in cinema has been especially influential over the last two decades. His concepts have informed several approaches to audiovisual composition including those by Grierson (2005), Coulter (2010), Basanta (2017), Boucher and Piché (2020). Chion's idea of the audiovisual contract puts forward the notion that when viewing a film, the 'audio-spectator' (Chion 1994: 60) enters a symbolic agreement with the director or producer wherein they perceive the separate audio and visual elements, respectively emanating from the sound system and screen, as one audiovisual entity (ibid. : 216, 222). This contract allows the audio and visuals to affect each other, to come together and enhance the audio-spectator's perception of the other. This enhancement of the senses is called added-value.

> By added-value I mean the expressive and informative value with which a sound enriches a given image so as to create the definite impression, in the immediate or remembered experience one has of it, that this information or expression "naturally" comes from what is seen, and is already contained in the image itself.
> (Ibid. 1994: 5)

Whilst he is not the first person to recognise that the combination of audio and visual media produces an effect that can be described as more than the sum of its parts, his coinage of the term added-value to describe this phenomenon has become established in the literature. Grierson describes his own practice of audiovisual composition as the 'process of composing audiovisual works which exploit added-value' (Grierson 2005: 10). In doing this he is borrowing Chion's terminology and re-contextualising it as a central concern of his abstract practice. He builds on this by adding compositional methods distinct to the field including 'audiovisual synthesis' and 'audiovisual cutting' (Grierson 2005: 2). The act of identifying techniques like this contributes to the development of an audiovisual vocabulary. Some abstract audiovisual practitioners will not be aware of Chion's terminology, as it is situated first and foremost in the field of film theory. However, the re-contextualisation of his ideas demonstrates the permeable boundaries between many audiovisual practices. The development of a compositional vocabulary like this, that is specific to audiovisual art, can provide direction and meaning for artists. Their work can then be discussed on its own terms rather than relying solely on musical or visual based frames of reference. This is the distinction of self-definition that Whitney speaks of.

Chion is also responsible for identifying a special case of added-value that he calls synchresis. He defines synchresis as 'the spontaneous and irresistible weld produced between a particular auditory phenomenon and visual phenomenon when they occur at the same time' (Chion 1994: 63). Boucher and Piché (2020) use Chion's concept of synchresis quite successfully within their Sound/Image Relationship Typology. Here they specify compositional uses of synchresis, diegesis, time and narration within a *vidéomusique* context. Their identification of synchresis as a defining principle of their work cements the concept as a primary concern of the *vidéomusique* artist. The concepts of synchresis and added-value are important to artists working across several areas of audiovisual art. They are also relevant to the ideas being discussed in this chapter.

3.2 Audio and Visual Equality

A recurring sentiment within audiovisual discourse is the desire to give equal importance to the composition of both the audio and visual elements of a piece. The frequent nature of this concept in the literature indicates that this is an important issue for audiovisual artists.

Lund and Lund (2009: 12) consider the balance between sound and image a fundamental concern of their definition of visual music, where they state that the basic objective of the practice is to achieve 'evenly balanced or equilibrated interplay between visual and acoustic components'. Battey (2015)

acknowledges ‘the commonly stated goal to have sound and image be of equal importance’ when composers apply the concept of ‘counterpoint as a metaphor for audiovisual composition’. Garro (2012) acknowledges the general primacy of sight over sound in human perception. He argues that upon experiencing an audiovisual work, the audience must be conscious of the fact that audiovisual artists ‘hold the primacy of both ear and eye together as their artistic credo’ (Garro 2012: 106). Ryo Ikeshiro’s live audiovisual pieces, *Construction in Zhuangzi* (2011) and *Construction in Kneading* (2013), are based on the simultaneous creation of audio and visual elements from a common source of data. As discussed in the previous chapter, he calls this approach audiovisualisation. He states that it ensures a non-hierarchical sensory structure in his pieces as ‘the moving image is no longer a score for performers but intended to be experienced in tandem with the sound’ (Ikeshiro 2013: 58). Rogers (2014: 80) talks about the simultaneous experience of sound and image as a ‘holistic form of engagement’, citing Norman McLaren’s *Synchromy* (1971) as an example in which he ‘achieves a single audiovisual voice where neither sound nor image can successfully be extricated from the other’. Mollaghan (2015) states that John and James Whitney regarded the equality between audio and visual elements in their *Five Film Exercises* (1943-1944) as a core direction for original audiovisual compositions. She states that they ‘were adamant that their films should be original audiovisual compositions in which the sound and image shared an equal partnership’ (Mollaghan 2015: 146). However, even if this is their intention, their approach to the *Five Film Exercises* were not always interpreted that way. Birtwistle (2006: 161) sees the Whitney brother’s optical process as a primarily visual practice where the sonic elements are in constant subjugation.

> With the Film Exercises the sonic is allowed into the Whitney project, but on condition that it is shaped, moulded, driven and curtailed by the visual. The film certainly proposes an audio-visual synthesis, but one in which the sonic is absorbed by the primary term of the audio-visual contract, which despite word order remains the visual.

Birtwistle’s analysis of the Whitney brother’s process illustrates the difficulty for the audiovisual composer in attempting to treat audio and visual material equally. Nevertheless, this evidence points to the desirability and importance of equality between the audio and visual material within an audiovisual composition. This is further supported by Weisling et al. (2018: 345), who found that participants in their survey ‘overwhelmingly agree that the aural and visual components are equally important’ in their practice.

3.3 Audiovisual Balance

The perceptual result of this compositional desire to treat both audio and visual material with equal importance could manifest as a sense of balance between the elements of a piece. If the elements of a piece achieve a sense of balance, perhaps there are forces that are responsible for maintaining or destroying it. Unpacking the concept, and trying to find some practical means by which the perceptual sense of balance could be achieved, three forces will be presented that may have the potential to affect the perceived balance within an audiovisual composition. These are *relative-temporal-motion*, *isolated-structural-incoherence* and *relative-expressive-range.* These forces could possibly be harnessed and used to purposefully alter the experience of tension, release and added-value across the duration of a composition. The concept of audiovisual balance could be described as follows:

> Audiovisual balance is the extent by which our attention is drawn to material being presented either visually or sonically, affecting our ability to experience the composed inter-relationships.

In order to further illuminate the concept, an example from outside of the field will be discussed. Mainstream documentary film-making is an example of popular media that is actually quite balanced, audiovisually. Similar to Ikeshiro's audiovisualisations (Ikeshiro 2013: 58), the documentary film-maker utilises the central narrative as their source of data from which they build the spoken audio and visual elements. The narration, interview questions and visual footage are composed meticulously to communicate the narrative to the audience in such a way that they will become emotionally involved. The audio and visual elements here play an equal role in uncovering the story that the director wants to tell.

An audiovisual work may have moments where one modality will dominate the other. During these moments the potential for added-value could be minimised as the perception of the audience is dominated by one side over the other. A strategy aimed at maximising the potential for added-value experiences, could be to find ways to perceptually balance the audio and visual material, within an audiovisual composition, as much as possible. During a perceptually balanced section, the potential for added-value experiences to occur could be increased due to the ability of the audience to perceive the audio and visual material simultaneously, rather than being fixated on one sensory mode over the other. In an ideal theoretical sense, perfectly balanced material would mean that the audio-spectator is not consciously focused on either modality in particular. Instead, they are allowing the material to cognitively bind in their perception as an audiovisual whole. The aesthetic choices and mapping approach of the composer could affect the audio-spectator's sense of balance.

There are certain cognitive characteristics that affect our perception in various ways so as to automatically give primacy to either sight or sound in a given situation. Sá (2016: 30) provides an in-depth discussion of what she calls 'sensory dominance' and provides strategies for minimising the natural dominance of sight over hearing so that sound can be brought to the forefront of the audience's perception during a live audiovisual performance. She identifies two situations in which the sense of sight will automatically dominate the sense of hearing. The first situation arises when there are extreme discontinuities within the visual material. The second situation arises when there is a strong correlation between the visual material and the source of the audio (ibid. 2016: 230). That is, when the sound and the image conceptually match, the sound can be easily perceived as belonging to the visual object. The two scenarios identified by Sá cause the audience's visual attention to become perceptually dominant. This in turn clouds their perception of the audio. For Sá, this is undesirable as she is primarily a musician and wants the visuals to augment the audio whilst maintaining the music at the forefront of the audience's perception.

However, in the context of this discussion, the motivation here is to explore the concept of equality between the senses. Therefore, the conceptualisation of audiovisual balance is appropriate here as it describes a relationship between audio and visual material in a non-hierarchical way. Sá's work suggests that this balance can be purposefully swayed by the composer to favour either the audio or the visuals. Taking this line of thinking further, it may be possible to purposefully compose material that is perceived equally, in a state of balance.

For an example of how audiovisual material can slip in and out of balance see the example video *ventriloquy_ex2.mov* in the accompanying media pack[14]. It is also uploaded to Vimeo[15]. This clip was recorded during preparation of the material for *Ventriloquy I*. The finished composition will be discussed in Chapter 6. The clip shows two cubes, with their structures displaced by noise. The audio is made up of harsh FM-influenced textures that pulsate at various rates. Here, due to the repetitive noisy movement in the visuals and the speed of the audio oscillations, the audio and visuals move in and out of synchronisation. This perceptual synchronisation is an example of what Nicholas Cook (1998: 78) calls 'ventriloquism', which refers to when the visuals adopt rhythmic qualities from the audio. The cross-modal binding of the material, due to similarity of motion, is also an example of synchresis and its adherence to 'the laws of gestalt psychology' (Chion 1994: 58). Boucher and Piché (2020: 20), further explore this phenomenon, describing it as 'gestural synchresis'. Regarding

[14] mediaFiles/ventriloquy/ventriloquy1.
[15] https://vimeo.com/251893597 (accessed 24/06/2020).

audiovisual balance, the clip starts with the audio and visuals in sync. The audio and visual material could be considered to be well-balanced at this point. However, as the material unfolds, the tight synchronisation seems to drift, causing the audio and visual streams to separate slightly. The balance of material becomes more unsteady as you focus on trying to see if the material is in sync, first contemplating the visual movement, then switching focus to the audio rhythm. Then at 0:08, the object morphs into a new shape. At this point the audio rises in pitch and the visual movement seems to speed up. The audio and visuals combine to form a single perceptual object. The balance is restored. This example relies on synchresis and audiovisual ventriloquism to highlight how audiovisual material can slip in and out of balance in a temporal way. The use of material as it is tending-to-noise seems to accentuate the shifting balance. However, audiovisual balance could also be affected by other factors such as the structural completeness of the media and also the relative-expressive-range of the material. These factors, along with the relative motion of the elements within the media as demonstrated above, will be discussed in the next three subsections.

In the event of achieving audiovisual balance in one of the above mentioned cases, an added-value experience may not be guaranteed due to another aspect of the composition being out of balance. Take *MetaVision*[16] (2011) by Colin Goldberg and Intersolar as an example. The individual elements that make up the audio and visual material could be analysed in terms of aesthetic richness and variation, and it could be argued that this piece is fairly balanced in terms of these elements. As will be seen below, this will be referred to as the relative-expressive-range. However, if we then analyse the piece in terms of how the visual material moves in relation to the audio material, it could be argued that the piece is unbalanced in this aspect. This will be referred to as relative-temporal-motion below. This imbalance in the material could be a barrier to the experience of added-value in the piece. Following from this, a conceptual relationship may be represented as follows:

> Longer periods of audiovisual balance may lead to a higher potential for added-value experiences.

3.3.1 Relative Temporal Motion

The temporal motion of elements within an audiovisual composition could be conceptualised as the perceptual manifestation of kinetic energy being applied to those elements. This energy can be

[16] https://www.goldberg.art/audiovisual-works/ (accessed 26/11/2021).

harnessed in one modality and transferred to the material in the other modality. For example, audio reactive visuals move due to the transfer of audio analysis data. The energy can also be generated within a third-party system and transferred to either modality in isolation or both modalities simultaneously. An example of this is the audiovisualisation of a dynamical system. These approaches can be achieved through mapping of data and parameters. The energy could also simply be generated in, and only affect one, modality. The motion within each sensory modality could then be aligned temporally, without specific mapping of data. This kinetic energy, or more precisely, the perceptual motion that arises from its distribution, could be a source of potential imbalance within a composition. As such, the concept of relative-temporal-motion describes:

> The perception of audio elements changing in time, relative to the perception of visual elements changing in time.

Music is an art of movement. A tonal melodic line contains latent energy that manifests itself perceptually as the musician moves from one note to the next, creating tension and release depending on the position of the notes in the scale. This is culturally dependent of course, but for anyone familiar with the major scale, there is an expectation of tonic resolution when the leading note is played in the context of that scale. Animation relies on the simulation of physical forces to move visual elements through space. These are examples of energy being manifested as motion, as it is perceived within each sensory modality in isolation. If the material in one modality exhibited motion that was completely at odds with the relative material in the other modality, it could be argued that there is an imbalance present in the relative motion of the material. However, if the material were to then align, this could be a way to create tension and release within a composition.

When employing mapping techniques, we often see audio analysis data mapped to visual elements, causing them to move. Here, the visual elements are deriving their motion from the relative motion in the audio material. This does not only have to be spatial motion, but can include dynamic morphing of textures and colours. This approach will create tightly synchronised motion across the modalities. It is also possible to harness visual movement and map that to audio processes. The goal of the mapping is to translate the movement in such a way that it is faithfully represented in the other modality. However, this approach presents some risks, if the mapping is too transparent, with a one-to-one relationship between the elements, this can render the material uninteresting (Dannenberg 2005: 28).

When this happens, there is only one source of energy responsible for driving the motion of the entire piece. The same could be said when motion is extracted from visual elements and used to drive audio processes. This could be seen as an imbalance in the energy distribution within the piece. If this is the case, this would appear to represent an inequality in the agency of the material. If this inequality pervades the entire piece, then it could be argued that the piece is not equally balanced. However, even in this context, if there are several layers of mappings happening, some transparent and some oblique, the relative movement of the piece could still be perceived as being balanced. When observing an audiovisual composition, it is not always obvious where the agency for motion is situated. Mapping approaches and terminology will be discussed in more detail in Chapter 5.

Consider a system programmed in such a way as to rely entirely on the direct mapping of audio features to the movement of visual content. A simple example of this would be any basic music visualiser such as the one included with the Windows Media Player. Here the visual movement is entirely derived from the audio content. The visual material has no kinetic energy of its own. With applications of this type there is sometimes not much thought given to deep structural cross-modal correspondences. This results in simple mappings of audio transients to visual expansion. No other elements of the music or the visuals are mapped or aligned. This results in movement in each modality that is not well-balanced, peppered here and there with tightly-synchronised transient mapping. This may demonstrate an imbalance between the audio and visual material if assessed through the lens of audiovisual composition. Callear (2012: 31) states that systems such as these result in a 'dominant' and 'dependent' relationship between the two modalities. In order to retain control of audiovisual balance, it may be helpful to avoid a system that depends entirely on a dominant/dependent relationship. Sound is the dominant modality in the context of music visualisation applications. The visuals are an illustration of the analysed audio signal. Systems such as this generally allow for simple, transparent connections between the audio and visual elements. This makes them successful as music visualisers. However, this would not be enough in the context of an audiovisual composition. When talking about audiovisual composition, the ability for the composer to manipulate the audience's experience of audiovisual balance could be restricted if the system only allows for mapping of parameters between modalities in a single direction. For example, the composer would not be able to influence the morphological alignment of audio and visual elements, which has been posited by Katan (2012: 185) as a possible avenue of interesting experimentation 'involving the temporal separation of sonic and visual events'.

Bret Battey's *Estuaries 3*[17] (2018), grounded in his 'idea of gestural and fluid counterpoint' (Battey 2020: 277), contains sections (4:51 - 5:56) that are populated with 'many note-to-motion correspondences' (ibid. : 278). During these sections, the same generative process that is used to define the musical gestures of the piece, is used to define visual motion through the use of animation keyframes. This demonstrates a correspondence of motion across the media. It could be argued here that there is an equality, or balance, in the relative temporal motion of the audio and visual elements of this piece. Indeed, Battey describes his approach as a 'phenomenology-based interdependence of gestures' (Battey 2015: 30).

The opening section (0:00 - 0:41) of *33*[18] (2014) by Scott Kiernan and Victoria Keddie plays with the audio-spectator's sense of relative temporal motion. Here the audio material is made up of several rhythmic audio textures in addition to some tonal electronic patterns. There is a regular tempo, unmatched in the movement of the visual elements. Here, the visual material is monochrome and consists of regions of visual static, some scrolling from left to right and some stationary. The relative temporal motion moves in and out of balance as the piece moves on, creating sections of cross-modal unity (2:37 - 3:44) compared to sections of more independent media streams (3:59 - 4:20). There is a sense of tension in the unbalanced sections that resolves when the material is gesturally aligned. Perhaps this is a compositional technique that can be utilised going forward.

3.3.2 Isolated Structural Incoherence

Isolated-structural-incoherence is a concept that has emerged through contemplation on the importance of unity between the audio and visual elements of a work. This desire for unity is also apparent in the work of John Whitney.

> We are attracted by the prospects of an idiom as unified, bi-sensorially, as the sound film can be. Naturally, we have wanted to avoid weakening that unity, which would be the very essence of an abstract film medium. (Whitney 1980: 145)

The concept of isolated-structural-incoherence examines the notion of structural completeness in the audio and visual material, when experienced separately from each other. The relative coherency of the material in each sensory mode may have an effect on the audiovisual balance of the piece. When assessing audiovisual work, the audio-spectator can ask themselves:

[17] https://vimeo.com/264837797 (accessed 02/12/2021).
[18] https://vimeo.com/88879966 (accessed 10/11/21).

- Could I listen to the audio on its own and be satisfied that I have experienced a fully developed piece of music or sound art?
- Could I watch the visuals and enjoy them as a fully developed work in their own right?

The concept of isolated-structural-incoherence posits that if either element can be isolated from the other and experienced as a self-sufficient work in its own right from start to finish, it could weaken the overall audiovisuality of the piece. The concept of isolated-structural-incoherence could be defined as:

> The potential for material that is structurally complete to dominate the audio-spectator's attention thereby causing an imbalance in the perception of the piece.

When considered in relation to added-value, this concept seems logical. If the intentional combination of audio and visual material results in an experience that can be said to be greater than the sum of its parts (i.e. exhibit added-value), then listening to, or watching each element in isolation will result in a reduced experience. Although the nature of added-value is such that the combination of an individually coherent visual work and an individually coherent musical work may well result in an added-value experience, it is argued here that the purposeful combination of structurally incoherent, or ambiguous media may be a useful strategy for achieving more unified and balanced audiovisual works.

Throughout an audiovisual work, there may be sections where the audio and visuals are individually more coherent before they return to ambiguity. This may create dynamic variation in the perceived audiovisual balance of the material. The relationship between congruent/incongruent material and tension/release is a common theme in audiovisual theory and can be seen in the writings of Grierson (2005: 22 - 23) and Callear (2012: 43 - 46). The very idea of an audiovisual compositional language is based on correspondence and interaction between audio and visual elements. This implies a mutual dependence between the two media. If this dependence, throughout the course of a whole piece, is absent, then it would be difficult to analyse the piece as an audiovisual composition as the term is understood here. If the audio or visual material are coherent on their own, then it could be argued that they are less likely to exhibit any dependence on additional material in any other sensory modality.

If the audio from an audiovisual composition can be listened to without its visual counterpart, then why have the visual counterpart at all? In this situation the visuals may act as a decoration of the audio, which is completely fine, but this would situate the piece in the world of the music video.

Similarly, if the visual element of an audiovisual composition can be viewed and understood in isolation from the audio, it is more akin to an animation or film where the music acts as a soundtrack in support of the visuals. If we consider the example of the documentary film discussed above, we can see that this form of media actually adheres quite well to the concept of isolated-structural-incoherence. If the script or visuals are divorced from each other, comprehension of the narrative would likely be severely diminished. Further, the script and visuals are completely dependent upon each other. The footage needs that exact spoken audio track. Likewise the spoken audio track needs that exact footage.

This example actually highlights an interesting characteristic of isolated-structural-incoherence. Abstraction frees both the audio and visual elements from representation, inherent association and narrative. It is because of this that the abstract audiovisual composer must work harder to ensure that both sensory modalities work together. With narration and footage in documentary film-making, isolated-structural-incoherence is guaranteed. With abstract audiovisual composition, it is not guaranteed.

The concept of isolated-structural-incoherence is in direct contradiction to the philosophy of Jean Piché and his practice of *vidéomusique*, who believes that the audio and visual elements should have separate, coherent identities. When speaking about *Sieves* (Piché 2004), he explains that for him the music should be able to exist as a coherent piece on its own. He states that this is to differentiate the audio in an audiovisual composition from the soundtrack to a film. If the film soundtrack were to be severed from the film 'the music loses its reason to be' (ibid.). He also states that the visual element should be able to work well on its own. However, when the two are combined, the audio-spectator should 'get the immediate impression that one cannot be without the other'. This seems like a contradiction in the context of the discussion presented here. If the audio-spectator was under the impression that one element could not exist without the other, then surely, in isolation, each element would seem incomplete.

The intention in exploring the concept of isolated-structural-incoherence, within the context of audiovisual balance, is not to present it as the only way to arrange audiovisual material. There are plenty of examples of audiovisual compositions that could be shown to contradict the idea. *Perspectrum*[19] (1975) by Ishu Patel is a well-balanced composition combining an instrumental tune played on the Japanese koto, with boldly coloured, solid visual shapes that move elegantly in time

[19] https://www.youtube.com/watch?v=M1f2D14TiDo (accessed 16/11/21).

with the music. The music does exist in it's own right as a traditional Japanese tune[20] whilst the visuals could also exist independently as they follow the structure of the music quite tightly. The kinetic energy of the music is integrated into the visual movement. Another example is *Jazz Orgie*[21] (2015) by Irina Rubina and Emanuel Hauptmann. The combination of jazz instrumentation and tightly synchronised visuals, reminiscent of Kandinsky, results in an audiovisually balanced expression even though the material in either modality could be enjoyed in isolation and still provide a coherent statement. With this in mind, the argument for isolated-structural-incoherence is that it is a possible strategy that the audiovisual composer can employ, in conjunction with the other concepts introduced here, to try to create a well-balanced, unified audiovisual work.

In order to illustrate the concept and discuss some issues that arise from it, some contemporary work, viewed through the lens of isolated-structural-incoherence, will now be discussed. *This City* (2015) by Mark Eats is an example of a piece in which the audio could exist quite sufficiently on its own. The visuals are representational and show a network of roads with cars, streetlights and traffic lights. The music is performed live by Mark Eats on a range of synths and midi controllers. Certain parameters of the audio are mapped to the visuals and affect them in real time. For instance, as he opens the filter on his Sub39 synth at 1:18, the cars lose gravity and float into the sky. They hang there weightless as the music builds tension underneath. An ascending scale reaches the leading note before resolving on the tonic as the cars drop back onto the road at 1:40. This is a visual representation of the drop that is a staple of electronic music. This build up and resolution of tension is crafted quite well in this instance. However, this is a fully formed musical piece in and of itself. It follows its own chord progression and obeys the laws of tonal harmony. The correspondences here between the music and the visuals are transparent one-to-one mappings and therefore demonstrate a clear cause and effect relationship. For example, at 2:15 the performer introduces a delay effect into the music. The visuals at this point become blurred with the cars leaving white and red trails across the screen. Although these mappings are effective, the self-contained nature of the audio and visual material suggests that they are not necessarily dependent on each other. Further, the regular beat structure and lack of visual elements to balance this, places the music in a dominant position. It could be argued that this piece might successfully be described as a musical composition with supporting visuals. It works very well in this context. However, if we were to analyse it in terms of isolated-structural-incoherence, the well-defined solidity of the music could work against it. Indeed, the artist here is not working in an abstract context, so perhaps such a critical analysis is unfair. However, the intention in using it here is simply to try to illustrate the idea of isolated-structural-incoherence.

[20] https://www.youtube.com/watch?v=3wvQk57vwHY (accessed 16/11/21).
[21] https://www.puntoyrayafestival.com/en/tv/films/jazz-orgie (accessed 16/11/21).

An abstract audiovisual composition needs to demonstrate a deep and necessary connection between the audio and the visuals. In terms of isolated-structural-incoherence, it could be argued that representational visuals lend themselves more readily to narrative, so in the absence of one they can be quite incoherent structurally. However, popular music that is tonal and follows a regular beat has a very strong structure. This is the case with the above piece. Perhaps this structural solidity is perceptually complete, thus acting as a barrier and preventing a close structural bond with the visuals. This completeness is possibly skewing the audiovisual balance in favour of the musical material for the entire piece.

However, this may not be the full explanation. Perhaps the barrier to comprehensive perceptual binding is not simply because both audio and visuals could exist on their own. This may be a high-level observation that sets the context for further exploration. The mappings in this particular piece are made by the performer at a high perceptual level and are narrative or semantic in character. If we look at the structural level, there doesn't seem to be any robust mapping between the prominent rhythmic and harmonic elements of the music and the activity of the visuals. This may be the deeper issue that isolated-structural-incoherence points to. It was shown above that compositions can contain material that is structurally coherent in isolation yet still present a unified audiovisual expression. Each of these pieces display tight rhythmic, structural mappings. This suggests that when a piece does not exhibit a sense of isolated-structural-incoherence, for example, when the music displays a regular beat, the rhythmic mapping schema becomes more important for creating a deep bond across the material. Conversely, when a piece does exhibit a sense of isolated-structural-incoherence, perhaps the more ambiguous structure of the material allows the observer to perceive both modalities in a clearer way, thereby allowing the observer to create their own bindings. In this scenario, perhaps the audiovisual composer has more freedom to create metaphorical correspondences without having to worry so much about structural or rhythmic mappings.

Therefore, it may be more difficult to create audiovisual compositions using strongly-structured and narrative material. This is similar to Hyde's (2012) focus on material as it was tending-to-noise. As material tends towards noise it loses structure and also representation. Using these examples from outside the context of abstract audiovisual composition reinforces the point that isolated-structural-incoherence and audiovisual balance are unique problems that the audiovisual composer needs to address. This helps to refine and focus the task that the audiovisual composer is faced with when approaching their work.

Paul Prudence's *Cyclotone III*[22] (2015) is an example of a generative abstract audiovisual composition. The visuals are abstract monochrome shapes, mainly rectangles, arranged in various circular and spherical formations. The audio is made up of mechanical clicks and machine-like noises. The audio is very tightly synchronised in parts to certain visual movements, creating strong synchretic audiovisual correspondences. The audio here acts in an almost diegetic fashion. This purposeful blurring of lines between diegetic sound and music is a common approach in audiovisual media. Rogers (2019: 261) identifies the creative use of diegesis in the work of David Lynch. Walter Murch (2000) also speaks about the metaphorical use of diegesis to create ambiguity between the sound and image.

There are also less tightly-synchronised ambient sounds, within *Cyclotone III,* that align themselves metaphorically with the floating characteristics of the spheres and the smooth movement of the circular arrangements. When experiencing this piece there is a sense of interdependence between the audio and visual material. Perceiving the audio in isolation, it loses a certain amount of structure and meaning, becoming more ambiguous. This suggests that the audio here would not be able to survive as a coherent piece in isolation. The visuals could be more suited to isolation than the audio as they are structurally solid. However, the combined aesthetic coherence between the material seems to balance this out somewhat. The audio certainly imbues the visuals with an ethereal yet mechanical personality that enhances and instils character in them. At the same time, the visuals lend the audio a definite structure and direction. It could be argued that the sense of isolated-structural-incoherence experienced within the audio of this piece may leave room for the elements to unite within a well-balanced audiovisual expression.

Cook's (1998) concept of 'gaps' in a media stream could be a useful way to visualise the concept of coherence within the material. He states that a media stream is gapped where there is an 'implication but not realization', or 'absence of closure' (Cook 1998: 141). Perhaps it is this quality that makes either the audio or visual material incoherent when analysed in isolation. The incoherent nature of the material could contribute to an increase in potential for interaction within composed audiovisual content. This would suggest that an autonomous, fully realised musical or visual work may leave less room for interaction with complementary material. Cook states that any example of multimedia where 'one or more of the constituent media has its own closure and autonomy is likely to be characterised by contest' (ibid.: 103). By 'contest', Cook means 'the sense in which different media are, so to speak, vying for the same terrain, each attempting to impose its own characteristics upon the other' (ibid.:

[22] https://www.transphormetic.com/Cyclotone-III (accessed 14/04/2022).

103). Although Cook sees this state of contest as desirable, when attempting to create a unified and balanced audiovisual composition, it may be helpful to create material that leaves room for cross-modal integration.

3.3.3 Relative Expressive Range

When considering the concept of equality, put into practice as an effort to find a balance between two different senses, we could consider the aesthetic richness and thematic variation of the material, presented in each of the sensory modalities, as a factor that could cause an imbalance in the perception of the finished piece. Aesthetic richness describes the way in which individual elements of each modality come together to create the final perceived output. For example, a complex sound wave emitted from a violin is generally considered to possess a richer timbre than a pure sine wave. This is due to the combination of many individual sine waves and their phases. In the visual domain, a multi-coloured shape with surface detail, lighting and shadows, could be described as texturally richer than a monotone shape with no surface detail, flat lighting and no shadows. In addition to these individual elements, the amount of variation of this material, across the composition, could be analysed as a factor that could affect the audiovisual balance of the piece. Consider a situation where the visual material moves through several scenes, with many permutations of the constituent elements, compared to a soundworld that is relatively static and unchanging. This could be seen as an example of inequality within the piece. The use of the term *expressive range* refers to both:

> The perceived richness of the constituent elements and the amount of variation of these elements, observed throughout the composition.

John Whitney's *Moon Drum*[23] (1991), the first in a series of twelve pieces inspired by Native American culture, is a work that could be analysed in terms of relative-expressive-range. The series as a whole is a substantial addition to Whitney's catalogue and belongs to his later output. As such, it represents the culmination of decades of work with differential dynamics. It is also an important link in the evolution of visual music practice towards utilising the computer as the main creative tool. It could be seen as a direct link between contemporary generative work and the film based visual music tradition. Whitney composed both the visuals and music for this series of works, thereby realising his dream of having a system that allowed for the audio and visuals to be composed together. Previously he made a conscious decision to only concentrate on his visual practice.

> For the time being, I elected to put aside the musical problem as it bore along my own long-term plans while I would concentrate upon new prospects for optical differential dynamics. I

[23] https://archive.org/details/JohnWhitneyMoonDrum1991 (accessed 14/04/2022).

> would settle for whatever music I might find for each new graphic composition since my optical studies were the immediate challenge. (Whitney 1980: 44)

The audio in the first *Moon Drum* section contains poorly-rendered drum samples and some basic synthesis which severely limits the richness of the audio aesthetic. The visuals are colourful and are presented with a wide range of motion and form, in keeping with Whitney's unique style. They could be said to form an aesthetically rich palette of visual forms and motion. However, the lack of aesthetic richness in the audio, perhaps caused by low-bitrate sampling due to hardware limitations of the time, creates an inequality in the relative-expressive-range of the material, causing a perceptual imbalance within the piece.

Test Pattern[24] (2008 - ongoing) by Ryoji Ikeda is a long-running series of installations and performances. The compositional material consists of data, taken from various sources, converted to binary and directly sonified. The visual material consists of black and white lines separated into two or more strips. The audio material could be said to be more expressive than the visuals here. The live performances[25], in particular, suffer from this imbalance. At a performance in London in 2017, the author found the audio performance to be quite expressive, varied and aesthetically rich, whereas the visuals consisted of simple black and white square patterns. This skewed the expressive balance of the piece.

In contrast, *[DUST]*[26] (2011) by Mariska De Groot and Yannis Tsirikoglou maintains a pleasant balance of expressive range throughout the piece. The audio and visual material is balanced quite well in terms of the amount of elements present in each modality. For example, when there is a sparse audio section (2:00 - 2:11 and 4:22 - 5:19), the visuals are also appropriately sparse. Conversely, when the number of elements in the audio material increases (2:15 - 4:20 and 6:16 - 8:14), they are matched in the visual material. In addition to this, there is a balance of aesthetic richness in the elements that make up the material. There are textural correlations between the analog audio crackle at 4:13 and the analog spot artefacts in the video-strip animation.

3.4 Conclusion

This chapter examined the concept of equality between audio and visual material in audiovisual compositions. This concept was interpreted, in practical implementation, as a sense of perceptual balance within a piece. Three characteristics of audiovisual compositions were posited as potential

[24] https://www.ryojiikeda.com/project/testpattern/ (accessed 10/11/21).
[25] https://www.youtube.com/watch?v=B6eocxPgnbQ (accessed 10/11/21).
[26] https://vimeo.com/24742293 (accessed 10/11/21).

areas through which the audiovisual balance of a piece may be controlled. Fig. 3.3 shows a visual conceptualisation of the forces that may affect audiovisual balance. The relationship between these forces could be summarised like so:

> Relative temporal motion, isolated structural incoherence and relative expressive range are perceptual forces that may affect the audiovisual balance within a composition. Longer periods of well-balanced material may help to achieve a unified audiovisual expression with the intention of maximising the potential for added-value experiences.

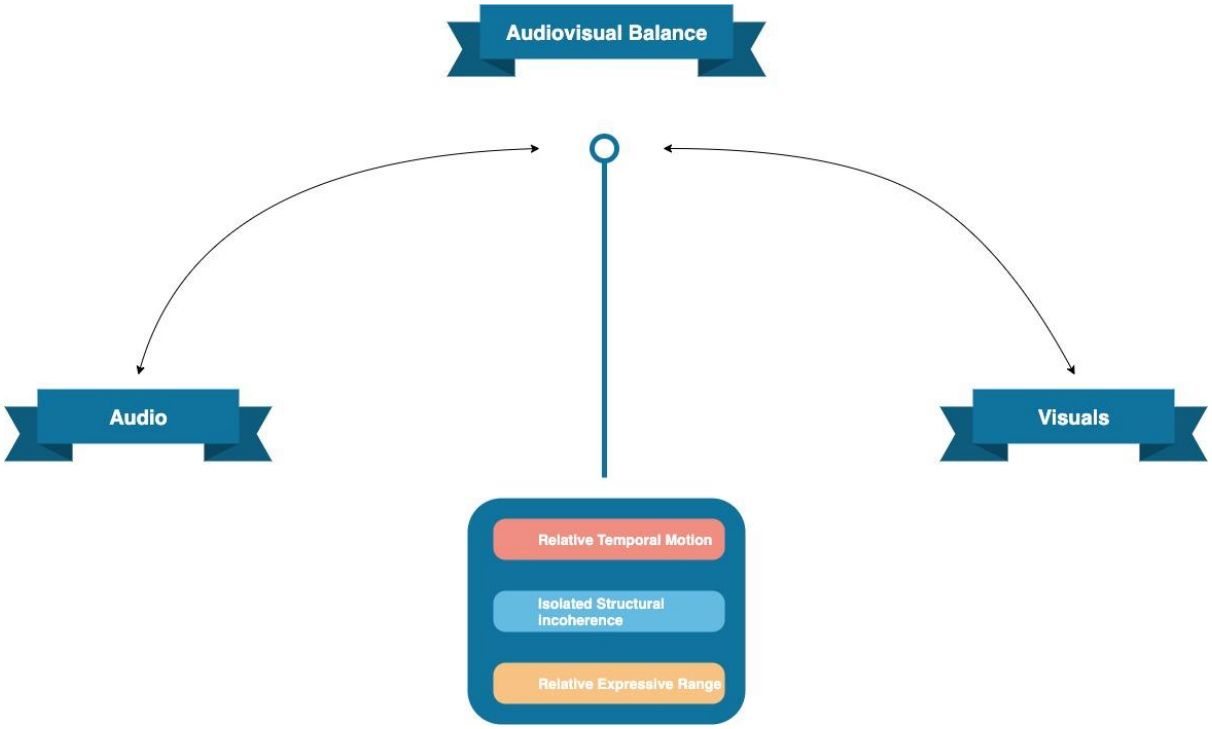

Figure 3.3 Audiovisual balance and proposed forces.

When analysing a piece through the lens of audiovisual balance, it is necessary to take a holistic approach to the experience of the piece. Also, a piece may be well-balanced in regard to relative-temporal-motion but not well-balanced in terms of relative-expressive-range. This would mean that the piece may seem unbalanced when viewed holistically.

During the discussion on relative-expressive-range, it was suggested that moving from areas of imbalance, to balance, could be an effective strategy for introducing moments of tension and release within a composition. In terms of isolated-structural-incoherence this could be achieved by introducing a regular beat in the audio material and then destroying it whilst at the same time aligning visual elements with non-rhythmic audio elements. In terms of relative-temporal-motion it

could mean varying the visual motion of textures and shapes to go in and out of sync with timbral, harmonic or rhythmic audio motion. This would necessitate independent control of the audio and visual processes, with the ability to map data and parameters freely between them.

Utilising a machine learning approach to controlling both audio and visual material could provide a useful way to achieve this. In the context of virtual environments, this approach may also provide an opportunity for the audiovisual composer to create fully immersive works, allowing the audio-spectator to explore a composition in real-time and to control all aspects of the audiovisual world they are placed in. The transition from screen-based to fully immersive environments is a significant undertaking both technically and theoretically. The concepts discussed here will be augmented by issues specific to the emerging medium of virtual reality. The next chapter will look at the theoretical issues facing the audiovisual composer when creating fully immersive work.

Chapter 4 Presence and Place

Virtual Reality (VR) is a multi-sensory medium, just as audiovisual art is a multi-sensory art form. The VR headset encloses the user in a visual and aural environment. This medium is inherently audiovisual and so it seems that it would be particularly suitable for abstract audiovisual work. Despite a relatively long history in technological terms (Slater 2009: 3549), VR is still an evolving medium. The technology has become more affordable and has also developed at quite a fast pace in the last decade. There are now a range of consumer VR systems such as those offered by Oculus[27], HTC[28], Valve[29], Hewlett Packard[30] and Pimax[31]. This means that these technologies are increasingly accessible to the audiovisual composer. This availability has led to a rise in the creation of VR art, examples of which will be discussed in this chapter. As the technology advances, there are increasing possibilities for the artist, not only in technical terms, but also in compositional terms. There are still many questions to be answered when creating VR experiences for artistic ends. Some of these questions were introduced in Chapter 1, and will be explored in more detail here and through the compositions presented in Chapters 8 and 9. This unexplored potential could provide a fertile area of practice for the audiovisual composer and is one of the main motivations for exploring this path.

4.1 Language of Immersion

VR has been defined as 'a high-end user-computer interface that involves real time simulation and interactions through multiple sensorial channels. These sensory modalities are visual, auditory, tactile, smell and taste' (Burdea and Coiffet 2003: 3). This definition doesn't necessarily imply the use of a head-mounted-display (HMD) and tracking system. The term immersive virtual reality (IVR) has been used to specify the context in which a HMD and tracking system is used (Slater 2009: 3549). Further to this, the term virtual environment (VE) has been used to describe environments that utilise precise tracking and high fidelity displays (Nilsson, Nordhal and Serafin 2016: 109). In the context of the work in this thesis, the acronym VR, will be used to discuss IVR systems that are used to immerse the user in a VE.

When working in a fully immersive context, there are a number of medium-specific issues that the audiovisual composer needs to be aware of. These include issues relating to:

[27] https://www.oculus.com/rift-s/ (accessed 12/09/2020).
[28] https://www.vive.com/uk/product/vive-cosmos-elite/overview/ (accessed 12/09/2020).
[29] https://store.steampowered.com/valveindex (accessed 12/09/2020).
[30] https://www.hp.com/us-en/vr/reverb-g2-vr-headset.html (accessed 29/09/2021).
[31] https://www.pimax.com/ (accessed 12/09/2020).

- Presence
- Locomotion
- Interaction
- Body representation

These issues are central to the experience of the VR user. They are characteristics of the medium of VR and as such need to be addressed by the audiovisual composer when creating work within that context. In order to explore these questions there needs to be some understanding of how they work. The following section will begin this task by unpacking the term, *immersion*.

4.1.1 Immersion and Presence

Immersion is a multifaceted term that is used in several fields, such as VR, computer games, film theory, music and literature (ibid. 2016: 109). This multidisciplinary interest in the concept has led to several interpretations that create uncertainty about what is precisely meant by the term (ibid. 2016: 108, Agrawal et al. 2020: 404). The ambiguity noted by Agrawal et al. hinders research and justifies their efforts at clarifying the term.

> To conduct research on immersive audiovisual experiences, there is a need to establish a clear definition of immersion. (Agrawal et al. 2020: 404)

This is similar to the argument for clarification relating to discourse surrounding audiovisual art that was put forward in Chapter 2 of this thesis. Agrawal et al. identify two main strands of thought when it comes to immersion; those that see immersion as an objective characteristic determined only by the system used to mediate the material, and those that see immersion as being predominantly defined as a psychological experience (Agrawal et al. 2020: 405). Servotte at al. (2020: 36) and Slater et al. (1996: 164 -165) have defined immersion as an objective description of the system that is used to envelop the user's senses. There are also different levels of immersion depending on the sophistication of the system and the number of senses that it affects (Slater 2009: 3550).

Taking a non-mediated perspective on immersion, Agrawal et al. (2020: 407) settle on the following definition:

> Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world.

By defining immersion in these terms they consciously reject the argument that immersion is dependent on the technology used. Rather, it is based on the user's psychological state. They divide this perspective on immersion into the three categories shown in Fig. 4.1.

Figure 4.1 Agrawal et al. (2020) conceptualisation of immersion.

These categories may or may not require a full VR system. They focus on the sense of immersion regardless of medium. Working in the field of computer games, McMahan (2003: 68) states that the 'most accepted definition of *immersion* is Janet Murray's'.

> Immersion is a metaphorical term derived from the physical experience of being submerged in water. We seek the same feeling from a psychologically immersive experience that we do from a plunge in the ocean or swimming pool: the sensation of being surrounded by a completely other reality, as different as water is from air, that takes over all of our attention, our whole perceptual apparatus. (Murray 1997: 98)

Murray's definition first refers to the original meaning of the word, to convey the feeling of being submerged in water. She takes this as a metaphor and applies it to a virtual world. The virtual world must give the participant the impression that they are completely surrounded by it, or submerged in it.

Crucially, Murray specifies that her interpretation of the term has to allow for active participation in the digital world. She states that 'in a participatory medium, immersion implies learning to swim' (Murray 1997: 99). This characteristic of immersion differentiates the experience from that of being immersed in music or a book. This idea of a participatory medium is particularly prescient in the area of VR, where Burdea and Coiffet posit that VR experiences are built on 'an integrated trio of immersion-interaction-imagination' (Burdea and Coiffet 2003: 4). They call this combination of concepts the 'three I's' of virtual reality (ibid.). This acknowledgement of active participation is also noted by Slater and Sanchez-Vives in their conscious use of the term 'participant' rather than 'user'.

They qualify this decision due to the fact that 'VR is different from other forms of human-computer interface since the human *participates in the virtual world rather than uses it*' (Slater and Sanchez-Vives 2016: 3). In Chapters 7 and 8, where the *ImmersAV* toolkit and immersive audiovisual composition, *Obj_#3*, are discussed, this terminology will be adopted and adapted to align with the audiovisual terminology of Chion. The prefix 'AV-*'* will be used before 'participant'. This is intended to acknowledge Chion's usage of the term, 'audio-viewer', to specifically describe those experiencing both audio and visual stimuli as a single unit (Chion 1994: 215-216). In this way the term *AV-participant* refers to an *audio-viewer* that is a *participant* in an immersive virtual environment.

This distinction between the immersion felt when engaged with a gripping narrative in a book, or a beautiful piece of music, and the immersion experienced in an interactive digital environment is important. Agrawal et al. (2020) avoid this distinction and abstract their definition even further than Murray, by omitting any mention of a participatory medium. This works for their context, providing a multidisciplinary definition that can be applied across the diverse fields of the arts and creative technology. The acknowledgement of participation re-focuses the term for a particular context. It describes participatory immersion as a central characteristic of digital environments such as computer games, or other virtual environments, experienced in a VR headset.

Witmer and Singer (1998: 227) consider immersion to be both a function of the participant's psychology and the system used to deliver sensory stimulus.

> *Immersion* is a psychological state characterized by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences.

Slater, Usoh and Steed (1995: 204) position their definition of *immersion* in an even more refined way than Witmer and Singer by focusing only on the technical system.

> We use "immersion" as a description of a technology, rather than as a psychological characterization of what the system supplies to the human participant.

In fact, Witmer and Singer (1998: 227) specifically distance themselves from this position when they state that they 'do not agree with Slater's view that immersion is an objective description of the VE technology'. This disagreement of terms would continue for several years, with Slater's (1999) response criticising their lack of separation between the concepts of immersion and presence. His notes on terminology (Slater 2003) attempt to provide further clarity. Witmer, Jerome and Singer (2005: 310) further criticise Slater's approach for discarding 'variables simply because they do not conform to our preconceived notions of immersion and presence'.

Slater et al. transfer the psychological characteristics of immersion over to the concept of presence. It seems that as the concept of immersion becomes more specialised within the field of VR, aspects of it morph into the concept of presence. When talking about immersion in a VR context, the concept of presence becomes extremely important. In fact Slater, Usoh and Steed (1995: 204) identify it as 'the central issue for virtual reality'. Witmer and Singer (1998: 225) also point out that the 'effectiveness of virtual environments (VEs) has often been linked to the sense of presence'.

The concept of telepresence describes a situation where a human is operating machinery from a remote location but is able to experience the sense of 'being there' (Minsky 1980). This definition has been adopted by VR researchers as a suitable description of the desired sensation of presence in virtual environments (Weech, Kenny and Barnett-Cowan 2019: 2, McCreery et al. 2013: 1635). It has also been defined as 'the subjective experience of being in one place or environment, even when one is situated in another' (Witmer and Singer 1998: 225).

Slater et al. (1996: 165) justify their separation of immersion and presence as a way 'to study the possible effects of the former on the latter', and that 'there may be other responses that are associated with immersion, independently of presence'. This is a criticism that Slater makes of Witmer and Singer's (1998) presence questionnaire. He states that he would not use it in his research as it 'does not give a measure of presence that is constructed independently from the factors that might influence it' (Slater 1999: 9). Slater (2003: 1) acknowledges the fact that these disagreements over terminology were 'hampering progress in the field'. Here, Slater again argues that the term, immersion, should 'stand simply for what the technology delivers from an objective point of view' (ibid.). Slater et al. (1996: 165) put forward that an immersive system is assessed based on the objective attributes of its displays. These attributes are shown in Fig. 4.2.

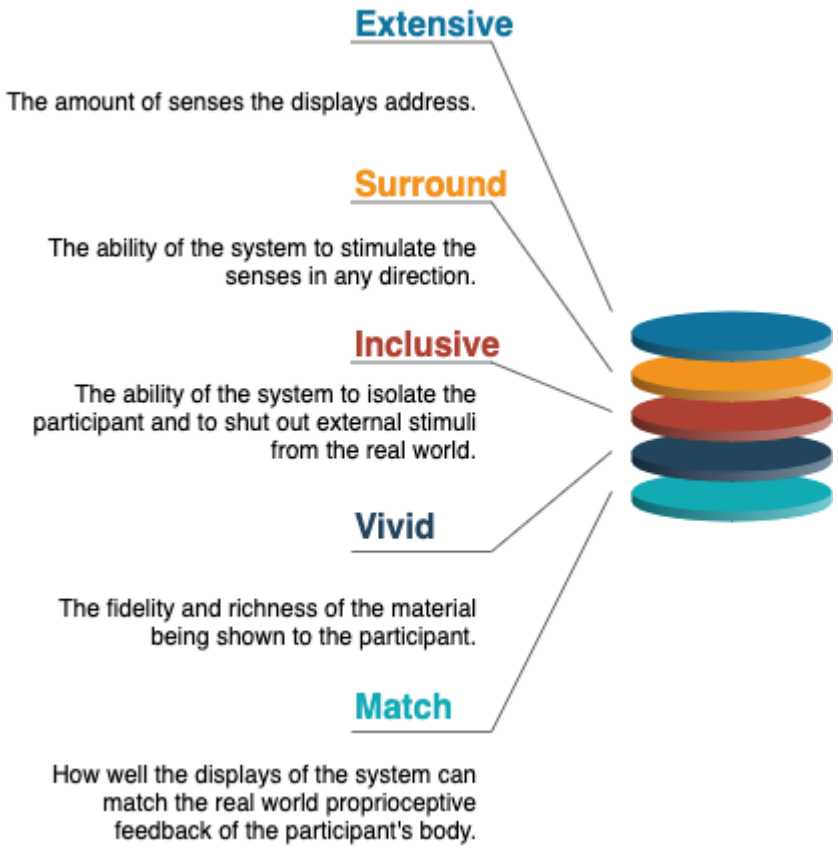

Figure 4.2 The criteria for immersive systems based on Slater et al. (1996).

These are the standards by which a system can be assessed for its level of immersion. The empirical relationship between the immersive capabilities of a system and the resulting sensation of presence are 'probably strongly related' (Slater 2003: 2). This intimate connection between immersion and presence is also acknowledged by Witmer and Singer (1998: 227) when they state that a 'VE that produces a greater sense of immersion will produce higher levels of presence'. Although this is a claim that Slater may object to on terminological grounds, it does highlight a consensus that immersion and presence are in some way linked.

In order to clarify the separation of presence from immersion, Slater offers an analogy to highlight the difference between the subjective perception of colours on one hand, and the empirical measurement of those colours on the other hand. Slater points out that although colours can 'be described objectively in terms of a wavelength distribution' (Slater 2003: 1), those colours may be perceived differently depending on the physiology of the person in question. He highlights the existence of 'metamers, where objectively different wavelength distributions are perceived as the same colour by human observers' (ibid.). In terms of the colour spectrum, there are separate constructs for objective measurement and subjective perception. The same is true for VR, in that presence 'is a

human reaction to immersion' (ibid. 2003: 2). In other words, presence is the perception of immersion.

He further clarifies the position of presence by distinguishing it from concepts such as *involvement* and *emotion.* Presence is concerned with what Slater calls the *form* of the system rather than the *content* (ibid.*)*. He gives an example of listening to a recording of a classical performance. The listener may feel like they are in the concert hall due to the reverb characteristics of the sound, and the frequency response of the headphones. This would be a perception of presence. However, they may then become bored because they don't like the piece of music being played. This has nothing to do with presence. The listener still feels present whether or not they like the music. The feeling of being in the concert hall is related to the form of the system allowing them to feel the perception of presence. The value judgement about the music is related to the actual content of the material being presented through the system. The concepts of *involvement* and *emotion* relate to content rather than form (see Fig. 4.3). A participant may be present but not involved or interested in what is happening. Conversely, a participant may be involved but not present. Here, Slater gives the example of watching a TV show or reading a book. In the case of a book, he states that it is 'at a certain low level of immersive "technology", and maybe can induce presence in some people' (ibid.*)*. Similarly, the participant may be present in a situation but experience different emotions in the place depending on what is happening and where their focus lies. Therefore, they are conceptually separated from *presence.*

> Presence is a response. Separate from presence are aspects of an experience such as involvement, interest and emotion. (Ibid. 2003: 4)

When engaging in a virtual experience, a sign that someone is present is when they behave 'in a way that is similar to what their behaviour would have been in a similar real life situation' (ibid. 2003: 3). Further, in the case of virtual experiences that are not based on reality, we are 'able to explore what presence would be like if such worlds existed' (ibid. 2003: 4). This concept of exploring what presence would be like in non-realistic environments, or 'unrealities' (Slater and Sanchez-Vives 2016), will play an important part in the exploration of immersive audiovisual composition discussed in Chapter 7.

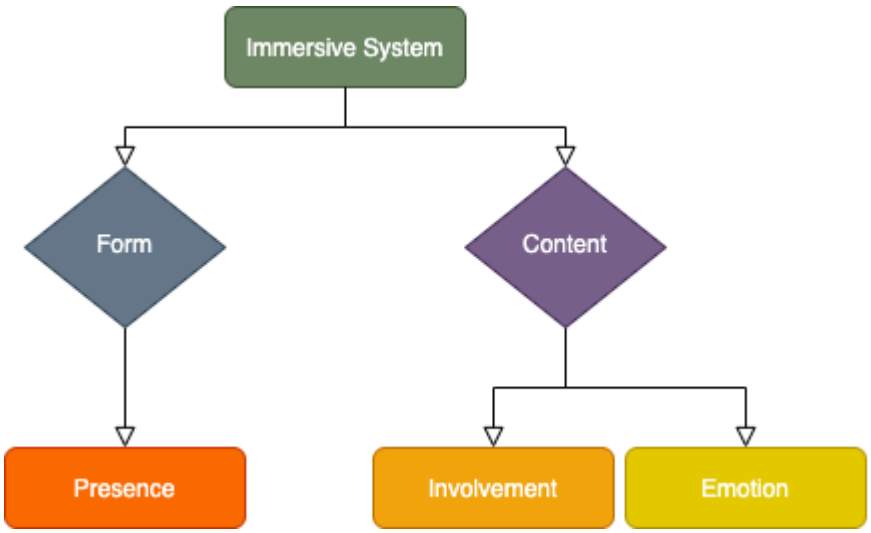

Figure 4.3 The relationship between feelings of presence, involvement and emotion.

4.1.2 Place and Plausibility

Motivated by the confusion surrounding the conceptualisation of *presence*, Slater (2009) introduced the *response-as-if-real* (RAIR) framework to describe a VR system and its potential for creating presence-like sensations. This framework is based on the elements illustrated in Fig. 4.4.

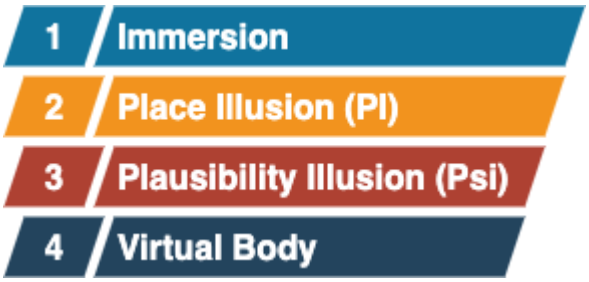

Figure 4.4 Elements of the RAIR framework.

Taking Slater's interpretation, immersion is defined by the capabilities of the system hardware. He then introduces a new concept to differentiate between immersive systems. He characterises systems based on 'the sensorimotor contingencies (SCs) that they support' (Slater 2009: 3550). He describes these SCs as 'the actions that we know to carry out in order to perceive' and defines the set of actions that are possible in a given VE as the 'set of *valid actions'* (ibid.*)*. Immersive systems can then be assessed based on the amount of valid actions they allow. The highest level of immersive system is a '*first-order system*' that is capable of supporting a range of SCs that 'approximate reality' (ibid. 2009: 3551). He then defines a '*second-order system'* as 'one that has valid actions as a proper subset of a first-order system, and so on for lower orders' (ibid.*)*. A fundamental difference between system levels is that a higher order system would be able to completely simulate a lower order system.

> So a system that supported being able to perceive using the whole body (bending down to look underneath something, reaching out, looking around an object, etc.) would be at a higher level of immersion than one that just afforded looking at a screen (for as soon as you turn your head away from the screen you are no longer perceiving the virtual world). (Slater 2018: 432)

Slater (2009) also introduces two new terms to use as further elements of *presence.* They are *place illusion* (PI) and *plausibility illusion* (Psi). The definitions are shown in Fig. 4.5.

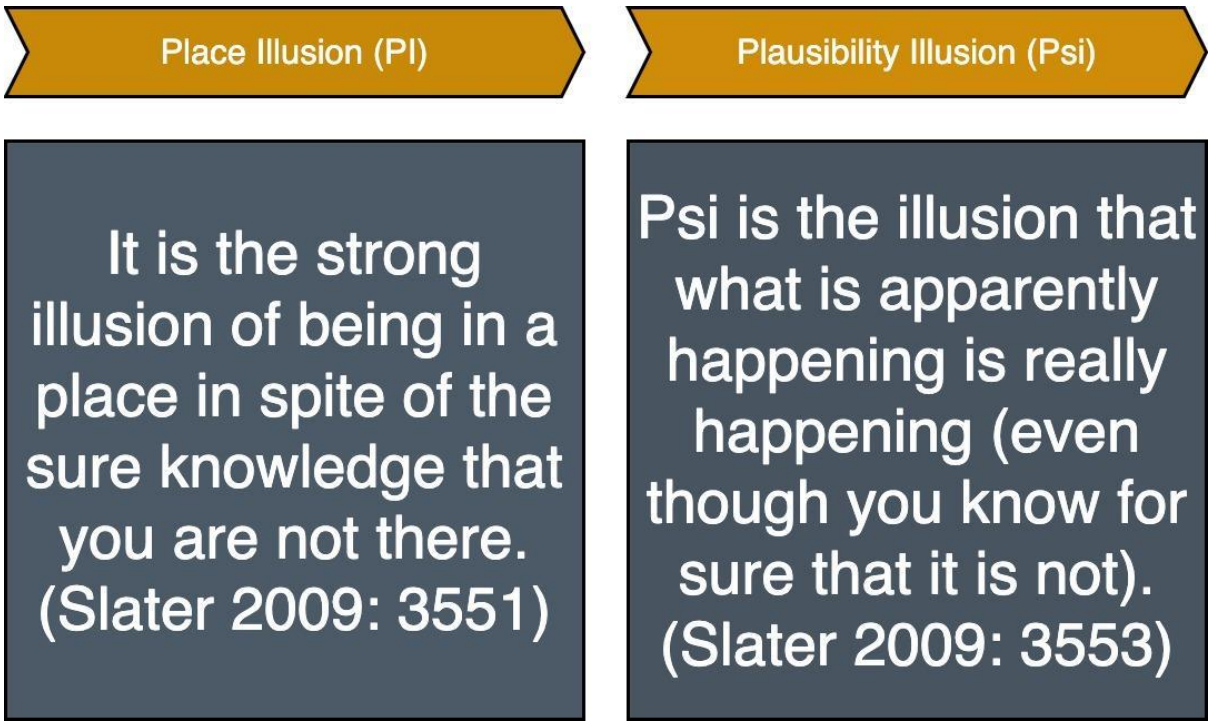

Figure 4.5 PI and Psi definitions.

The final element to Slater's framework is the inclusion of a virtual body in the VE. He identifies the body as 'the focal point where PI and Psi are fused' (ibid. 2009: 3554). He states that the simple act of 'looking down at your own body provides very powerful evidence of PI' (ibid.).

Marco Gillies (2016) recognises the importance of Slater's account of immersive systems and elements of presence. However, he also acknowledges that 'sensorimotor contingencies can break down when they are too complex to model and recognise easily in code (or are not feasible for other reasons, like direct walking)' (Gillies 2016: 4). Schirm, Tullius and Habgood (2019: 672) describe the conceptualisation of PI and Psi as being 'highly influential'. Nilsson, Nordhal and Serafin (2016: 130) acknowledge that there are 'large differences between existing views of what characterizes and causes presence'. They identify large overlaps between certain definitions of immersion, including Witmer and Singer's (1998) definition, and Slater's definition of PI. Due to this they advise that 'it may not be particularly fruitful to rely on these definitions of immersion' (Nilsson, Nordhal and

Serafin 2016: 126). Ultimately, the consensus in this thesis is that the concepts of PI and Psi, in particular, would be helpful for trying to create audiovisual compositions in a VE. These concepts will be discussed in the context of immersive audiovisual composition later in the chapter.

4.2 Contemporary work

The increased availability of immersive systems has led to an increase in the amount of artistic work being produced for the medium. One of the core characteristics of fully immersive VR is the ability to interact with the virtual world and objects within it. *Mutator VR: Vortex*[32] is an immersive experience that 'gives the viewer a direct sense of presence through cause and effect by attracting or repelling' procedurally generated 'swirling organic agents' (Putnam, Latham and Todd 2017: 139). The motivation behind this experience is 'to investigate unified approaches to defining the sounds, geometry, and interactive dynamics of a world populated with agents' (ibid.). The main goal of the interaction is 'that it be emergent without predetermined goals so that the user has more opportunity to construct their own experience' (ibid.). Latham et al. (2021: 277) further elaborate on these interaction principles by explaining that even in the event that the participant does not follow the given instructions, they inherently avoid confusion in a tightly-timed exhibition context. This is achieved by 'ensuring that (a) each interaction has discoverable consequences and (b) as far as possible there is a natural (kinaesthetic) correspondence between cause and effect'. The unified approach to graphics and sound generation in *Mutator VR* clearly shows a desire to create a strong audiovisual experience.

Ox and Britton (2000) approached the development of the *Virtual Color Organ* with specific *intermedial* ambitions. Intermedia is a term, coined by Dick Higgins in 1966 (Higgins and Higgins 2001), that underpins the philosophy of the *Fluxus* art movement. Ox and Britton (2000: 2) differentiate *intermedia* from *multimedia*.

> With multimedia, content/information is presented in more than one medium simultaneously. However, intermedia is a combinatory structure of syntactical elements that come from more than one medium but combine into one. The final form can only be seen after going through the entire process.

The *Virtual Color Organ* was created to be experienced in a range of virtual environments such as the *CAVE* (Cruz-Neira, Sandin and DeFanti 1993), the *ImmersaDesk* (Czernuszenko et al. 1997) and the *VisionDome* (Colucci et al. 1999). These environments don't require the use of a head mounted display (HMD) and were more popular before mass production of HMD-based systems was possible.

[32] https://mutatorvr.co.uk/ (accessed 08/04/2022).

The *Virtual Color Organ* maps structural characteristics of pre-existing musical compositions to visual textures and colours in 'a metaphorical way' (Ox and Britton 2000: 3). This mapping scheme was developed by Jack Ox and emerged out of 'a 20-year history of visualizing music by devising systems of equivalences that "translate" organized collections of data, gleaned from preexisting compositions' (Ox and Britton 2000: 2). The *Virtual Color Organ* is a more recent incarnation of the colour organs that have been produced by artists and inventors for hundreds of years. Ox's practice is firmly situated within the tradition of visual music, and the *Virtual Color Organ* is a link between that tradition and generative immersive audiovisual practice such as *Mutator VR*.

Another important characteristic of artistic work in VR is the ability to create environments and experiences that would be impossible in the real world. The exploration of non-euclidean space is a compelling possibility with VR (Hart et al. 2017a, 2017b; Coulon et al. 2020a, 2020b). Hart et al. (2017a: 33), through their *Hypernom*[33] work, point out that these 'spaces are still seen as unintuitive and exotic', but they 'believe that with direct immersive experience we can get a better "feel" for them'. Their mathematically accurate exploration of non-euclidean spaces demonstrates the powerful potential for VR to create completely impossible worlds that are entirely divorced from reality.

However, for an engaging artistic experience it has been posited that there needs to be a balance between the real and unreal. Latham et al. (2021: 276) have found that too 'real an experience is artistically boring, too unreal an experience leaves users uninterested and disoriented'. They place their *Mutator VR* project within the lineage of surrealist artworks. This contextualisation is grounded in the 'successful mixing' of 'real and unreal elements in the same scene', the same technique 'used in the surrealist paintings of Dali, Magritte and Max Ernst' (ibid.).

There are exciting possibilities for 'music-led virtual reality experiences' outlined by Buckley and Carlson (2019). VR opens up possibilities for musical compositions that are not limited by real-world practicalities such as 'sound systems, the number of musicians available to them, and environmental acoustic considerations' (ibid. 2019: 1497). Apart from technical affordances, there are extended aesthetic possibilities in immersive environments for repurposing 20th century composition techniques such as the 'event scores of Fluxus artists' (ibid. 2019: 1499), and the aleatoric approach of Witold Lutoslawski in his piece *Venetian Games* (1961). Through his creative notation he granted the 'performers the ability to control speed and exact timing' (ibid.). In a VR context, the performer would be the AV-participant, in control of aspects of the environment, whilst the composer is

[33] http://h3.hypernom.com/ (accessed 12/09/2020).

responsible for creating the larger world. Buckley and Carlson draw on the philosophy of embodiment, design-goals including discoverability and enchantment, the utilisation of the 'techno-somatic dimension' (ibid. 2019: 1498) and concepts of digital instrument design, to propose a framework for musical composition in immersive environments. Their identification of the importance for the AV-participant to experience embodiment through 'hearing and being heard' feeds into the sensation of presence that will be explored in Chapter 8.

4.3 Compositional Challenges

The transition from a lower-order immersive system to a higher-order immersive system, using Slater's terminology, poses some interesting problems for the audiovisual composer. Some insight had been gained into these challenges during the composition of *Ventriloquy II*. The system on which that piece was presented, was of a higher order, in terms of immersion, than the system used to perform *Ventriloquy I*. See Fig. 4.6 for a graphical representation of this.

Figure 4.6 Ventriloquy I vs. Ventriloquy II system comparison.

The spatial capabilities, of both the audio and visual displays, were increased, going from *Ventriloquy I* to *Ventriloquy II*. The focal-point of the single screen was expanded to six screens arranged all around the audience. The audio system was expanded from stereo to 12.2 surround sound. The expansion of the system introduced a spatial element to the composition. The relative security of the single screen focal point was gone. The stereo safety net was also gone. 360 degree space was suddenly a central part of the composition. As will be discussed in Chapter 6, the experience of creating work for that context provided a learning opportunity in the practical implications of creating immersive audiovisual art.

4.3.1 Presence and Abstract Environments

As discussed above, the concept of presence is a key element within immersive environments. To a certain extent this defines the medium of VR. It therefore must be a primary concern for the portfolio compositions presented in this thesis.

Presence is spoken about in terms of being a sensation. A sense of presence is sometimes felt when the AV-participant is immersed in a virtual world. Taking this conceptualisation of presence, when working in an immersive context, the audiovisual composer is trying to integrate a third, embodied sense into a traditionally bi-sensory practice. Instead of arranging visual and sonic elements to engage the senses of sight and sound, the composer must now arrange visual and sonic elements to engage the senses of sight, sound and presence.

Consider PI and Psi, the components of presence identified by Slater. Slater states that PI should 'be treated as binary - it is a qualia associated with an illusion. Either you get the illusion or you do not - you cannot partially get an illusion' (Slater 2009: 3554). Bergstrom et al. (2017: 1338) argue that 'the problem of attaining presence as Place Illusion has a solution with broad outlines known'. They further explain that 'the more that "real world" sensorimotor contingencies are afforded in VR the greater the likelihood that' (ibid.) presence will be achieved through PI. They suggest that the real area of interest now is in achieving presence through Psi.

Slater (2009: 3553) identifies several factors that may influence Psi. Firstly he states that 'events in the virtual environment' may affect Psi when they 'refer directly to you' even though you have 'no direct control' over them. He also identifies realistic elements of the environment as possible factors. Bergstrom et al. (2017: 1336) present these as three separate factors. These factors are shown in Fig. 4.7.

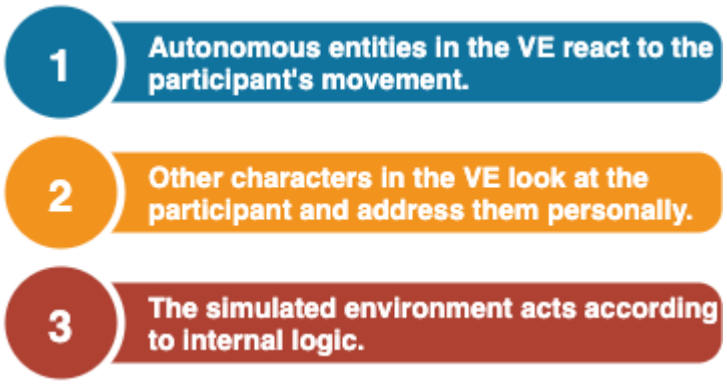

Figure 4.7 Elements of Psi.

The third point here is especially relevant to the work in this thesis. Bergstrom et al. identify the 'field of entertainment and fantasy' (ibid.) as a major context within which VR is being explored. This field relies on unrealistic settings in which they place the participant. However, even though these settings are not based on the real world, they still adhere to the internal logic of that particular world. This seems like a salient point for abstract audiovisual composition. It would suggest that the audiovisual composer should try to create an environment that is abstract yet still conveys its own sense of realism. This point will be explored in the piece *Obj_#3* in Chapter 8.

4.3.2 The Immersive Contract

The relationship between sight and sound was discussed in terms of audiovisual balance in Chapter 3. It was posited that when the audio and visuals are in a balanced state the audio-viewer is more likely to experience added-value. The effect of the sense of presence, on audiovisual material within a VE, will be explored below.

Regarding audiovisual balance, it seems intuitive that a participant who does not feel present in a VE will be less engaged with the content of the VE. Lack of presence means the participant is aware of the real world outside of the VE. This may cause conflict in the senses between the two realities. Slater and Steed (2000: 417) note that the 'issue of presence becomes interesting only when there are competing environments'. This includes both external and internal environments. They further note that the sensation of presence then determines which environment 'the individual responds to and acts within at any given moment' (ibid.).

This suggests that if the participant is not experiencing the sensation of presence within the VE at a particular moment, they will be less-able to perceive added-value or correspondences between the audio and visual material. Therefore, the audio-visual contract must take into account the sense of presence. Chion's audio-visual contract requires the audience to cognitively accept that what they hear and see in the cinema (or whatever venue they are watching the film) is one, unified construct, even though it is not.

> The audiovisual relationship is not natural but rather a sort of symbolic pact to which the audio-spectator agrees when she or he considers the elements of sound and image to be participating in one and the same entity or world. (Chion 1994: 222, note 6)

Considering presence in this definition, the affected variable is the *world* in which the elements of sound and image are participating. This world is no longer the cinema theatre or audiovisual performance venue. It is a virtual world that demands a further suspension of

disbelief from the participant. The participant may have to enter into an initial contract agreeing to accept the VE as real. Only after accepting this, and experiencing the sensation of presence, can they fully experience the forces at work in the piece.

Once the AV-participant has accepted the VE as a plausible reality, there is an extra axis of balance to be considered between the *real* and the *unreal* as identified above. This new axis doesn't relate to a balance between audio and visual material as such. Rather it is a balance that needs to be struck within the environment as a single entity. This concept is discussed further in Chapter 8.

4.4 Conclusion

This chapter introduced concepts particular to immersive virtual environments. It established the concept of presence as a particularly important element in the language of immersion. The exploration of the sensation of presence was identified as a major concern for the audiovisual composer working in an immersive context. Now, the audiovisual composer has the power to affect the sense of presence in addition to the senses of sight and sound. This could be a fundamental concern for audiovisual composers working in immersive contexts going forward.

An additional axis of balance was identified that has the potential to guide the composer in creating immersive work. This axis highlights the importance of balancing the experience such that it is real enough to allow for a sense of presence, and unreal enough to maintain interest and a sense of 'enchantment' (Buckley and Carlson 2019: 1498). This additional axis can potentially support the audiovisual balance framework by reinforcing interest in the environment as a whole, which will then allow the AV-participant to remain engaged with the material.

Chapter 5 Neural Audiovisual Mapping

The previous two chapters dealt with theories of audiovisual composition. The concept of equality as it relates to audio and visual material was explored in Chapter 3. The implications of creating audiovisual compositions in the emerging medium of VR were then discussed in Chapter 4. These discussions laid the theoretical foundations and artistic goals for the development of the software and artworks in the rest of the thesis. This chapter presents an approach to mapping real-time input data to audio and visual parameters simultaneously which results in a novel control paradigm for audiovisual compositions.

The chapter will open with a discussion around mapping in audiovisual systems. Interactive machine learning (IML), will then be proposed as a way to quickly build a mapping layer between real-time input data and output audiovisual parameter data. It is hoped that an audiovisual performance tool built using this approach could be an appropriate way to create work that is audiovisually balanced. A bespoke system was developed to explore this approach to audiovisual mapping. This is called the Neural AV Mapper and will be discussed alongside four compositional studies that were created and performed using the system. The source code for the software presented in this chapter is included in the media pack in the following directory: *sourceCode/neuralAVMapper*. The source code can also be found at the GitHub repository[34]. Video recordings of the studies are included in the media pack in the following directory: *mediaFiles/iml_studies*. These videos can also be found on YouTube and are linked throughout the text.

5.1 Mapping Terminology

Parameter mapping is an important aspect of generative audiovisual art, as the very nature of the artform involves the manipulation and interaction of material perceived across separate modalities. Data or information is often mapped from one medium to another in order to create close structural bonds. In order to contextualise the practical work presented later in this chapter, some ideas and terminology related to the issue of mapping within audiovisual art will be discussed.

Stephen Callear has identified a taxonomy of terms drawn from ‘the fields of instrument design, algorithmic composition and audiovisual art’ (Callear 2012: 26). Here we again see the audiovisual metadiscipline drawing influence from related fields. Callear identified three classifications that are useful in discussing parameter mapping. These include mapping types (Fig. 5.1), mapping hierarchy

[34] https://github.com/bDunph/neuralAvMapper (accessed 14/04/2022).

(Fig. 5.2) and mapping perception (Fig. 5.3). These concepts have been grouped into graphical charts to aid the understanding of the conceptual area.

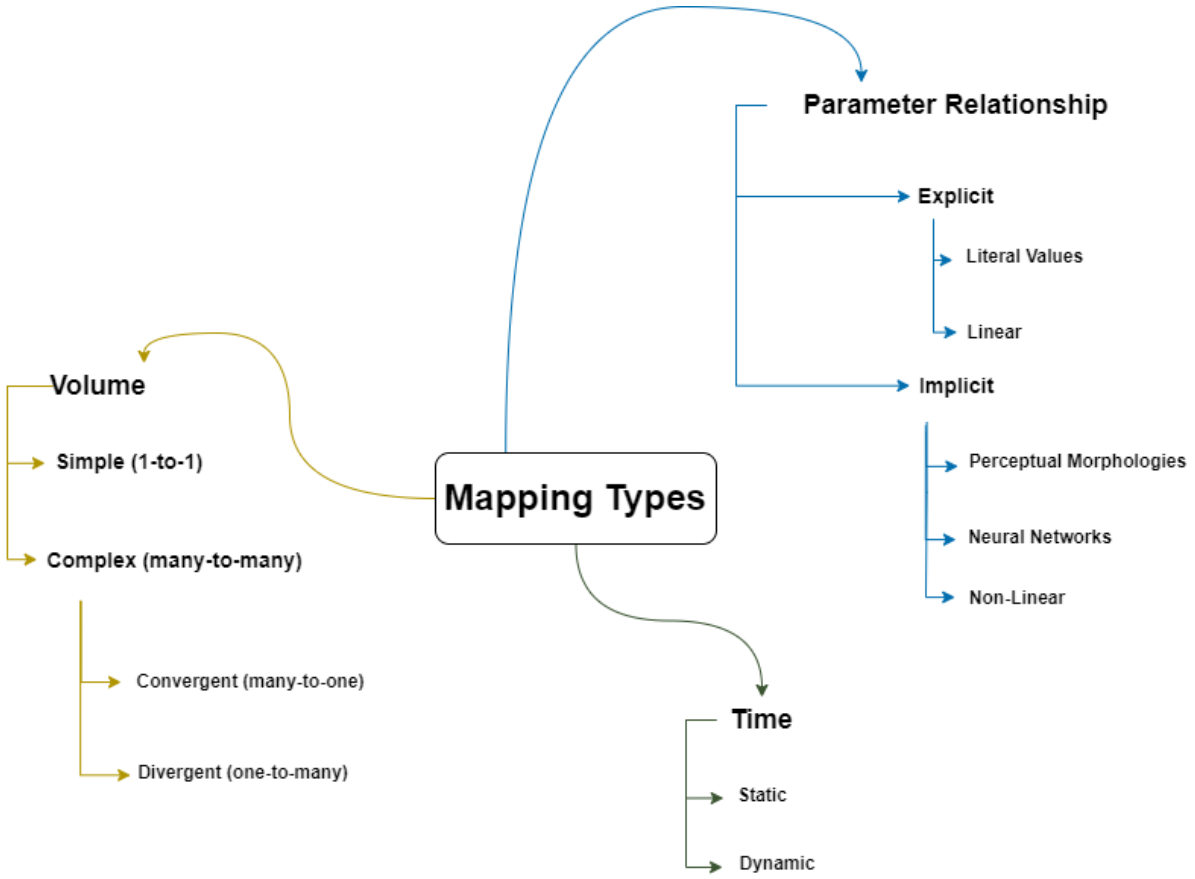

Figure 5.1 Mapping types within audiovisual art.

As shown in Fig. 5.1, mappings can be described in terms of; the relationship between input and output parameters, time variance and the volume of parameters. These are not exclusive in that a mapping could be described using all of these categories. For example, a mapping of pitch to height on screen could be described as simple, explicit and static. Where the mapping is based on the analysis of the input/output relationship, Hunt, Wanderly and Kirk (2000: 209) identify two approaches. The first is an explicit approach, where the relationships between the parameters are clearly defined. The second is a generative approach, which makes use of a 'method that provides a mapping strategy by means of internal adaptations of the system through training'. These distinctions are further developed by Arfib et al. (2002: 130) to identify explicit mappings as those that 'exactly describe the links between the input and output mapping parameters, thanks to mathematical formulae', whereas implicit mapping strategies 'define behaviour rules but not precise value rules'. Relevant to the work presented below, Arfib et al. further state that the use of artificial neural networks would be described as an implicit mapping strategy (Arfib et al. 2002: 131). Callear expands on the idea of the implicit strategy within audiovisual composition by stating that the 'perception of change in one parameter

can be used to inform relative change in another without the need for a direct translation of actual values' (Callear 2012: 26). He further states that an implicit approach to mapping can involve transference of 'perceived morphologies' between media (ibid. 2012: 31). This is indicated in Fig. 5.1 under the label *perceptual morphologies.*

Mappings between input and output parameters can also be described as linear or nonlinear. A linear mapping is a direct mapping of one input parameter to an output parameter. Doornbusch (2002: 146) cites the mapping of gas velocities to violin glissandi in Xenakis's *Pithoprakta* (1955-56) as an example of a 'linear and direct' mapping. A nonlinear mapping, on the other hand, is the indirect mapping of one or more input parameters to one or more outputs. Hunt and Kirk give the violin as an acoustic example of nonlinear mapping, as they describe how the combination of input parameters such as finger position and bow pressure affect multiple output parameters such as amplitude, pitch and timbre.

> The total effect of all these convergent and divergent mappings, with various weighting and biasing, is to make a traditional acoustic instrument into a highly non-linear device. (Hunt and Kirk 2000: 235)

Mappings can also be described with reference to their state over time. Momeni and Henri (2006: 49) describe an approach to the concurrent control of audio and video material using 'an independent layer of algorithms with time-varying behaviour that is affected, explored, or observed in some way with gesture'. As an example of this they describe the manipulation of a three dimensional parameter space with only two inputs. Here, a static mapping strategy would only allow the manipulation of 'one two-dimensional manifold in the three-dimensional parameter space of our instrument' (Momeni and Henri 2006: 56). However, in order to capitalise on the entire three dimensional space, the 'two-dimensional manifold' could change over time according to some 'internal algorithm that is intuitive and controllable' (ibid.). They call this a dynamic mapping layer. According to Arfib et al. (2002: 131), dynamism can be interpreted as the 'the ability of the mapping to evolve in time, to learn from the input data over time', or the use 'of dynamic description parameters for gestures'. If this time variance is not a feature of the mapping strategy, it would be classed as static, because the 'relationship between input and output parameters remains constant' (Callear 2012: 26).

In Fig. 5.1, the interpretation of linear/nonlinear and static/dynamic mappings diverges somewhat from Callear's. He states that these terms 'can be used interchangeably as descriptors for the time variant behaviour of a mapping strategy' (ibid. 2012: 27). This statement is only partially correct, in that a dynamic mapping is always nonlinear. However, it is possible to have a nonlinear mapping that

is not time-variant, thus indicating that the terms are not completely interchangeable. As will be seen below, a neural network performing regression analysis would be an example of this.

Mapping terminology can also relate to the number of parameters the mapping strategy engages with. This can mean a single input can be mapped to a single output (one-to-one), a single input can be mapped to many outputs (one-to-many), many inputs can be mapped to a single output (many-to-one) or many inputs can be mapped to many outputs (many-to-many). According to Callear, a one-to-one mapping can be referred to as a 'simple' mapping whilst the other can be referred to as 'complex' (ibid.). Complex mappings such as one-to-many and many-to-one can also be described as divergent and convergent respectively. According to Rovan et al. (1997: 69), a divergent mapping is such that 'one gestural output is used to control more than one simultaneous musical parameter', whereas a convergent mapping is such that 'many gestures are coupled to produce one musical parameter'.

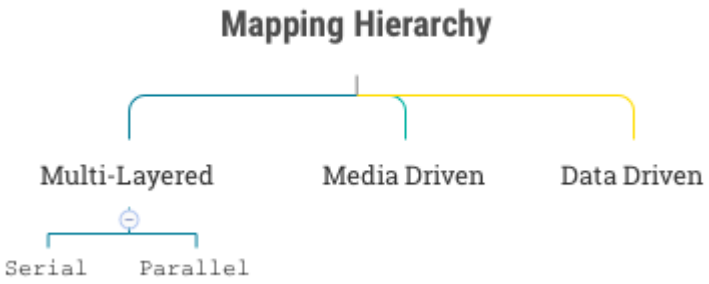

Figure 5.2 Mapping hierarchy within audiovisual art.

According to Callear (2012: 27) 'the term layer is used to describe an independent mapping system within the context of a parameter mapping strategy'. He goes on to explain that an audiovisual system may contain multiple mapping layers which can be arranged in a serial or parallel configuration. In a serial configuration, the data would be sent from input to output, resulting in a 'cumulative effect of multiple mapping layers', whereas 'a parallel mapping strategy allows the dissemination of data from a single input set to multiple output parameter sets' (Callear 2012: 27-28). These layers can be independently implemented to function as any of the mapping types discussed above.

Callear (2012: 28) has identified two types of audiovisual system, namely 'media driven or data driven'. Media driven audiovisual systems derive either the audio or the visual content from the complementary modality. For example, the movement and appearance of the visual elements might be derived from analysis of audio features such as frequency, timbre or transients. In this situation the visuals could be described as audio-reactive. Similarly, audio features may be dictated by visual characteristics such as shape, hue and on-screen coordinates. The audio here could be described as a

sonification of the visual parameters. As discussed in Chapter 3, Callear states that the relationship between the audio and visual material in this type of media-driven system is hierarchical. He states that the dominant medium is the one whose data drives the parameters of the dependent medium (Callear 2012: 31). When viewed through the lens of audiovisual balance and media equality, this could create an unbalanced relationship within the architecture of the system. In contrast, a data driven audiovisual system enables the 'concurrent synthesis of aural and visual media derived from a single set of control parameters' (ibid. 2012: 36). Callear further states that the control parameters can be 'derived from any external data-set from which a temporal morphology can be extracted' (ibid.). Perhaps an approach to the composition of an audiovisual work in this manner may help, in the construction at least, to avoid a hierarchical relationship between the media, as defined above by Callear. However, it should be noted that this approach would still not guarantee a non-hierarchical perceptual result.

In a theoretical sense, the idea of a purely data-driven approach is conceptually attractive when thinking about the equality of the material. In terms of the relative perceptual motion of the material, the source of the kinetic energy of the piece would be centralised, driving both the audio and visual material simultaneously. This concept was encountered in Chapter 3 with the discussion of Ryo Ikeshiro's technique of audiovisualisation. Unfortunately, in practice, there is a danger that the central data source will not translate well to both audio and visual material. The raw expressive material may not exist sufficiently in the data alone. Battey (2020: 272) states that:

> Audiovisualisation can be both a fascinating and problematic prospect - arguably something likely to prove difficult to achieve to a convincing standard, given the differences between visual and musical perception. How and to what degree and depth can music and image cohere if arising from a single underlying abstraction that is neither musical nor visual in its essence, nor perceptually informed?

In *Estuaries 3* (2018), Battey tackles the issue of perceptual coherence by manually arranging material, after it has been produced by his generative system, according to his own artistic perception. In this way he calls it an 'audiovisualisation-assisted' composition (Battey 2020: 273). Perhaps a hybrid approach such as this would give the artist more control over the relative expressive range of the material, ensuring that it is well-balanced throughout the work.

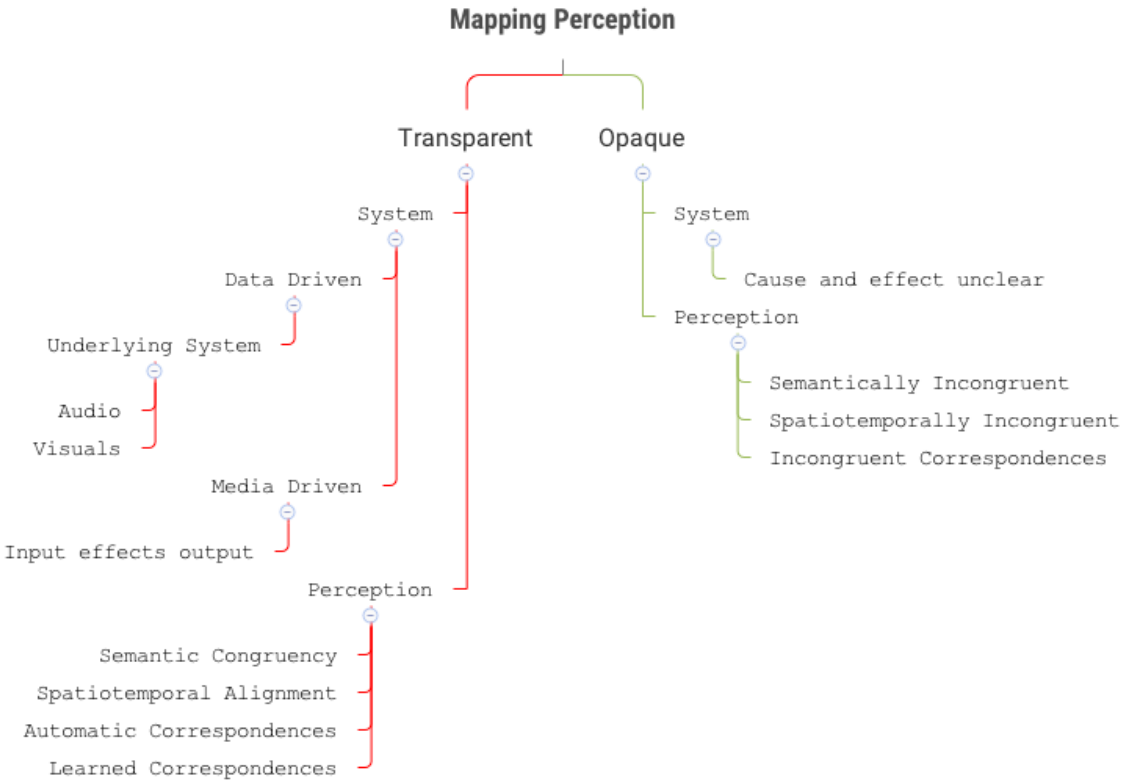

Figure 5.3 Mapping perception from the audio-spectator's perspective.

Mappings can be described as transparent or opaque (Fig. 5.3). Within an audiovisual system, a transparent mapping is one where the 'relationship between input and output parameters is clearly perceivable' (Callear 2012: 29). In a media driven piece, this describes the ability to clearly perceive the cause and effect between dominant and dependent media. As an example, consider a transient sound making a circle's circumference momentarily larger. Here it is obvious that the circle is reacting to the beat. This type of mapping can create solid connections within the perception of the audience, but with overuse, can become tiresome due to its simplicity. The use of this type of mapping has also been criticised as being 'superficial' (Dannenberg 2005: 28) as it offers 'only what is readily apparent in the music itself' (ibid.). However, within a data-driven work, a transparent mapping 'specifies the perceptual strength of correspondences between aural and visual events and also between input data and the resultant medium structures' (Callear 2012: 29). In some cases, this may mean that the audience can clearly perceive the underlying system and how it is driving both the audio and visuals.

Within generative audiovisual work, this is actually desirable as some artists, such as Ryo Ikeshiro, aim to reveal details of the underlying system through audiovisualisation. At the technical level, an opaque mapping obfuscates the cause and effect between the modalities leading to a ‘relationship between input and output parameters’ that is ‘imperceptible' (ibid.).

Callear (ibid.) states that a balance needs to be struck within an audiovisual piece between opacity

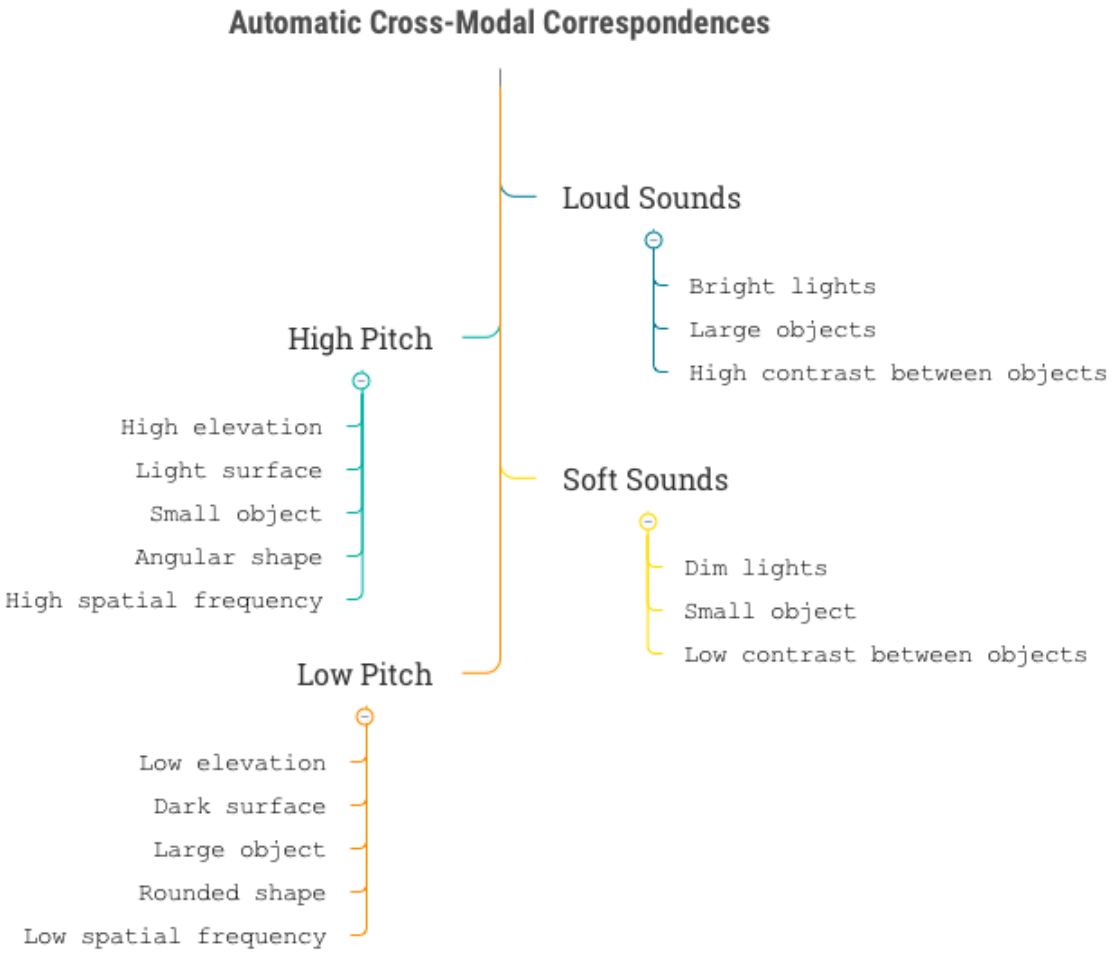

Figure 5.4 Automatic cross-modal correspondences.

and transparency, as using only opaque mappings would ‘be inappropriate for compositions that seek to define an interpretable correspondence between mediums’. This is reinforced by Sá, Caramieux and Tanaka (2014: 86) when they state that ‘clearly perceivable cause-effect relationships’ are ‘problematic for the music’, in that they create a perceptual hierarchy that elevates the visual and subordinates the audio. When viewed through the lens of audiovisual balance, this would not be desirable. If the mappings within an audiovisual work were completely opaque, there is a chance that the audience will perceive the audio and visuals as separate entities, thereby weakening the audiovisuality of the work. This would suggest that a balance between opaque and transparent mappings would be desirable. In order to address this issue, Callear (2012: 29) finds that an

‘audiovisual composition is best served by parameter mapping strategies that exhibit temporal variance in mapping transparency’. He further states that the temporal manipulation of mappings in such a way can be seen to be analogous to ‘the use of consonance and dissonance in tonal music’ (ibid.). This could be interpreted as a possible way to create tension and release within an audiovisual work. This desire to manipulate tension and release is also found in Evans’s (2005) approach to composing visual music. Sá, Caramieux and Tanaka suggest implementing a fungible mapping strategy, that produces ‘a sense of causation’, whilst at the same time managing to ‘confound the actual cause-effect relationship’ (Sá, Caramieux and Tanaka 2014: 85). This has the benefit of giving the audio-spectator enough transparency to bind the material perceptually whilst maintaining enough opacity to keep the interactions interesting.

Some audiovisual objects exhibit features that allow them to be automatically bound in our perception. These perceptual bindings can also be used to create transparency within an audiovisual system (Callear 2012: 30). Experiments within the field of cognitive science have shown us that we tend to bind audio and visual events when they are temporally, spatially and semantically congruent.

> *Semantic congruency* usually refers to those situations in which pairs of auditory and visual stimuli are presented that vary (i.e. match vs. mismatch) in terms of their *identity* and/or *meaning*. (Spence 2011: 972)

There is also a large amount of evidence showing that ‘many non-arbitrary crossmodal correspondences exist between a variety of auditory and visual stimulus features’ and ‘dimensions’ (ibid. 2011: 975). These automatic crossmodal correspondences (Fig. 5.4) include the binding of pitch to visual phenomena such as spatial height, the lightness or darkness of a surface, the size and shape of an object and the spatial frequency of a visual pattern (Evans and Treisman 2010). The amplitude of a sound has also been found to affect visual phenomena such that ‘loud sounds can improve the perception of bright lights and large objects, whereas soft sounds facilitate the perception of dim lights and small objects’ (ibid. 2010: 1-2). Loud sounds have also been found to correspond to ‘visual stimuli that have a higher contrast’ (Spence 2011: 974). Further to this, Brunel, Carvalho and Goldstone (2015) were able to show that unrelated audiovisual stimuli can be ‘integrated within a single memory trace’ (ibid. 2015: 2) such that ‘each component is no longer accessible individually without an effect of the other component’ (ibid. 2015: 8). They further state that the ‘greater the prior knowledge in the system about the fact that two stimuli belong together, the stronger these stimuli will be coupled’ (ibid. 2015: 9). Repetition could be a way to increase the audio-spectator’s knowledge of the stimuli. This could be evidence for Chion’s claim that synchresis is ‘Pavlovian’ (Chion 1994: 63). In the context of abstract audiovisual objects, this repetition could be used as a

compositional technique to bind seemingly unrelated audio and visual media through repeated association. Further, in order to balance the transparency of congruent associations, audio and visual events that are not spatiotemporally congruent, semantically congruent or automatic could contribute to the opacity of a mapping system.

5.2 Interactive Machine Learning

IML tools such as the Wekinator (Fiebrink, Trueman and Cook 2009), the Rapidmix API (Bernardo et al. 2017) and the MIMIC[35] platform provide easy access to machine learning algorithms for non-experts in the field. An IML approach to real-time audiovisual composition enables the relatively straight-forward creation of complex audiovisual mappings within a reasonable time frame and fosters an intuitive approach to exploring audiovisual relations.

Machine learning is a well-established field that initially grew out of research into artificial intelligence (AI) but also makes use of knowledge from fields such as 'probability and statistics, computational complexity theory, control theory, information theory, philosophy, psychology and neurobiology' (Mitchell 1997: 2). At the heart of machine learning are algorithms that allow a system to improve its performance relative to a specific goal. Tom Mitchell defines machine learning in the following way:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. (Mitchell 1997: 2)

Gerhard Widmer defines machine learning like so:

> Machine Learning may be defined as the subfield of Artificial Intelligence that studies the phenomenon of learning, both by constructing formal theoretical models, and by developing operational algorithms and computer programs that can learn. (Widmer in Miranda 2013: 70)

Russell and Norvig similarly describe machine learning as 'the subfield of AI concerned with programs that learn from experience' (Russell and Norvig 2010: 523). Each of these definitions focuses on the concept of learning, which can be interpreted in a number of ways. Russell and Norvig describe learning in terms of behaviour optimisation, whereas Michalski talks about learning in terms of gaining new knowledge about the world. However, both of these goals arise as a result of experience. Widmer notes that, despite the multiple approaches to learning, all algorithms and

[35] https://mimicproject.com/about (accessed 09/09/2020).

methods fundamentally arise as a result of ‘generalisations that are consistent with the given observations (ie. the given data)’ (Widmer in Miranda 2013: 70-71).

Classic machine learning, or CML (Bernardo et al. 2017: 2), involves ‘lengthy and asynchronous iterations’, resulting in models that the end-user generally has little control over (Amershi 2014: 1). In contrast, IML allows for rapid development and training of machine learning models by the end-users themselves. Further to this, the end-user does not need to be a machine learning expert (Bernardo et al. 2017: 2). This makes the exploration of IML attractive for artists who are less interested in the inner workings of machine learning algorithms than in how they can be used to produce art. Rebecca Fiebrink states that ‘interactive machine learning has marvellous potential as a creativity support tool’, citing examples of composers working with the Wekinator (Fiebrink and Trueman 2012). Fiebrink notes that users were able to focus ‘on crafting and evaluating the relationships between gesture and sound’ rather than expending time and energy ‘writing code or designing mathematical functions’ (ibid.). Considering the advantages of IML outlined above, it must be pointed out that placing the design of machine learning models into the hands of non-expert users also presents challenges. Bernardo et al. (2017) have identified these challenges, outlining difficulties in:

- predicting future performance.
- the selection of ‘appropriate features’.
- the collection and understanding of data.
- connecting ‘machine learning tools to other tools of interest’.
- identifying whether a design is suitable for the ‘available algorithms, features and data’.

With the above in mind, it seems the advantages eclipse the challenges when using IML techniques in an artistic context. It is possible to create complex audiovisual relations that would have otherwise taken a significant amount of time to code or would have been outright impossible. As evidenced throughout history, from the development of overdrive distortion to the use of the Roland TB-303, the misuse of technology sometimes creates unexpected and interesting artistic results.

These above challenges were taken into consideration in the development of the Rapidmix API, which is a toolkit that incorporates IML technologies such as the Wekinator, and ‘aims to make IML efficient and accessible to developers who are new to machine learning’ (Bernardo et al. 2017). It provides access to algorithms and features that are commonly used in IML applications such as supervised regression and classification models. The term ‘supervised’ here refers to the learning

approach. That is, the user provides the ideal examples that the model should be looking for in the training data. This is opposed to unsupervised learning, where the algorithm must infer the goals itself from the dataset. Supervised learning algorithms can take the form of regression or classification models. A classification model 'encodes a function mapping the input space to a discrete set of output classes' (Fiebrink, Trueman and Cook 2009: 2), whereas a regression model implemented as a neural network 'can map inputs to a continuous output space' (ibid.). By providing access to pre-built models such as these, Rapidmix API lets the user focus on the output of the application rather than the construction of algorithms.

There are some examples of artists and musicians leveraging the mapping capabilities of neural networks. Lee, Freed and Wessel (1991) introduced *MAXNet,* an object for Miller Puckette's MAX environment (Puckette and Zicarelli 1990), that allowed for the mapping of musical gestures to audio synthesis. Zbyszynski et al. (2021) present methods for exploring gestural to audio timbral mappings using interactive machine learning approaches and electromyography sensors. Laetitia Sonami's *Spring Spyre* instrument is an example of the artistic use of IML techniques. Sonami has been developing this instrument since 2012 in collaboration with Rebecca Fiebrink (Fiebrink and Sonami 2020). Sonami built the instrument using 'thin springs attached to audio pickups' (ibid. : 239). Features from these signals are then extracted and mapped to audio synthesis control parameters using a set of multilayer perceptron neural networks. These practitioners and artists are using interactive machine learning techniques specifically within the musical domain. The work presented in this thesis differs from other work in this area in that it extends into the visual domain, and in doing so, not only creates complex, nonlinear mappings between input data and sonic material, but also perceptual mappings between the audio and visual material itself. Further, the practice is specifically located within the wider audiovisual art context of generative visual music, as mapped out in Chapter 2.

Chris Kiefer's *10K Video*[36] (2018) is a real-time improvised audiovisual performance in which ten thousand individual neural networks have been trained on the pixel behaviour of a short video. These neural networks also act as oscillators that produce clouds of noise-based sound. Kiefer controls the neural networks in real-time using conceptors (Magnusson, Eldridge and Kiefer 2020: 511), which are high-level mechanisms, sometimes conceptualised as filters, that enable the user to 'control a multiplicity of processing modes' of recurrent neural networks (Jaeger 2014: 10). Recurrent neural networks (RNN) are a specific type of neural network that can generate time varying signals and patterns. Kiefer is here using neural networks to create an audiovisual piece. Neural networks are also

[36] https://vimeo.com/268980331 (accessed 17/12/2021).

being used in this thesis to create audiovisual art. However, the research presented here is focused on a different type of neural network and explores a different control paradigm.

IML technology has been shown to foster creativity and provide a successful platform from which to create artistic work (Fiebrink and Trueman 2012). This paradigm will be used as the basis for controlling and interacting with the audiovisual compositions that are presented later in the thesis. This approach has not been significantly explored within the field of audiovisual art and thus represents a fertile area within which new approaches to the practice of audiovisual composition can be explored.

5.3 Neural AV Mapper

As discussed above, the issue of mapping permeates the entire field of audiovisual art. There is a balance to be struck between implying a cause-and-effect relationship between the audio and the visuals and creating enough complexity to hide the nature of the relationship (Sá, Caramieux and Tanaka 2014: 85). The task of creating a mapping network between parameters that has enough complexity and flexibility to achieve this goal traditionally takes a lot of programming and manual adjustment. However, with the use of IML algorithms, this large workload can be reduced, and the mappings can be explored intuitively through playing with the system in real time. In this way, the algorithm takes care of the mapping so the artist can concentrate on the aesthetic result.

The system described below was developed as a live-performance tool. In the prototype state presented here, it is made up of the following performance interface elements:

1. Real-time mouse or trackpad input
2. On-screen graphical user interface (GUI) showing audio and visual parameters
3. On-screen GUI allowing the user to save and load regression models
4. On-screen feedback displaying operating state of the current neural network
5. Visible mouse-pointer showing location of input coordinates on the screen

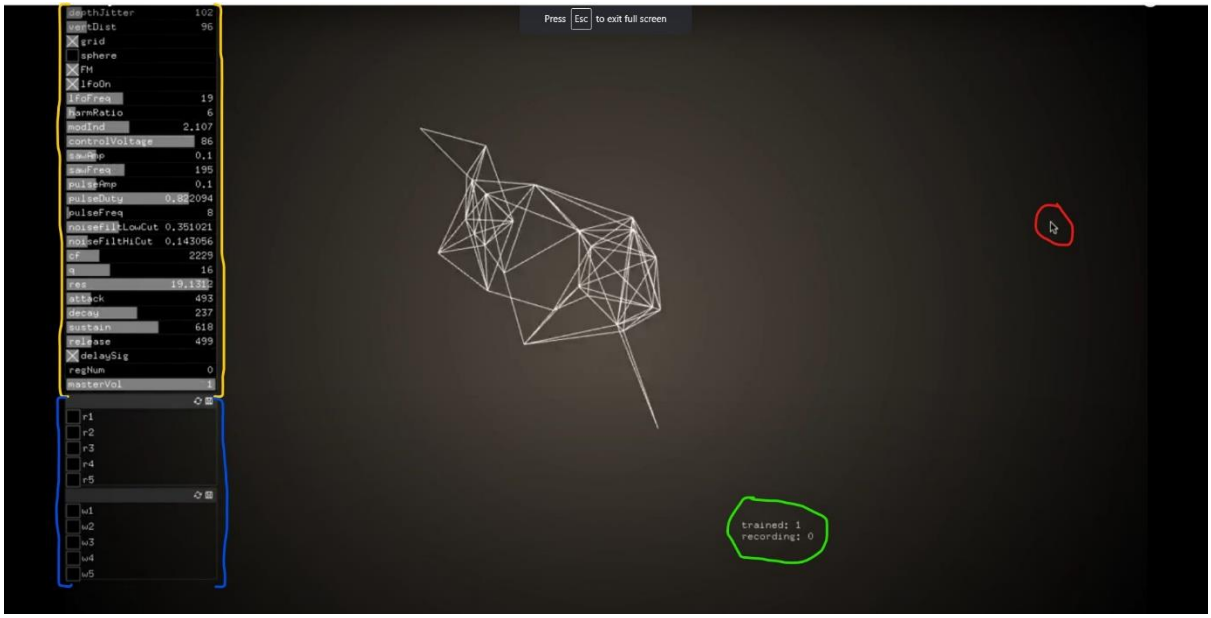

Figure 5.5 Still from Study No.2 showing performance interface elements.

Fig. 5.5 shows some of the interface elements that were used during the development of *Study No.2*. On the left side of the image, the audio and visual parameter values are displayed (bracketed in yellow). The parameters emerged gradually as both the audio and visual processing code was developed. Some of the GUI elements take the form of sliders and others take the form of toggle buttons. Below that (bracketed in blue), are buttons that are used to read and write the trained models. This allows the user to save and load trained models. At the bottom of the screen, circled in green, there is simple graphical feedback indicating to the user whether the current neural network model is trained or if recording is taking place. Finally, circled in red, the mouse position is shown with the standard arrow. The coordinates located at the position of the mouse are used as input data for the neural networks.

The aesthetic output of the system (visually displayed in the centre of Fig. 5.5) consists of wireframe three-dimensional shapes and synthesised audio. In *Study No.3* and *Study No.4*, when simultaneous audiovisual objects are introduced, each parameter is technically implemented as an array with different values for each audiovisual object. During development, a specific workflow emerged when preparing content and performing with the system. This workflow involves the following steps:

1. Randomise, or manually adjust, audio and visual parameters until the system outputs aesthetic material that is deemed to be well-matched
2. Use the mouse or trackpad to move the cursor to a specific location on the screen

3. Store the screen coordinates as *input* training data
4. Store the audio and visual parameters as *output* training data
5. Combine input and output training data and store within a single training example
6. Repeat steps 1 to 5 until a sufficient amount of training examples are stored
7. Train the neural networks with the training data
8. Activate the neural network models
9. Use the mouse or trackpad to change the on-screen position of the cursor, observing the audiovisual output in real-time
10. If necessary, repeat the above steps until the aesthetic output is satisfactory

Steps 1 to 4 above represent the data collection and training phases of the workflow. These steps happen prior to performance with the system. Steps 2 to 5 are demonstrated in the recording of *Study No.3* below.

5.3.1 Regression Analysis

Regression analysis involves using a weighted function to calculate a line of best fit between sets of training data. This line of best fit can then be used to calculate the corresponding relationship between new input and output data in a continuous manner. Multiple inputs can be mapped to multiple outputs. The RapidLib[37] machine learning library was used in the development of the system discussed here. This library provides algorithms and data structures specifically intended for use in an IML system. It is a component of the wider RapidMix API discussed in section 5.2 above. In order to support multiple outputs, RapidLib creates a new neural network for each parameter. This is a result of the underlying Weka library, which 'only supports architectures with a single output node' (Fiebrink, Trueman and Cook 2009: 4). The creation of multiple neural networks is hidden from the user by the rapidLib library, and, in this way, a large number of parameters can be controlled simultaneously. This characteristic means that regression mapping is particularly suited to the real-time control of multi-parametric, generative audiovisuals.

There are some differences between this approach and a linear interpolation approach that are worth noting. In the case of the system presented here, a linear interpolation approach would mean mapping the two dimensional coordinate range of the performance interface, to each of the parameter ranges individually. This would create a direct, linear mapping between the input data and the output parameter value. The linear nature of the mapping could present a perceptual problem when trying to

[37] https://github.com/mzed/RapidLib (accessed 21/12/2021).

use the system as an expressive tool. As discussed above, clearly transparent mappings are at risk of becoming perceptually uninteresting. The use of a multilayered perceptron neural network, with backpropagation and a nonlinear activation function, results in nonlinear associations between the input and output data. This introduces an element of uncertainty to the system, which is often desirable in an artistic context. It is hoped that this will help to create opaque mappings that may aid in maintaining interest. It is also hoped that this system will support the ability to perform expressively. Another difference between the two approaches is the required time taken to map parameters as the dimensionality of the model increases. A linear interpolation approach may be appropriate for mapping lower dimensional spaces. However, as the dimensionality grows, the time needed to map the space would grow prohibitively large.

5.3.2 Mapping Characteristics

With reference to the mapping terminology discussed above, regression mapping could be described as an example of a complex, implicit mapping strategy (Arfib et al. 2002). This mapping can also be described as nonlinear in that variation in the input data does not necessarily correspond to a perceived proportional variation in the output data. The implementation described below is divergent in nature but could also be utilised in a convergent manner. Even though only *x* and *y* coordinates were used in these studies, the input could easily consist of information represented by data in higher dimensions whilst the output data could just as easily be made up of fewer elements.

The audiovisual system used to construct these studies could be described as data-driven. Both the audio and visual material is synthesised through code within separate algorithms. Parameters of these algorithms are then controlled concurrently through a single input method. The input data is therefore driving the audio and visual events. From a perceptual perspective, the mapping system could be described as opaque. There are no hard-coded correspondences between the audio and visual parameters. Perceptual transparency can be integrated into the system through cross-modal correspondences as described above. For example, a noisy audio texture can be coupled with a disjointed, rapidly moving shape. Or a high-pitched tone can be coupled with a small, relatively solid object (Abbado 1988). This could create interest for the audience whilst also maintaining ambiguity in the actual audiovisual mapping. This is what Sá, Caramieux and Tanaka (2014) call ‘fungible audio-visual mapping’.

5.3.3 Theoretical Relationships

Regarding audiovisual balance, this approach to controlling audiovisuals aims to avoid a dominant/dependent relationship between the material. The audio and visual algorithms were

developed separately and are then controlled simultaneously. Neither modality is dependent on the other for its temporal motion or as the source of its energy. They are both equally as dependent on the input data. Keeping the idea of isolated-structural-incoherence in mind, the aim is to make the media appear perceptually and aesthetically interdependent. It is commonly held that vision will dominate hearing wherever possible. Sá, Caramieux and Tanaka (2014) have outlined ways of redressing this imbalance 'by directing sensory organs towards a target, and/or modulating the sensitivity of neural circuits accordingly' (ibid. 2014: 86). They also found that by simplifying the visual elements using Gestalt principles and reducing discontinuities, the audience's 'perceptual resolution can be optimised for the music' (ibid.). Whereas Sá, Caramieux and Tanaka are intentionally attempting to 'keep the music in the foreground' (ibid.), the intention here is to create a perceptual equality between the two senses where the audience can experience the interaction between the two modalities.

5.3.4 Technical Structure

The system was built using openFrameworks[38], ofxRapidLib[39] and Maximilian[40]. The source code can be found at the Neural AV Mapper[41] GitHub repository or in the accompanying media pack. OpenFrameworks provides a flexible, open source environment within which all the elements of the system could be developed without the need to communicate between programs. It is built in C++, which offers low-level access that optimises CPU and GPU performance. There is a vibrant online

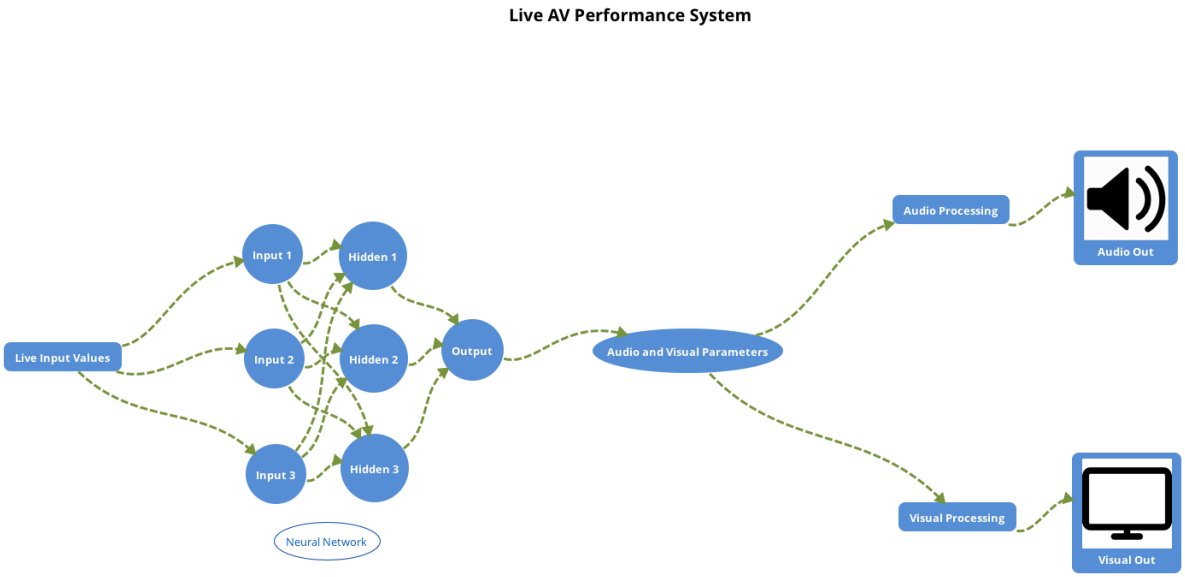

Figure 5.6 Neural AV Mapper system.

[38] https://openframeworks.cc/ (accessed 30/07/2020).
[39] https://github.com/mzed/ofxRapidLib (accessed 30/07/2020).
[40] https://github.com/micknoise/Maximilian (accessed 30/07/2020).
[41] https://github.com/bDunph/neuralAvMapper (accessed 02/08/2020).

community of artists and developers with an active forum. It also contains many easy to use classes aimed at creating real-time graphics. The audio library, Maximilian, was used as it was 'deliberately designed to integrate well' (Grierson and Kiefer 2011: 276) with openFrameworks. In order to integrate IML capabilities, the ofxRapidLib addon was used. This is an openFrameworks wrapper for the RapidLib IML library that makes up part of the Rapidmix API discussed above. This library allows for easy integration of several machine learning algorithms into an interactive workflow. A diagram of the system is shown in Fig. 5.6.

The core of the system is the neural network provided by ofxRapidLib. In the *avObject()* class, the function *randomParams()* randomises the audio and visual parameters. The parameters themselves will be discussed in the following section. Once a satisfactory combination of parameters has been found, the function, *trainingExamples()*, prepares the values to be processed as an individual training example (Ex. 5.1).

```
00 void avObject::trainingExamples(int _numVoices, int _x, int _y)
01 {
02
03     int numVoices = _numVoices;
04     int x = _x;
05     int y = _y;
06
07     for(int i = 0; i < numVoices; i++)
08     {
09         rapidlib::trainingExample tempExample[numVoices];
10         tempExample[i].input = {double(x), double(y)};
11         tempExample[i].output = {(double) avParameters[i] };
12         trainingSet[i].push_back(tempExample[i]);
13     }
14 }
```

Example 5.1 Function to create a model training set.

This function takes three int values. The first, *_numVoices*, is used as an argument in the *loop* declaration at line 07. This value refers to the number of audiovisual objects that are being processed. The second and third arguments to the function are the x and y mouse positions. At line 09 an array of type *trainingExample* is declared. The *x* and *y* mouse positions are assigned to the example input array at line 10. The current values of all the audio and visual parameters are then assigned to the output array. Finally, at line 12 the current *tempExample[i]* is assigned to the *trainingSet[i]* vector. Once all the training examples are recorded, the model is then trained. See Ex. 5.2.

```
01 bool avObject::trainModel(int _numVoices)
02 {
03
04         int numVoices = _numVoices;
05
06         for(int i = 0; i < numVoices; i++)
07         {
08                 result = reg[i][regNum[i]].train(trainingSet[i]);
09         }
10
11         return result;
12 }
```

Example 5.2 Training the model.

At line 08 above, *reg[][]* is a two-dimensional array of *rapidlib::regression* objects. The first row determines the current audiovisual object. The second row determines the current neural network selected for that object. It is possible to load multiple regression models for each object by switching between neural networks. These neural networks can be selected using the GUI buttons *r1* to *r5* (see Fig. 5.5). The *r* stands for *read*, which essentially loads a pre-saved neural network. This is explained in more detail in the next section. Here, the *train()* function is called and the current *trainingSet* is passed as its only argument. When the network is trained, it returns a *bool* which is assigned to *result*. This *bool* is then returned from *avObject::trainModel()*. Once the neural network is trained it runs automatically when *avObject::trainedOutput()* is called. See Ex. 5.3.

```
00 void avObject::trainedOutput(int _numVoices, int _x, int _y)
01 {
02     int numVoices = _numVoices;
03     int x = _x;
04     int y = _y;
05     std::vector<double> input[numVoices];
06     std::vector<double> output[numVoices];
07 
08         for(int i = 0; i < numVoices; i++)
09         {
10 
11                 input[i].push_back (double(x));
12                 input[i].push_back (double(y));
13                 output[i] = reg[i][regNum[i]].run(input[i]);
14 
15                 vertices[i] = output[i][0];
16                 if (output[i][0] < 2)
17                 {
18                     vertices[i] = 2;
19                 }
20 
21            etc......
22         }
23 }
```

Example 5.3 Running the trained model.

Similar to *avObject::trainingExamples()*, the above function takes three arguments: the number of audiovisual objects to iterate the *for loop* at line 08, and the *x* and *y* positions of the mouse. Two *vectors* are declared at lines 05 and 06. These contain the input and output data. The current *x* and *y* values are added to the input vector at lines 11 and 12. The currently selected regression model then calls the *run()* function that takes the *input[i]* vector as an argument. This function sends the input values to the neural network and returns the *output[i]* vector. The output vector contains the calculated parameter values for the particular input according to the trained regression model. At line 15 the first element of the output vector is accessed and assigned to the *vertices* parameter for that particular audiovisual object. Lines 16 to 19 define a lower limit for the output parameter value. This requirement was discovered during development, as any shape with less than two vertices is simply a point and was not useful in this context. Each of the parameters are retrieved in the same way. In this way, the output of each neural network is fed directly to both the audio and visual processing sections of code.

5.4 Mapping Studies

The aim of these studies is to explore the compositional possibilities provided by the system described above. It has been posited above that IML, and more specifically multilayer perceptron neural networks, have the potential to provide a fertile environment within which audiovisual art can be created. The approach to the development of the audiovisual material was to write separate code for the generation of the audio and visuals respectively. There are no direct mappings between the audio and visual parameters in these studies. The neural network is intended to act as the sole mapping layer. It is hoped that the audio-spectator will perceive audiovisual correspondences through their senses.

5.4.1 Compositional Approach

The compositional approach taken with the studies below involved initially choosing audiovisual objects that exhibited strong perceptual bonds. During performance, these were treated as cadential points within the audiovisual space. To achieve this, the system was trained using on-screen *x* and *y* coordinates as input data, and the parameter values of these chosen objects as output data. After training was completed, the user could then drag the mouse over the screen, observing smooth interpolation of output parameter values. The initial cadential audiovisual objects could be expressed by moving the cursor to the location that was given as a training example. By choosing different on-screen locations for the initial objects, the user could move the cursor from one to another, with the coordinates in-between causing the audiovisual parameters to interpolate smoothly.

For an example of this approach in practice see *Study No.1*[42] (included in the media pack at *mediaFiles/iml_studies/study1.mov*). At 0:04, the cursor is located in the top-left area of the performance interface. The screen coordinates at this location are fed into the input of the neural network. This results in the specific parameter values, shown on the left of the performance interface, being output from the neural network. The system was trained initially by pairing the mouse coordinates with the parameter values, and storing them as a training example. At 00:11, the mouse jumps to the bottom left of the screen. The parameter values then change, resulting in a different audio texture and visual form. This is another one of the audiovisual objects that were used to train the system. The mouse then moves to the bottom left area of the performance interface at 00:29. This again causes the parameters and resulting audiovisual figure to change. At 00:34, the mouse moves to the top right area of the performance interface to trigger the final audiovisual object used to train

[42] https://www.youtube.com/watch?v=jFXdvAaCQok (accessed 14/12/2021).

the system. Throughout the rest of the study, the mouse then travels throughout the performance interface. The changing coordinates of the mouse are mapped continuously to the parameter ranges resulting in a dynamic audiovisual texture. It was hoped that this approach to creating an audiovisual composition would give the composer the capability to manipulate the perception of audiovisual relationships through the exploration of the space between the cadential points. In this way, the composer could intentionally move away from and back to areas that are known to exhibit strong perceptual coupling.

In the studies below, the main criteria used to decide if the audio and visuals were well matched, was the relative-temporal-motion of the material in each modality. When this movement was similar enough across the material, it resulted in a sense of synchronicity that created strong cross-modal binding. However, the audio and visual material below is not directly mapped. Instead, the source of energy for each modality is generated through noise algorithms within each generative patch. This results in a loose synchronicity, where the material drifts away from, and towards, unity. This could be an example of Cook's concept of ventriloquism between audio and visual media, which was briefly discussed in Chapter 3. It could also be a demonstration of a cross-modal incarnation of the Gestalt principle of common fate.

> The common fate principle states that elements tend to be perceived as grouped together if they move together. (Todorovic 2008)

When the amplitude envelopes of the audio elements align with the visual motion of the vertices, the separate audio and visual elements appear to be grouped together. This grouping of similar motion across sensory modalities is also explored by Moody, Fells and Bailey (2007) in the context of Chion's synchresis. It is hoped that purposefully combining audiovisual objects in this way would increase the potential for a strong perceptual connection between the audio and visual material. Further, it was hoped that the use of structurally ambiguous noisy material would also foster a sense of balance in the context of isolated-structural-incoherence. An attempt was also made to balance the relative-expressive-range of the studies by implementing a minimal aesthetic in both the visuals and the audio. The generation of the material is discussed in more detail in the next section.

5.4.2 Generating Material

The functions responsible for generating the visual structures and audio textures were prepared prior to performance and placed within the *avObject()* class. During performance, the material was then

manipulated in real-time by controlling a range of parameters. The construction of the functions responsible for generating both the audio and visual material are now described.

Visuals

The visuals are made up of wireframe mesh cubes and spheres. The visuals alternate between solid structures and chaotic arrangements of points and lines. This minimal aesthetic suited the purposes of these initial studies in that they are flexible and enabled the exploration of solid and chaotic states. The visual processing takes place in the *avObject::visual()* function. See Ex. 5.4.

```
00 if(shape[i] == 0)
01 {
02     //********* Cube **********/
03     for(int x=0; x<vertices[i]; ++x)
04     {
05             for(int y=0; y<vertices[i]; ++y)
06             {
07                     for(int z=0; z<vertices[i]; ++z)
08                     {
09                             int jit = ofRandom(-depthJitter[i],
depthJitter[i]);
10                             float xPos, yPos, zPos;
11                             int offset = 50;
12                             xPos = x * offset + jit;
13                             yPos = y * offset + jit;
14                             zPos = z * offset + jit;
15                             ofVec3f vertex(xPos, yPos, zPos);
16                             meshes[i][selVis].addVertex(vertex);
17
       meshes[i][selVis].addColor(ofColor::white);
18                      }
19             }
20      }
21 }
```

Example 5.4 Construction of a cube.

The cube is constructed manually according to the code above. At line 00, there is an *if-statement* checking whether the shape parameter is set to 0. If it is, three nested *for-loops* are used to place the vertices. The number of iterations for each of the *for-loops* at lines 03, 05 and 07 are determined by the *vertices[i]* parameter of the particular audiovisual object instance. At line 09, the *depthJitter[i]* parameter is used as an argument to *ofRandom()*. This returns a random value in the range *-depthJitter* to *depthJitter*. This value is then assigned to *jit*. This *int* is then added to each vertex calculation at lines 12, 13 and 14. This randomisation process creates the noisy movement in the visual material.

The visual structures are created using an *ofMesh*[43] object. This object is contained in the two-dimensional array *meshes[i][selVis]*. Here, *i* refers to the number of audiovisual object instances. The index value *selVis* is a bool. This means there are two meshes for each object instance. These meshes are switched at the end of each frame to allow the structures to change dimensions. At line 16 the vertex is added to the mesh. At line 17 the vertex is given a colour. Ex. 5.5 details the construction of a sphere.

```
00 else if(shape[i] == 1)
01 {
02         //********* Sphere ***********/
03         //ref: https://github.com/nicohsieh/sphere-
freq/blob/master/src/testApp.cpp
04         int radius = 250;
05         meshes[i][selVis].addVertex(ofVec3f(0,0,1*radius));
06
07         for (int j=1; j<vertices[i]; j++)
08         {
09                 double xPos, yPos, zPos;
10                 double phi = PI * double(j)/(vertices[i]);
11                 double cosPhi = cos(phi);
12                 double sinPhi = sin(phi);
13                 for (int k=0; k<vertices[i]; k++)
14                 {
15                     float jitter = ofRandom(-depthJitter[i],
depthJitter[i]);
16                     double theta = TWO_PI * double(k)/(vertices[i]);
17                     xPos = cos(theta)*sinPhi*radius + jitter;
18                     yPos = sin(theta)*sinPhi*radius + jitter;
19                     zPos = cosPhi*radius + jitter;
20                     meshes[i][selVis].addColor(ofColor::red);
21                     meshes[i][selVis].addVertex(ofVec3f(xPos, yPos,
zPos));
22                 }
23         }
24         meshes[i][selVis].addVertex(ofVec3f(0,0,-1*radius));
25 }
```

Example 5.5 Construction of a sphere.

If *shape[i]* is equal to 1, the visual structure becomes a sphere. As indicated on line 03, the code for constructing the sphere was adapted from Nico Hsieh's *sphere-freq* repository. At line 05 the initial vertex is added to *meshes[i][selVis]* in the positive z direction at a distance determined by *radius*. Following this, a nested *for-loop* iterates through the rest of the vertices. Similar to the cube structure,

[43] OpenFrameworks classes can be found in the online documentation https://openframeworks.cc/documentation/ (accessed 29/09/2020).

the position of each vertex is summed with a *jitter* value from lines 17 to 19. This value is calculated at line 15 using *depthJitter[i]* as arguments for *ofRandom()*. At line 20 a red colour is applied to the mesh and at line 21 the vertex is added. Finally, at line 24, the last vertex in the negative *z* direction is added. The vertices are then indexed as shown in Ex. 5.6.

```
00 //********* From ofBook - Basics of Generating Meshes from an Image
*********//
01 float connectionDistance = vertDist[i];
02 int numVerts = meshes[i][selVis].getNumVertices();
03 for (int a=0; a<numVerts; ++a)
04 {
05    ofVec3f verta = meshes[i][selVis].getVertex(a);
06    for (int b=a+1; b<numVerts; ++b)
07    {
08          ofVec3f vertb = meshes[i][selVis].getVertex(b);
09          float distance = verta.distance(vertb);
10          if (distance <= connectionDistance)
11          {
12                // In OF_PRIMITIVE_LINES, every pair of vertices or
indices will be
13                // connected to form a line
14                meshes[i][selVis].addIndex(a);
15                meshes[i][selVis].addIndex(b);
16          }
17    }
18 }
```

Example 5.6 Adding indices to mesh vertices.

After the vertex positions and colour are added to the mesh objects, the indices are added using the above code. Indices are used in OpenGL[44] to label vertices, which allows the same vertex to be used on connected triangles. Otherwise, duplicate vertices would need to be generated for the same corner of the shape. As indicated on line 00, this code is adapted from Hadley (2020). At line 01, the parameter *vertDist[i]* is assigned to *connectionDistance*. This parameter is received from the output of the neural network. At line 02 the number of vertices is returned from *getNumVertices()*. This value is used to set the iteration limit for the nested for loops at lines 03 and 06. At line 05, the first vertex is returned from *getVertex()* and stored in *verta*. The inner *for-loop* is one step ahead of the outer loop. The next vertex is returned at line 08 and stored in *vertb*. At line 09 the distance between the two vertices is calculated and stored in the variable *distance*. This is then used at line 10 as a comparison to *connectionDistance*. If *distance* is less than *connectionDistance*, indices are added to the current vertices at lines 14 and 15. Earlier in *avObject::avSetup()*, each mesh object was set to

[44] http://www.opengl-tutorial.org/intermediate-tutorials/tutorial-9-vbo-indexing/ (accessed 21/12/2021).

OF_PRIMITIVE_LINES. This is a drawing mode that instructs OpenGL to draw lines between each vertex. This is responsible for the wireframe appearance of the structures. Once the structures are ready, *draw()* is called for each mesh inside *avObject::drawVisual()*. This is the command that actually draws the visuals on the screen.

The visual parameters used here are limited to the number of vertices (*vertices*), amount of movement on the z axis (*depthJitter*), distance between vertices (*vertDist*) and the underlying primitive shape (*shape*). The shapes are all rotating at a constant rate in the videos. This was implemented to try to convey the 3D nature of the shapes as there is no lighting or shading on the wireframe objects.

Audio

Generation of the audio textures was based on an FM approach to synthesis. Fig. 5.7 shows the list of audio parameters that are connected to the output of the neural network. The intention was to create complementary audio and visual spaces that would permit the exploration of audiovisual balance through the manipulation of the audio-spectator's perception of relative temporal motion. Ex. 5.7 shows the construction of the FM audio patch.

```
00 lfo[i] = lfOsc[i].sinewave(lfoFreq[i]);
01 harmonicity[i] = sawFreq[i] * (harmRatio[i] * lfo[i]);
02 modAmp[i] = harmonicity[i] * (modInd[i] * lfo[i]);
03 modulator[i] = mod[i].sinewave(harmonicity[i]) * modAmp[i];
04 saw[i] = sawOsc[i].saw(modulator[i] + sawFreq[i]) * sawAmp[i];
05 pulse[i] = pulseGen[i].pulse(pulseFreq[i] + saw[i], pulseDuty[i] *
lfo[i]) * pulseAmp[i];
```

Example 5.7 FM audio patch.

The audio signal is generated with the above code, which is executed within a *for-loop* that cycles through each audiovisual object. They are indexed using *[i]*. At line 00, a low frequency sine wave is generated. The object *lfOsc* is an oscillator of type *maxiOsc*. Similarly *mod*, *sawOsc* and *pulseGen* are all *maxiOsc* oscillators. The parameter *lfoFreq* is used to determine the frequency of *lfOsc*. This parameter is output from the neural network. The *lfOsc* signal is then stored in a double called *lfo*. At line 01, *lfo* is first multiplied by the neural network parameter *harmRatio*. The result is then multiplied by another parameter from the neural network, *sawFreq*. This calculation gives the harmonicity value. The *lfo* is then used to modulate the modulation index, *modInd*. This value is then used to modulate the harmonicity value to give the modulation amplitude *modAmp*. This value modulates the amplitude of the modulator signal. The harmonicity value is again used here as the frequency of the mod signal. The modulator signal is then added to the frequency of the carrier signal, which is generated using

sawOsc. The parameter *sawAmp* is also used here. At line 05 a pulse signal is added using *pulseGen*. Here, the frequency is modulated by the saw signal and the duty cycle is modulated by the LFO signal. The parameter *pulseAmp* is also used here. This synthesis algorithm was developed through experimentation rather than using precise methods. Due to this, there may be inefficient or even redundant calculations. However, the results were deemed to be appropriate for the application at this point in time. The output signal is then processed through several filters and a delay, to shape the sound. These filters and delay are also dependent on parameters from the neural network.

Audio Parameters

	Function
FM	Toggles FM synthesis on and off
lfoOn	Turns the lfo on and off
lfoFreq	Frequency of the lfo
harmRatio	Sets the value of the harmonic ratio for FM synthesis
modInd	Sets the modulation index for FM synthesis
controlVoltage	Sets the frequency of the phasor used to trigger the amplitude envelope
sawAmp	Amplitude of the saw wave generator
sawFreq	Frequency of the saw wave generator
pulseAmp	Amplitude of the pulse wave generator
pulseFreq	Frequency of the pulse wave generator
noiseFiltLowCut	Threshold of the low pass filter
noiseFiltHiCut	Threshold of the high pass filter
cf	Centre frequency of the bandpass filter
q	Q width of the bandpass filter
res	Resonance amount of the bandpass filter
attack	Attack value for adsr envelope
decay	Decay value for the adsr envelope
sustain	Sustain value for the adsr envelope
release	Release value for the adsr envelope
delaySig	Toggle delay on or off

Figure 5.7 Audio parameters output from the neural network.

5.4.3 Studies

This section analyses four studies that were created to test the compositional and performative potential of the system. The studies can be viewed by following the YouTube links contained in the footnotes, or in the accompanying media pack at *mediaFiles/iml_studies*.

Study No.1[45]

This study is limited to the exploration of a single regression model. The intention behind this limitation is to test the expressive capabilities of one regression model before expanding into multiple models. The goal was to see how much variation of material and expressive range a single model was capable of providing.

The setup of the model first involved picking four distinct audiovisual objects. The objects were chosen through manual adjustment of the audio and visual parameters. The choice of visuals and audio was, for the most part, left to artistic intuition. However, some universal correspondences that have been proven to exist in the literature were intentionally followed as it is clear that these principles create strong cross-modal correspondences.

The audiovisual object *av1* (0:00 - 0:10) was decided upon as it has been shown above that smaller solid objects correspond to higher pitched noises. Following this, *av2* (0:11 - 0:17) expands on *av1* structurally (where *av1* is a single cube, *av2* is made up of many) but is more unstable. The audio matches this by dropping in pitch and introducing more noise to correspond to the more unstable visual structure. Further, *av3* (00:29 - 00:33) is sparse in terms of audio and visual content. The disjointed visuals are matched in the audio domain with glitchy discontinuous sounds. Finally, *av4* (0:34 - 0:40) is an expansion of *av3* (similar to how *av2* is an expansion of *av1*) in that it fills out the space by creating an undefined chaotic mass accompanied by a loud and spectrally rich sound texture. During training, these four objects were associated with coordinates situated in each of the four corners of the performance interface and, during performance, were intended to act as perceptual anchor points or cadential figures that could help to resolve tension (Fig. 5.8).

[45] https://youtu.be/jFXdvAaCQok (accessed 20/08/2020).

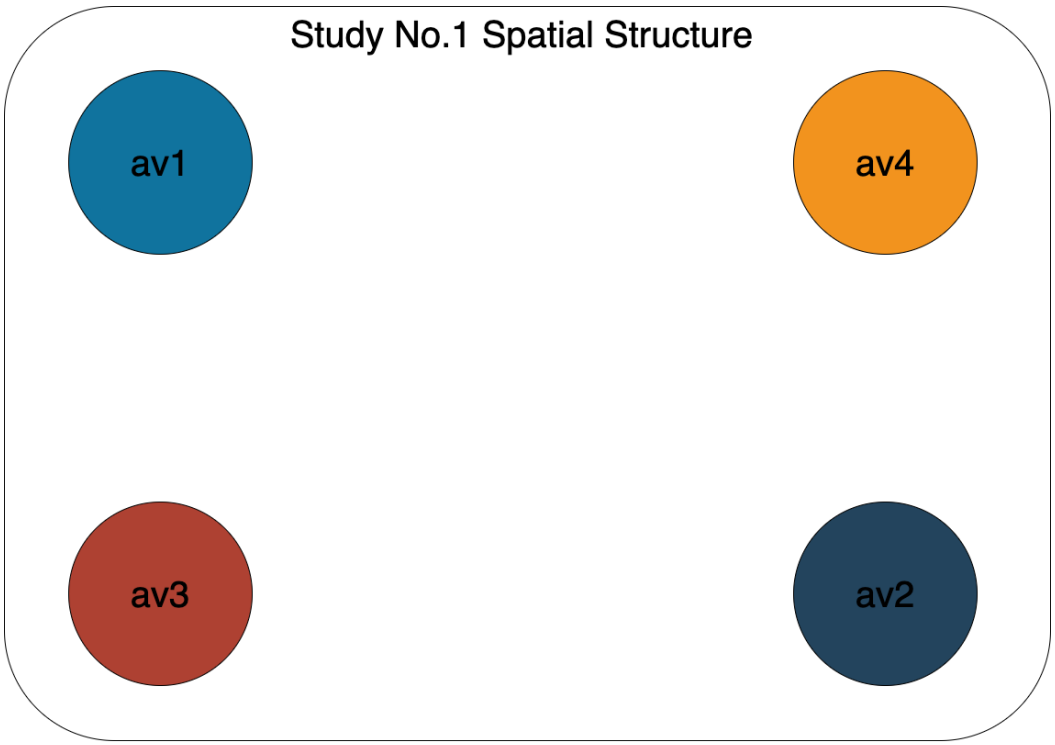

Figure 5.8 Study No.1 performance interface spatial structure.

The study is in ternary form made up of an exposition, development and improvisation. The audiovisual objects are introduced in the exposition which runs from 0:00 to 0:54 (Fig. 5.9). The piece then enters a development section (0:55 - 3:45) where the area around each couple is explored. Each couple in this section is explored in turn. An early performative technique revealed itself here. During the development of *av1* (0:55 - 1:13) it was found that by slowly moving away from the tonic area, the gradual morphing of audio and visuals created a degree of tension.

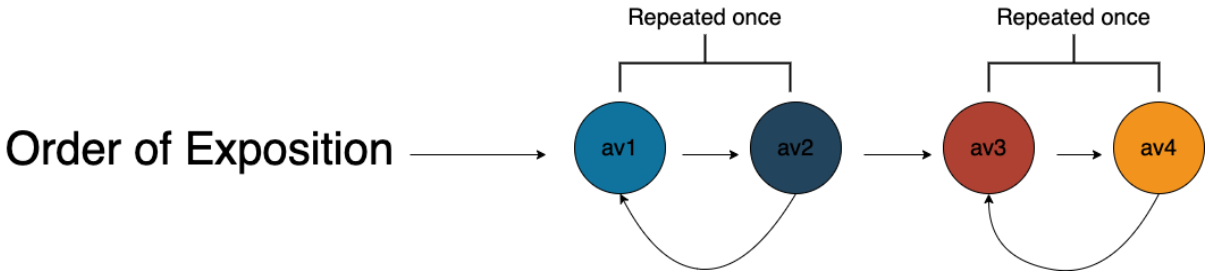

Figure 5.9 Study No. 1 order of exposition.

The term, *tonic area,* is used here to describe the *x/y* location on screen that produces the original audiovisual object. Each couple has a tonic area that is returned to in order to resolve tension. This is an analogy to the concept of the tonic chord in music which indicates the first chord in the key signature. Any movement in the harmony often resolves to this chord. The tension that arises from moving away from the tonic area can be resolved by jumping rapidly back to that point. This technique is called a *jump-resolve* here. This technique is also used during the third section (4:20 -

4:26) using *av3* as the tonic. The third section consists of a free-form improvisation exploring the space between the objects (3:46 - 5:12). A slight shift of focus occurs in this section as the mid-points of each side become the location of exploration. This came about purely through playing with the system and can be seen as a development of the original corner tonic areas. As the mouse travels from the centre-right side of the screen to the centre bottom (3:51 - 3:58), the audio shifts upwards in pitch. This is coupled with a perceptual speed up of the visual movement. It should be remembered that the audio and visuals are not directly connected. This variation in visual speed could be a manifestation of added-value. It is possible that the shift in sonic pitch is affecting the visual speed. Shifting focus like this and discovering the potential added-value effects suggested an extension of the expressive potential of the material and begs the question; what would happen if a subsequent regression model was trained using these new objects? This is explored in the following study.

This study represents an early exploration of the expressive potential in the system. Just using one regression model, it was felt that there was ultimately limited potential for interesting development. Also, the performance structure was quite rigid and simplistic. The improvisation in section three uncovered some interesting audiovisual correspondences when the performance was focused on the mid-points of each side of the interface. The audiovisual objects that were generated when the cursor was at these points were similar to the cadential figures but slightly different. They had a character about them that suggested they were in transition or in-between. For instance, the cursor during the section from 4:18 to 4:31 alternates between the middle-left side of the interface and the bottom-left corner. The audiovisual figure associated with the middle-left side is almost the same as *av1*. However, it is more unstable and can be seen to disintegrate into *av3* when the cursor moves to the bottom-left. The relative temporal movement during this passage is quite satisfying. As the cursor moves towards the bottom-left corner the sound breaks up into staccato type fluttering sounds. The cube shape accordingly stretches out and starts to break apart, with edges flying off on their own.

Although the visuals could benefit from more parameters and the introduction of more regression models could increase complexity and create more varied performance spaces,it was nevertheless felt that the relative expressive range is quite well-balanced *within* the study. The minimal audio texture suits the monochrome wireframe visuals. There is a nice contrast in character between the initial objects also. This facilitated contrasting output, varying from chaotic noise to lighter, quieter moments. The contrast between the passage just discussed, and the following passage from 4:30 to the end is a good example of this contrast. Although the expressive range was well-balanced within the study, the richness of the material could be expanded significantly, whilst remaining balanced.

Study No.2[46]

The aim of this study was to extend certain aspects of *Study No.1*. These include the method of deciding on audiovisual pairs, the placement of initial audiovisual objects and the implementation of a serial chain of regression models. It was hoped that the exploration of these features would reveal promising compositional and performative directions that would contribute to a better understanding of how to compose and perform with the system.

Figure 5.10 Regression models in serial arrangement.

In order to provide access to a wider range of potential material, the system was extended to include the ability to train and save multiple regression models which the user could then access through the performance interface. A suitable way to generate multiple sets of training data was then decided on. A serial arrangement of regression models seemed an appropriate way to initially explore this functionality. This is shown in Fig. 5.10. The development of the pairing procedure illustrated in Fig. 5.11 emerged from the serial structure. Suitable audiovisual objects were initially picked using the random function, *ofRandom()*, triggered using the keyboard, as opposed to manually working out specific audio and visual parameters. Following observation of the randomly generated audiovisual objects, four appropriate examples were chosen for exploration. The spatial arrangement (Fig. 5.12) was then decided on. The input coordinates and output parameters, associated with the chosen objects, were then used as training data for the first regression model (R1).

[46] https://youtu.be/D5f3EFdu4zo (accessed 20/08/2020).

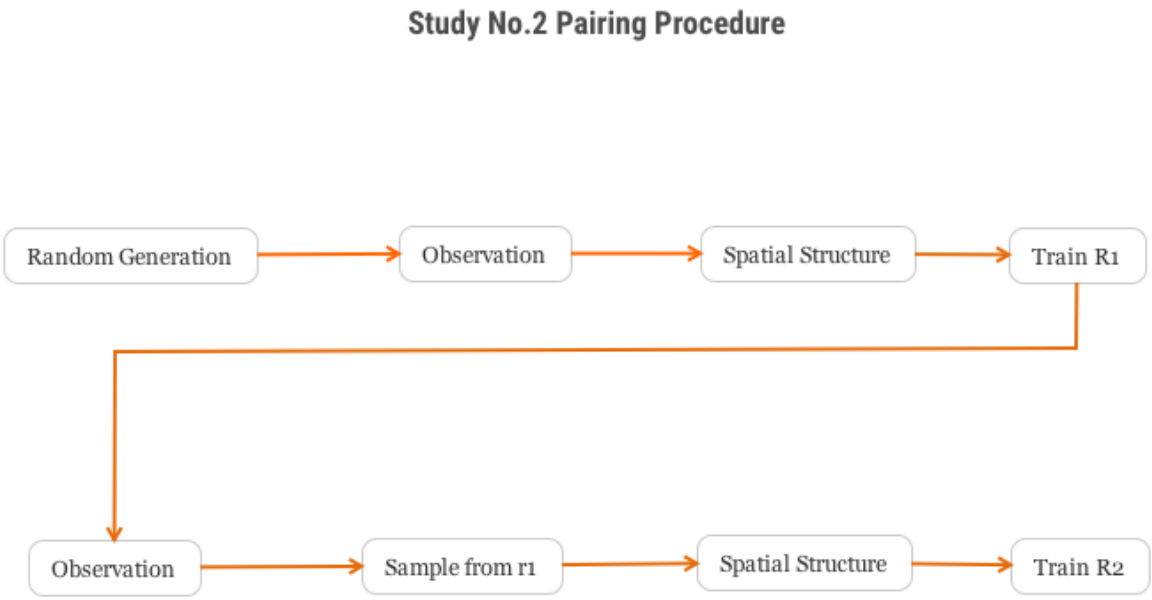

Figure 5.11 Workflow for Study No.2.

After exploring R1, areas in the space were identified that produced appropriate material for use in the second model, R2. The spatial layout of R1 influenced the sampling area explored for new training examples. As the audiovisual objects of R1 were positioned around the midpoint of each side, the four corner areas of the screen became attractive for sampling, as these areas represented the transitional phases of R1. These areas displayed new output from the regression model itself. In this way, the output of R1 was used as the training data for R2. It was hoped that connecting the models in this way would create a continuum within which the material could perceptually evolve. After identifying these rough areas, each one was explored to find suitable audiovisual objects. The centre of the screen was also chosen as a sampling point as the figure at (4:54) presented quite strong cross-modal binding. This meant that R2 was trained with five distinct audiovisual objects. The parameters were then stored and coordinates were chosen which resulted in a spatial arrangement as shown in Fig. 5.13. The coordinates and parameters were then used as training data for R2.

Figure 5.12 Spatial arrangement of initial training objects for the first regression model.

This study is in binary (AB) form with each section split into an exposition, a development and an improvisation. The exposition of each section follows a similar pattern. In section A the objects are presented in turn with repeats similar to *Study No.1*. In section B, the use of a central anchor point means that after each corner couple is presented, the cursor returns to the centre to emphasise its structural importance. This could also be seen as an attempt to strengthen the cross-modal correspondence through repetition. The development sections also follow similar patterns as each anchor point is developed sequentially. Again, in section B the cursor returns continually to the centre point. This exposition-development structure suggested itself naturally in the first study and it is used again here. However, whilst it enables the exploration of all the material, the structure was again quite restrictive and rigid. It may be beneficial to find a more fluid structure within which to develop the material.

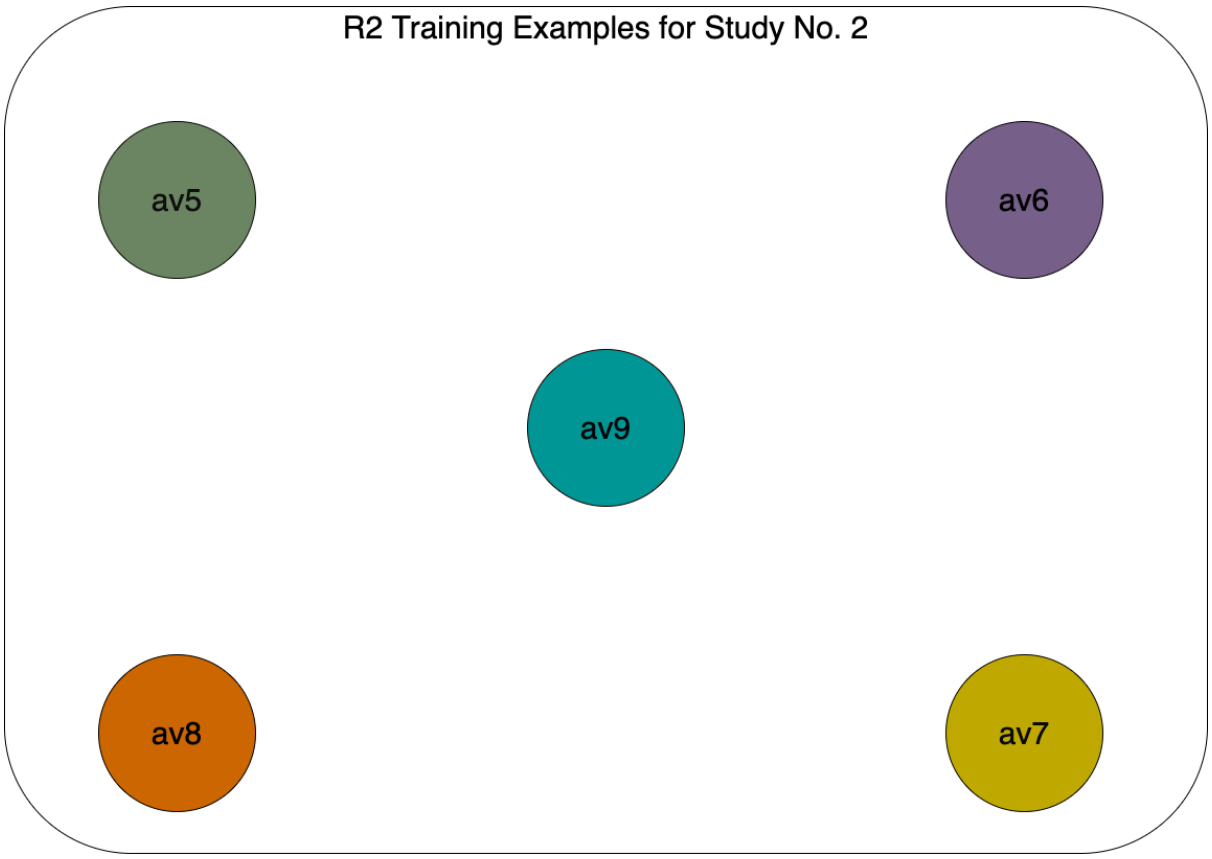

Figure 5.13 Spatial arrangement of training objects for the second regression model.

The parameter randomisation approach gave rise to some unexpected couplings. At first, these appeared to be weak pairings, but closer inspection revealed perceptual correspondences that were not immediately obvious. For example the couple introduced at 0:16 is made up of fragmented visual movement paired with an audio texture dominated by a siren-like sound. Initially, this couple exhibits weak correspondence as the fragmented appearance contradicts the continuous nature of the siren. However, upon observation, a second audio texture, made up of noisy fragmented sound becomes apparent. An implicit correspondence between the oscillating character of the siren and the moderate-speed rotation of the whole mass also emerges. The underlying, fragmented texture binds more explicitly with the visuals than the dominant siren. The nature of its emergence is quite pleasing and adds an emergent depth to the mapping. It also creates a sense of dissonance that could be resolved by introducing a further visual element that corresponds strongly with the siren texture. This suggests a direction for further exploration as a method to create and resolve tension.

The couple, introduced at 0:55 also corresponds in an unexpected way. Here, the object is visually small and relatively solid. The audio consists of a rough grinding texture, with a relatively low-frequency emphasis that initially seems to go against the pairing rationale outlined in *Study No.1* (i.e. that smaller, more solid visual objects correspond to higher frequency sounds). Upon observation, it is apparent that the underlying pulsing character of the audio texture has a similar amplitude envelope

to the motion of the vertices. This similarity in movement across modalities seems to foster a strong cross-modal correspondence suggesting the influence of the Gestalt principle of common fate. This ventriloquism effect is also apparent in the couple at 4:51. This suggests a further technique that can be used within the context of this system whereby added-value could be exploited.

It became apparent through performance that at certain points, the material became unbalanced in favour of either the audio or the visuals. From 3:03 to 3:11 tonal material in the audio seems to dominate the visuals and demand attention. This also happens from 7:42 to 7:58. At 8:28, the appearance of solid cubes becomes a focal point that is returned to a number of times. On reflection, the visuals seem to dominate the audio here as there is a very weak correspondence between them at this point. These events suggest a method for manipulating the sense of audiovisual balance. In this particular context, clear, tonal content, and solid, recognisable shapes, seem to cause the audiovisual balance to become skewed in the direction of the relevant content. This could be an indication that the sense of isolated-structural-incoherence is unbalanced at these points.

Study No.3[47]

Study No.2 expanded on *Study No.1* by introducing the ability to sample from the initial regression model and use those samples as training data for a second model. This sampling and training took place as a preparatory step before the performance of the study.

Study No.3 took an alternative approach to the previous studies by introducing the ability to train and run multiple regression models using the same x/y input. This parallel arrangement of models is shown in Fig. 5.14. In order to achieve this structure, an object-oriented approach was taken for the development of the system. This allowed for the control of multiple regression models, and multiple audio and visual synthesis patches, by the same *x/y* input. This, in effect, creates a parallel mapping architecture as opposed to the serial architecture explored in *Study No.2*. In terms of perceptual output, this architecture allows multiple audiovisual objects to be seen and heard simultaneously. For instance, visually there may be a cube and a sphere being rendered at the same time, along with an audio texture associated with each of the shapes.

[47] https://youtu.be/Abhwrk9O_pw (accessed 20/08/2020).

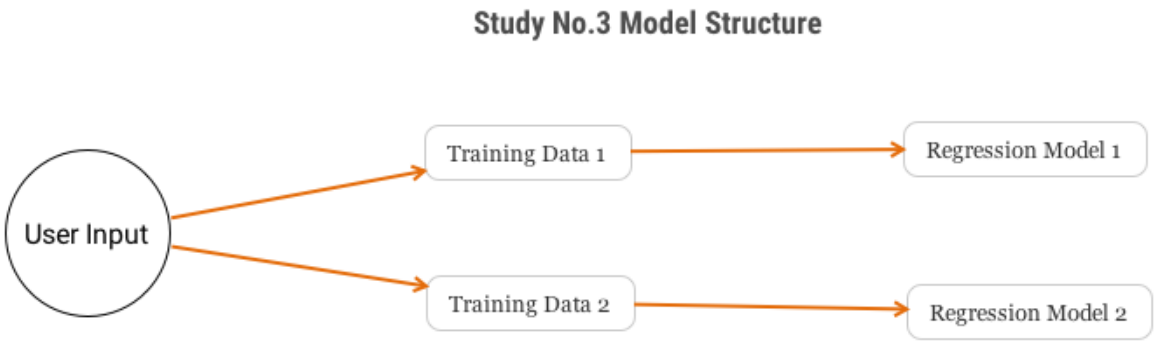

Figure 5.14 Regression models in parallel arrangement.

The initial audiovisual objects were chosen according to the same method described in *Study No.1* above. However, this time there was an extra layer of perceptual correspondence to be considered. Each individual object exhibits cross-modal correspondences within itself. At the same time the parallel presentation of two audiovisual objects means that there are cross-modal correspondences being exhibited between each object. This was carefully considered in the preparation of material. When collecting training data, the coordinates on the user interface now act as training input for a set of two audiovisual objects, as opposed to the single audiovisual objects in studies one and two. To reflect this, the diagrams in Figures 31 and 32 are labelled *cf* for *cadential figure.* This is to indicate that these cadential figures are constructed of more than one audiovisual object. Borrowing from musical terminology to describe a resolving chord progression, these combinations of audiovisual objects are intended to act as resolving points within the composition. It was decided to only utilise two simultaneous objects within this study but this could be expanded. A sphere is used instead of a cube as the primitive shape in order to provide some visual variation. It is unclear as to how many audiovisual objects it would be possible to generate before putting too much pressure on the capabilities of the computer hardware and also on the audio-spectator's perception. This is a direction for further experimentation.

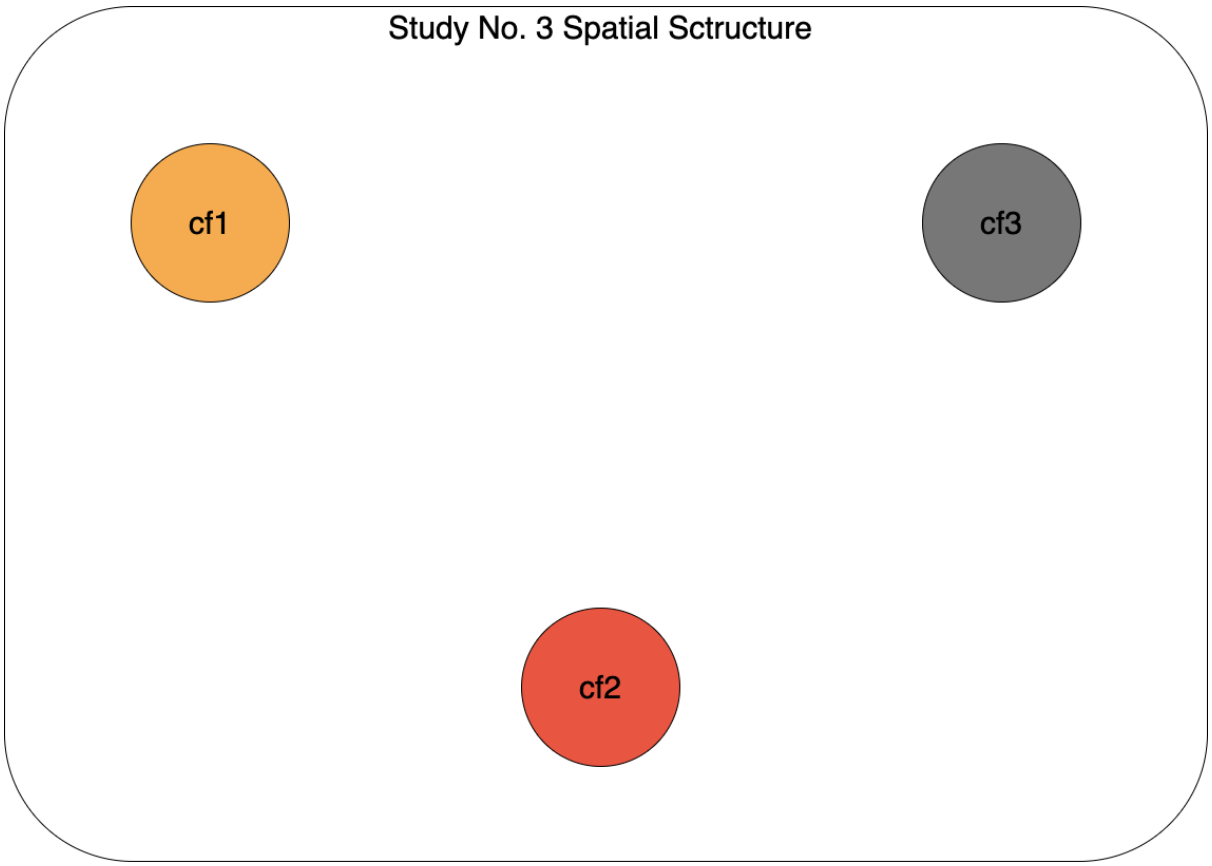

Figure 5.15 Spatial arrangement of initial training objects for Study No.3.

Three cadential figures were arranged within the space (0:00 - 1:34). The diagram above (Fig. 5.15) shows the approximate coordinate locations of the input training data for each of the audiovisual objects. The recording shows the data collection and training steps before the performance starts. From 0:27 to 0:31, *cf1* is prepared. The cursor circles around an area towards the top-left of the performance interface. While this is happening, the GUI element at the top of the interface, titled *recording*, changes to 1. This indicates that the system is storing the *x* and *y* coordinates of the cursor in a training example input vector, and storing both sets of parameters as training example output vectors. From 0:51 to 0:54 *cf2* is prepared. Finally, from 1:12 to 1:15 *cf3* is prepared. Then from 1:21 to 1:32 the system freezes as the neural networks are trained.

A more improvisational approach was taken to the performance of this study as it was felt that the rigid structures of the previous studies were too restrictive. Once the network is trained there is a short exposition section (1:48 - 2:09) where the mouse travels anti-clockwise from the top right corner. Each cadential figure is then explored in an improvisational manner. Throughout this exploration, the area at the top-centre of the screen becomes a focal point that the cursor returns to (3:19 and 4:19) several times. At 3:44 toward the bottom-left of the screen, a visual square becomes another focal

point. On reflection, as in *Study No.2*, the visual material seems to dominate the audio, and perhaps skews the audiovisual balance of the material.

Regarding the initial objects chosen for this study, *cf1* and *cf2* seem to exhibit weak cross-modal correspondences. In the case of *cf1* the solidity of the visual shape again seems to dominate perception. The presence of a strong siren in the audio in both objects also adds to the disconnection. Careful attention should possibly be given to the initial choice of objects to ensure strong cross-modal correspondence. If an audiovisually unbalanced focal point emerges during performance, it could be used as a focus of dissonance or tension. However, the initial cadential objects need to exhibit strong perceptual bonds to form a solid structure within which the material can be explored.

Study No.4[48]

Study No.4 expands on the ideas introduced in *Study No.3*. It again utilises a parallel model architecture, but this time alternates the underlying visual structures between spherical shapes and cuboid ones. Switching forms like this exposes some unexpected areas of interest within the environment. Two simultaneous audiovisual objects are utilised, with three cadential figures laid out in the same spatial arrangement, within the interface, as the previous study. This spatial arrangement is shown in Fig. 5.16. Each cadential figure is again a combination of two audiovisual objects. The initial objects were chosen manually. They are similar to those in *Study No.3* but have been adjusted to exhibit stronger cross-modal correspondences. The first cadential figure (*cf1*) can be seen from 0:02 to 0:05, the second (*cf2*) from 0:06 to 0:10 and the third (*cf3*) from 0:11 to 0:18.

[48] https://youtu.be/soL9XpSpjMs (accessed 20/08/2020).

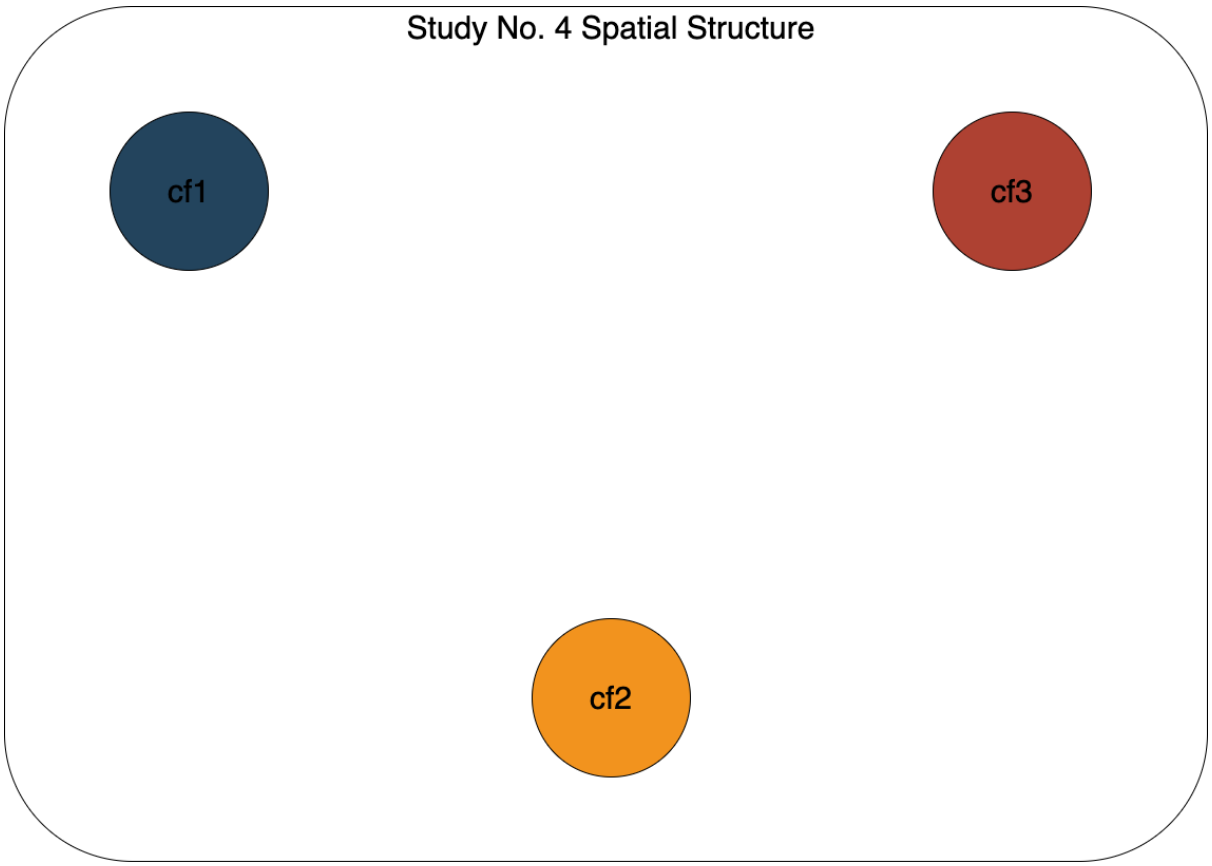

Figure 5.16 Spatial arrangement of initial training objects for Study No.4.

The study is structured in four sections, ABCA'. Section A (0:00 - 1:00) opens with a very short exposition (0:00 - 0:19). This is followed by an improvisational section in which the mouse gradually travels clockwise through the space. This section culminates in rapid jumping between *cf1* and *cf2* as it was found that the contrasting sounds and visual movement were aesthetically pleasing.

Section B (1:01 - 3:00) begins after the underlying visual primitive of one of the models is changed from a sphere to a cube. Here, the cube is coloured white whilst the sphere is red. As the space is explored, even though no new training has taken place, the perceptual interaction between the models has changed significantly. At 1:18 a new focal point (*fp1*) is found in the bottom-left portion of the screen. This becomes the primary centre of focus for this section. The emergence of these new focal points is shown in Fig. 5.17.

Figure 5.17. Emergence of new focal areas through improvisation.

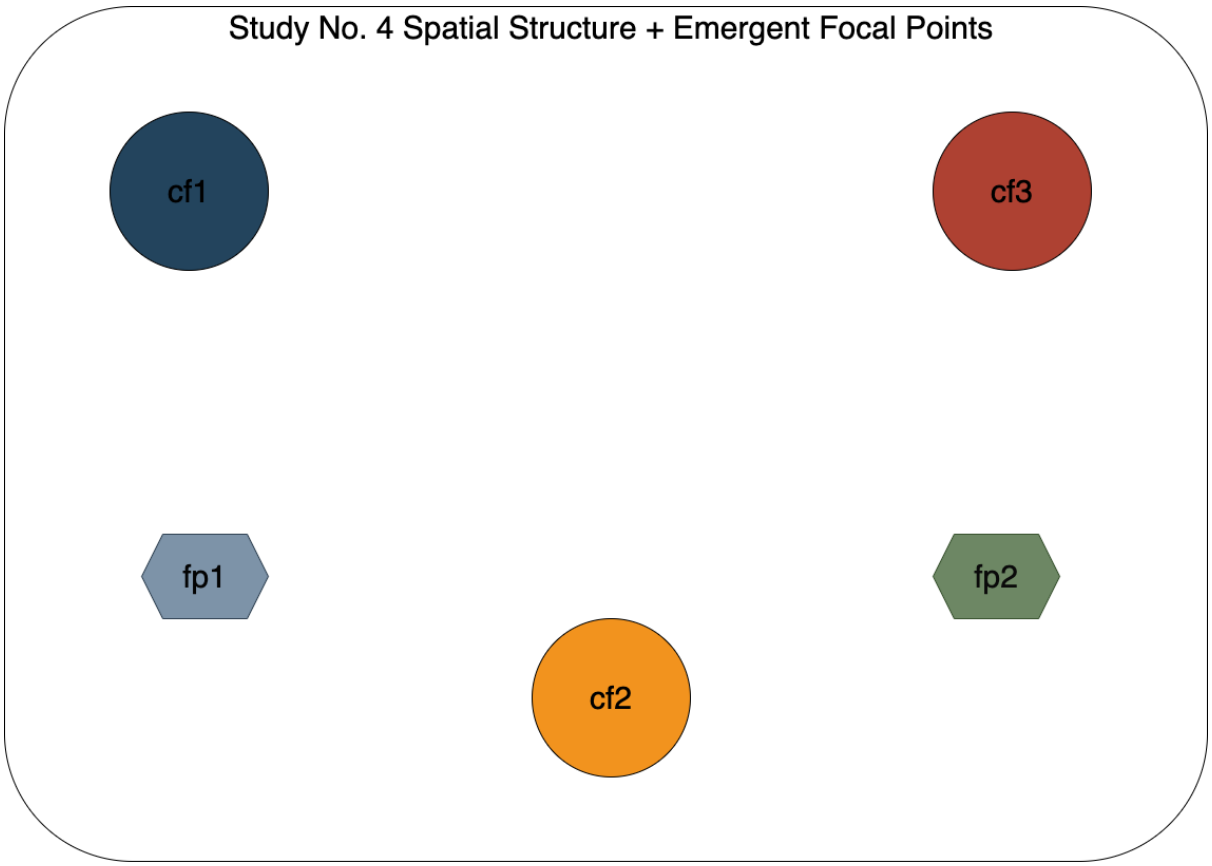

Figure 5.17 Emergence of new focal areas through improvisation.

On reflection this centre of focus is heavily influenced by the symmetry of the visual form rather than the audio texture. As above, this tendency for strong regular shapes to dominate perception should be utilised carefully throughout the piece, contrasting them with the more audiovisually balanced initial objects. This area is explored before rapidly jumping across the screen to again accentuate interesting contrasting textures. At 2:14 another interesting area (*fp2*) is found along the right-hand side of the screen. Again, upon reflection, it was realised that within this area the audio seems to dominate the perception rather than the visuals. Although this does represent an audiovisual imbalance in the material, this type of strategy, moving from a visually-dominated area to a sonically-dominated area before transitioning to a strong cadential figure as a resolution, could prove effective in practice. The section closes with rapid jumping between areas coming to rest at *fp1*.

Section C begins at 3:01 where the models are reversed visually. What had been the white cube was now a white sphere. Conversely, what had been the red sphere was now a red cube. The cursor gravitates towards the right-hand side of the area for most of this section where the red cube is highly unstable. This corresponds strongly with a disjointed glitchy audio texture. At 3:56 the cursor begins jumping rapidly between areas, mainly coming to rest at *cf1*. At 3:54 an interesting visual figure is found, again in the bottom-left of the screen, that is briefly explored before entering the final section

at 4:00. This final section (4:02 - 4:39) returns to the double sphere visual formation. This time *cf1 is used* as the anchor from which the cursor rapidly jumps across to other areas.

This study mainly consists of an improvisational exploration of the material. It was found that switching the visual shape extended the possibilities of the material even though the audio remained the same. This is a strong demonstration of the power of cross-modal correspondences and their perceptual binding power. A future extension of this might be to implement multiple audio patches whereby the audio can be switched just like the visuals were here. It was felt that each section naturally guided the performance. For example, the emergence of the new focal points in section B dictated the exploration of that area to a large extent. Perhaps a combination of this natural approach with the structured approach of the first two studies could provide a balanced framework within which pieces like this could be composed and performed.

5.5 Reflection

Throughout the development of the system, a log was kept, containing notes of ideas and observations related to the performance of the system. Potential compositional and technical parameters were identified that could be experimented with. In this way, possible avenues of exploration were identified. When it was time to test the performance of the system a study would be planned, performed and recorded. After each one, the recording was analysed, paying close attention to the perceptual aesthetic output of the system. It was found that this reflective analysis usually revealed interesting perceptual features that were not apparent when performing. Following these analyses, technical improvements would then be implemented, followed by another study. The following technical and compositional observations emerged from this cyclical process.

Technical:

- The ability to sample from models and save trained models can extend the creative potential of the system.
- The use of serial and parallel regression models provides interesting structures within which to develop material.

Compositional:

- The jump resolve technique can be useful to resolve tension.

- Compared to the manual selection of initial audiovisual objects, random selection can provide unexpected results. However, the selected objects need to exhibit strong cross-modal correspondences, as they will form the basis of the structural material.
- Perceptual ventriloquism can form the basis of strong cross-modal correspondences.
- The arrangement of the initial objects within the interface can help to inform the structure of the performance. However, the performance structure can freely change whilst maintaining the same interface layout. .
- A mix of improvisation with structural points leaves room for exploration of the parameter space whilst maintaining a loose compositional structure.
- New focal points may emerge through exploration of the area. These emergent focal points can be audiovisually unbalanced.
- The layered character of the audio textures can create partial correspondences with the visuals, initially exhibiting an incongruent or dissonant correspondence that resolves into partial congruency or consonance after observation.
- Audiovisual balance can potentially be skewed by solid, regular shapes and strong tonal content in some contexts.

The aesthetic and technical realisation of the system was influenced by the compositional principles described in Chapter 3. The lack of hard-coded data mappings between the audio and the visuals allowed for experimentation with both strong and weak audiovisual correspondences. In specifying audiovisual objects that appear to exhibit strong cross-modal correspondences, the intention was to create areas within which the audiovisual material was well balanced. By moving in and out of these areas the intention was to manipulate the experience of audiovisual balance, like a music composer would use consonance and dissonance to create tension and release within a piece of music.

With regard to the perception of audiovisual balance, the use of a data-driven system aims to eliminate a dominant/dependent relationship between the audio and the visuals. By using neural mapping as the sole technical bond between the audio and visuals, a nonlinear, complex relationship between the input and output was created. This aims to confound causation (Sá, Caramieux and Tanaka 2014a), creating audiovisual correspondences that are opaque. However, in choosing audiovisual objects that exhibit tight perceptual bonds, enough transparency is introduced, aiming to create some sense of causation even if that cause/effect relationship is not entirely clear. This mapping strategy follows work done by Sá, Caramieux and Tanaka (2014) and by manipulating the opacity and transparency of the mappings the intention is to attempt to gain control over audiovisual balance through live performance. In some contexts, it was found that solid regular shapes seem to skew audiovisual

balance toward the visuals, whereas strong tonal content seems to skew the balance towards the audio. This knowledge could be used by the artist to manipulate the audience's perception during performance.

5.6 Future Developments

The system presented above has been shown to be a suitable control paradigm within which audiovisual compositions can be explored. There are many avenues for further exploration within this context.

The method of choosing the initial audiovisual objects could be refined. Rather than simply randomising the parameters it might be useful to implement an interactive genetic algorithm where audiovisual objects are selected and their features are used to narrow down the parameter space. The method of controlling the system was not developed at all. Above, the laptop trackpad was used. However, the system will accept any type of input. An exploration of gestural devices may be fruitful, as such an approach could lead to nuanced, expressive real-time performances. The extension of larger serial and parallel chains of regression models presents many possibilities for further exploration. The flexibility of the RapidLib library would allow many configurations leading to increasingly dense and layered connections between the audio and visuals.

The system does not depend on any certain type of visual or audio aesthetic. This allows vast potential for experimentation and development of numerous aesthetic choices. The audiovisual artist is fully empowered to explore whatever combination of audio and visuals they see fit. The visuals in particular above have not been developed extensively. They are very basic and minimal. Many more visual parameters could be added to create more complexity such as control of the camera, the use of textures and GLSL shaders. The audio aesthetic could also be extended in many ways to create richer and more varied sound worlds. Finally, the above approach could also be combined with media-driven techniques such as FFT analysis of the audio. If handled correctly, this could create a desirable balance between transparency and opacity that would create enough causation to keep the audience engaged whilst also providing enough complexity to sustain interest.

The studies presented here are initial explorations of how to compose and perform with the system. The next step is to test the control paradigm in a full performance context. Two performances will be explored in Chapter 6, where the system is used as the basis for the original compositions *Ventriloquy I* and *Ventriloquy II.* The performance contexts range from a single-screen live performance with

stereo sound, to a performance in a shared immersive space with 360 projection and surround sound. This provides a stepping-stone to realising the system within a fully-immersive VR environment. The development of this environment, and the integration of the neural network control paradigm, will be discussed in detail in Chapter 7.

Chapter 6 IML Control of AV Compositions in a Live Performance Context: *Ventriloquy I* and *II*

Chapter 5 introduced the Neural AV Mapper, which allows the quick mapping of many parameters between an input source and audiovisual material. This tool provides an intuitive way to explore audiovisual parameter spaces. This chapter will discuss the two original compositions *Ventriloquy I* and *Ventriloquy II*. They represent a practical implementation of the IML approach to controlling audiovisual compositions. Here, this method is being used to control these compositions in a live performance context. The source code for the two pieces discussed in this chapter can be found at their respective GitHub repositories or in the accompanying media pack at *sourceCode/ventriloquy1* and *sourceCode/ventriloquy2*. The example videos and performance footage can be found in the accompanying media pack at *mediaFiles/ventriloquy*. The videos have also been uploaded to YouTube and Vimeo. The links are provided as footnotes in the text.

6.1 Live Audiovisual Performance

As discussed in Chapter 2, the performance of live audiovisual material is conceptualised as an act that is analogous to the performance of music. There are many different styles of music and many of them can be performed live. Similarly there are many different styles of audiovisual art and many of them can be performed live. The liveness is a constituent part of the aesthetic of many styles rather than being a style in and of itself. The separation of practice and context was discussed in Chapter 2. The work in this chapter is generative visual music presented in a live context. This section will look at some of the issues involved in creating, and some desirable characteristics of, audiovisual performance tools.

Correia and Tanaka (2014) identify the following concepts that apply to the design of instruments and the practice of live audiovisual performance:

- Expressivity
- Flexibility
- Ease-of-use
- Audience involvement

They state that these ‘concepts can be useful for audiovisual performers, or designers of tools for audiovisual performance’ (Correia and Tanaka 2014: 98). Expressivity relates to the ability for the performer to express themselves through their chosen medium. Flexibility refers to the ability of the performer to adjust and customise their interface or software. Ease-of-use refers to how easy it is to pick up the tool or software and use it. This factor refers to commercial, or off-the-shelf, software that is targeted at a wide range of users, and tailored to workflows that are easy for people to use. Audience involvement refers to two concepts; ‘the importance for some artists of conveying the liveness of the performance to audiences; and how to have audiences participate in the performance’ (ibid. 2014: 98).

Bergstrom and Lotto highlight the lack of expressive instruments available for the performance of visual music (Bergstrom and Lotto 2016). They identify the embodied way in which musical instruments are used as a desirable characteristic for expressive controllers. Musical instruments encourage the musician to develop ‘advanced enactive knowledge: knowledge that can only be acquired and manifested through action’ (ibid. 2016: 399). The meaning of the term, *enactive,* is grounded in Varela, Thompson and Rosch’s usage, which emerged from the field of philosophical hermeneutics.

> The term hermeneutics originally referred to the discipline of interpreting ancient texts, but it has been extended to denote the entire phenomenon of interpretation, understood as the enactment or bringing forth of meaning from a background of understanding. (Varela, Thompson and Rosch 2016: 149)

A point that Bergstrom and Lotto touch on, is that the audience will react differently to a live performance depending on whether or not the performer is ‘employing advanced enactive knowledge’ (Bergstrom and Lotto 2016: 399). Here, they are stating that the audience shares an empathetic relationship with the performer, mentally mimicking their actions and visually processing the performance. This experience is changed when the controller is not a musical instrument, but rather a computer interface, or digital controller consisting of knobs and faders.

It could be argued that advanced enactive knowledge, as understood according to Varela, Thompson and Rosch, can be demonstrated with what Bergstrom and Lotto call ‘nonmusical controllers’ (ibid.

2016: 399). Examples of this would be DJs mixing tracks live using turntables, and the live performance of hip-hop and electronic music using controllers such as the Akai MPC, which has been described as ‘a bona fide, playable instrument with its own criteria for virtuosity’ (Brett 2016: 88). This capacity for virtuosity has been identified as an important factor in facilitating expression (Dobrian and Koppelman 2006: 279). It is clear then, that the method of controlling the material must allow the performer to be expressive, whether using a musical or non-musical device.

6.2 *Ventriloquy I*

Figure 6.1 Still from Ventriloquy I (2018).

Ventriloquy I was created using the Neural AV Mapper. It is a piece for single-screen live performance using a camera and the PSMove controller. The title is a concatenation of two words; ‘ventriloquist’ and ‘soliloquy’. It is the first of a two-piece suite of compositions. It was performed live at Seeing Sound in 2018. The source code can be found at the GitHub repository.[49] Fig. 6.1 is a still from the composition.

6.2.1 Motivation

[49] https://github.com/bDunph/ventriloquy1 (accessed 21/08/2020).

This piece emerged from the studies described in Chapter 5. It is intended to consolidate the earlier research with regression mapping, and attempt to create a piece of art informed by that work. There was also a desire to explore some of the theoretical concepts around audiovisual balance.

It was felt that the Gestalt principle of common fate might be a useful way to explore the concepts related to audiovisual balance. This was also demonstrated in the studies from Chapter 5. By utilising noisy movement in both the audio and visuals, the principle of common fate could be relied upon to create a perceptual bond in the audience's minds. It is this ventriloquism between audio and visual media streams that forms part of the title of the piece. Chion also acknowledges the effect of Gestalt principles when he states that the 'phenomenon of significant synchretic points generally obeys the laws of gestalt psychology' (Chion 1994: 58).

Within this piece, the Neural AV Mapper was used to explore the phenomenon of cross-modal Gestalt grouping as a perceptual binding foundation. It was hoped that this approach would foster a sense of isolated structural incoherence. As with the four studies, it was hoped that the indirect perceptual mapping between audio and visual parameters would create a subtle shifting sense of audiovisual balance as the material moves in and out of synchronicity. The use of repetition, through repeated audiovisual associations, was also explored, to further strengthen the cross-modal perceptual binding of the predefined audiovisual objects. Chion suggests that unrelated images and sounds have the potential to exhibit synchresis through repeated exposure.

> Certain experimental videos and films demonstrate that synchresis can even work out of thin air - that is, with images and sounds that strictly speaking have nothing to do with each other, forming monstrous yet inevitable and irresistible agglomerations in our perception. The syllable fa is heard over the shot of a dog, the sound of a blow with the sight of a triangle. Synchresis is Pavlovian. (Ibid. 1994: 63)

Ultimately, with *Ventriloquy I,* a core aim of the piece was the exploration of the Pavlovian characteristic of synchresis and how noisey rhythms in both modalities could align through the Gestalt principle of common fate. To achieve these aims, the Neural AV Mapper was utilised to perform regression analysis between input data from a controller and output data consisting of computer-generated audiovisual parameters.

6.2.2 Development of Material

In his *Estuaries* (2016 - 2018) series of audiovisual compositions, Bret Battey describes his compositional approach as 'audiovisualisation-assisted composition' (Battey 2020: 278). He generates material using 'variable-coupled map networks (VCMN)' (ibid. : 270) and notes that 'a great deal of editing of audiovisual results were required to achieve' (ibid. : 273) output that he found acceptable. The term audiovisualisation is taken from Ryo Ikeshiro's approach that was discussed in Chapter 2, where he generates the audio and visual material from a common source of data. Battey proposes that with his assisted method of composing, 'sensitive artistic perception remains essential for guiding selection and adaptation of algorithmically generated materials to achieve the delicate task of balancing unification and independence of materials' (ibid. : 278).

The approach to the composition of *Ventriloquy I* could also be described as being 'assisted'. Although an audiovisualisation approach is not being employed in this piece, mappings are being algorithmically generated, between the performer's input and the audiovisual material, through the use of a neural network. However, to ensure the mappings foster appropriate levels of dramatic expression in the material, there are perceptual mappings being arranged, between the audio and visual material, that are based on human observation during material preparation. This will be discussed in more detail below. Further, the perception of balance in the abstract audiovisual material is of interest here.

The system used to map the generated material was outlined in Chapter 5. However, before mapping any material, the potential for more aesthetic depth and richness of material, compared to the studies, needed to be generated. A new data input method, that fostered expressivity and felt natural in performance, also needed to be explored.

Visual Aesthetic

The visual shapes that feature in the initial studies are simple wireframe renders of cubes and spheres. Firstly, the expressive range, of both the audio and visual material, needed to be expanded. In order to do this, the faces of the shapes were rendered and more colour was introduced into the visual field. The first attempts at this can be seen in the *Neural AV Mapper - Colour Example*[50] video. The video file can also be found in the media pack under *mediaFiles/ventriloquy/ventriloquy1/neuralAVMapper_colourEx.mov*. This video demonstrates a

[50] https://vimeo.com/250616028 (accessed 02/06/2020).

cube and sphere combined structurally. The colours seemed to work well together and along with the shape's textures, were reminiscent of some of the classic visual music work of Mary Ellen Bute (*Colour Rhapsodie*, 1948) and Jordan Belson (*World*, 1970). As discussed in Chapter 2, the work is located in the generative visual music area of the wider audiovisual field. Therefore, this aesthetic seemed appropriate. The basic visual shapes consist of a cube and a sphere. These structures were generated using the basic *ofBoxPrimitive* and *ofSpherePrimitive* native openFrameworks functions. It was decided to use the native primitive functions as they have in-built members to allow for simple texture mapping. As shown in Ex. 6.1, the textures are initialised in *avObject::avSetup()*.

```
00 normMap.load("textures/normMap.jpg");
01 normMap.getTexture().setTextureWrap(GL_REPEAT, GL_REPEAT);
02
03 ofTexture tex;
04 unsigned char texData[256 * 256 * 4];
05 unsigned char * texPtr;
06
07 tex.allocate(256, 256, GL_RGBA);
08 tex.setTextureWrap(GL_REPEAT, GL_REPEAT);
09
10 for(int i = 0; i < 256 * 256 * 4; i++){
11           texData[i] = 0;
12 }
13 texPtr = &texData[0];
14
15 tex.loadData(texPtr, 256, 256, GL_RGBA);
```

Example 6.1 Initialisation of textures.

The code above sets up the texture objects to be used in rendering. At lines 00 and 01 *normMap*, an object of type *ofImage*, is loaded with *normMap.jpg*. This is a normal map that creates the illusion of

a bumpy surface. They are also called bump maps. See the example video *Ventriloquy (excerpt1)*[51] (labelled *Ventriloquy_ex1* in the accompanying media file under *mediaFiles/ventriloquy/ventriloquy1*), which shows an orange sphere with the bump map rendered. There was a desire to move away from the flatness of the basic structures and it was found that adding surface details like this created a pleasing aesthetic, giving the otherwise simple objects more aesthetic depth. The following code from line 03 to 15 sets up an *ofTexture* object called *tex*. The array *texData* will hold colour values for the texture accessed through *texPtr*. As shown in Ex. 6.2, the texture colour is updated within *avObject::visual()*.

```
00 for(int k = 0; k < 256 * 256 * 4; k+=4){
01     texData[k+0] = texColR;
02     texData[k+1] = texColG;
03     texData[k+2] = texColB;
04     texData[k+3] = texColA;
05 }

06 texPtr = &texData[0];
07 tex.loadData(texPtr, 256, 256, GL_RGBA);

08 box.mapTexCoordsFromTexture(tex);
09 sphere.mapTexCoordsFromTexture(tex);
```

Example 6.2 Update texture colour.

The *for* loop at line 00 iterates through every fourth element of the *texData[]* array. The texture is formatted as RGBA, so the variables *texCol** from line 01 to 04 are assigned to the relevant elements in the array. These variables are part of the GUI and are connected to sliders that are updated each frame. In this way, it was possible to interactively adjust the colour of the shape during the preparation phase. At line 06, *texPtr* points to the address of the first element of *texData[]*. The function *loadData()* is then used to apply the values of the array to the *tex* object. Lines 08 and 09 use the function, *mapTexCoordsFromTexture()*, to map the texture coordinates to the *box* and *sphere* primitives. The texture is then bound in *avObject::drawVisual()*, where the sphere and cube are both

[51] https://vimeo.com/251820270 (accessed 11/09/2020)

drawn. Here, the forms are drawn using the *noise_light_bump* shader. This shader draws the bump map, applies Phong lighting and applies a Perlin noise algorithm to the vertices. A more sophisticated noise algorithm, than the one used in the studies, is used here. It was found that the Perlin noise algorithm gives a more natural character to the movement of the vertices than the use of the basic *ofRandom()* function.

Audio Aesthetic

As mentioned above, one of the main themes being explored in this piece is the experience of cross-modal ventriloquism. This seemed to work quite well in the initial studies using noise-based visual movement with rough saw and square wave FM synthesis. It was decided to continue exploring this audio aesthetic. However, the texture of the sound needed to contain more depth and richness. In order to achieve this, FM-synthesised tones were mixed with some inharmonic additive partials. This is shown in Ex. 6.3. In practice, this is not pure FM synthesis. It would instead be more accurate to say that aspects of the patch are influenced by FM synthesis. This is why some of the parameters listed in Fig. 6.3 are not standard FM parameters.

```
00 square1 = squareOsc1.square(oscilFreq + modulator);

01 square2 = squareOsc2.square((oscilFreq * 1.17) + lfo) + square1;

02 square3 = squareOsc3.square(oscilFreq * 1.69) + square2;

03 square4 = squareOsc4.square(oscilFreq * 2.04) + square3;

04 subSig1 = ((square1 + square2 + square3 + square4) * (oscilAmp *
mappedVol)) / 4;
```

Example 6.3 FM and additive partials.

The first signal's frequency is modulated by the variable *modulator*. This *modulator* signal is audio-rate and is an extrapolation of the *modulator* variable detailed in Ex. 5.7 in Chapter 5. The resulting signal, *square1*, is then added to another frequency-modulated signal, *square2*. Here, the *lfo* variable acts as an audio-rate modulator. This, in turn, is added to *square3*, which is then added to *square4*. The frequencies of signals 2, 3 and 4 are multiplied by arbitrary floats. These float values emerged through experimentation and observation of the resulting sound. At line 04, the signals are then

summed together again to create *subSig1*. This final summing has the perceptual effect of masking high-frequency, pitched tones that are quite overpowering, with noisy textures. For an example of this, see the files *Vent1_Ex6_3_WithoutSum.m4a* and *Vent1_Ex6_3_WithSum.m4a* in the accompanying media pack under *mediaFiles/ventriloquy/ventriloquy1*. This signal is also multiplied by *oscilAmp* and *mappedVol*. These control the amplitude and distance-related volume of the signal. Each object's volume is mapped to its distance from the camera. It was hoped that the combination of synthesis techniques would provide more variety and depth than was present in the audio of the initial studies. The signals described above represent the main signals generated for the soundworld. They are then processed through various filters and envelopes to further shape the sound. The audio signal path is described in Fig. 6.2.

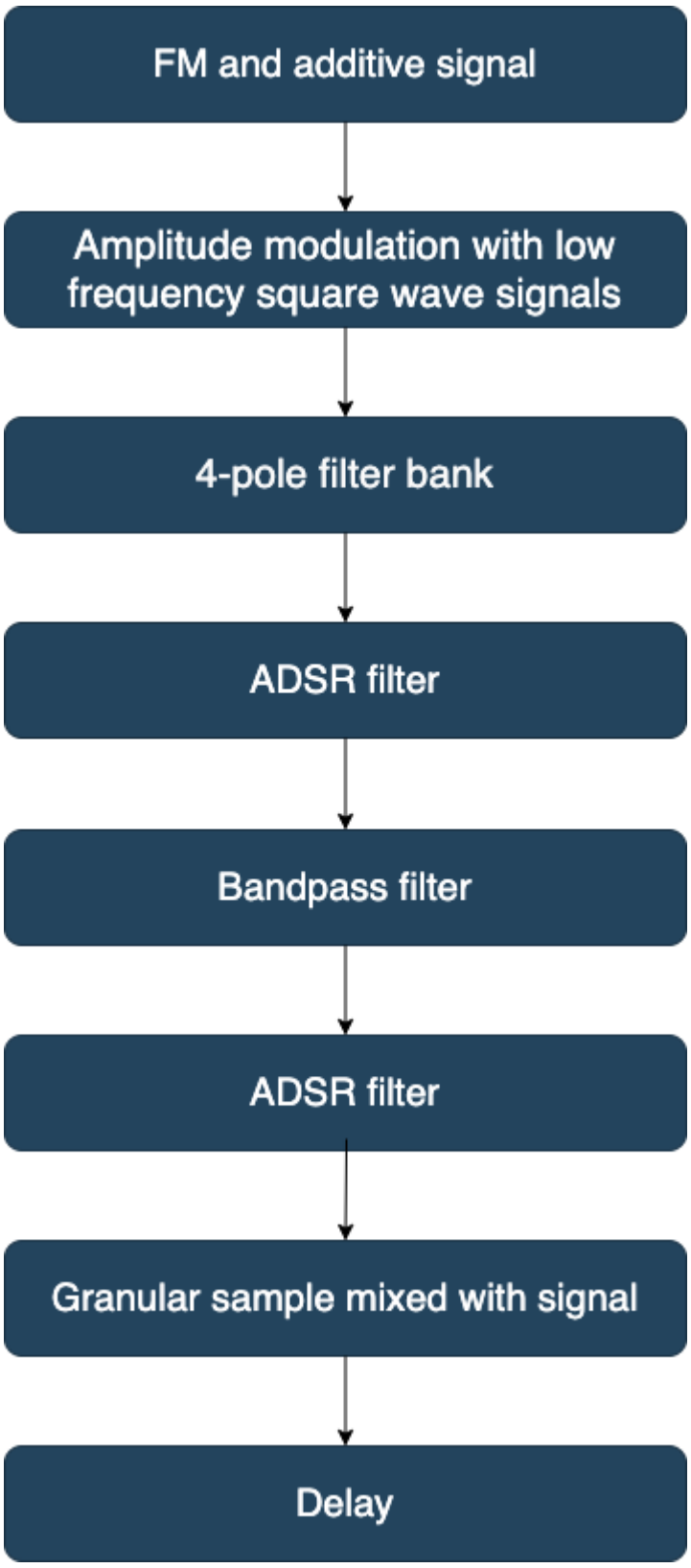

Figure 6.2 Ventriloquy I audio signal chain.

The amplitude modulation included in the figure above was intended to add some further low frequency complexity to the soundworld. After the second ADSR filter, there is a simple granular synthesis patch that takes a sample file called *lowRumbleLoop.wav* as input. The granular signal is then mixed with the filtered signal before being sent through a delay loop. The granular signal is very subtle. On reflection, it is perhaps too subtle, as it gets overwhelmed by the filtered signal in the final performance. However, the intention in adding it here was to give the signal some more substantial low-end frequencies.

Mapping

The most significant method of mapping in this composition was the use of several neural networks to map the controller input to a range of audio and visual parameters. These parameters are listed in Fig. 6.3.

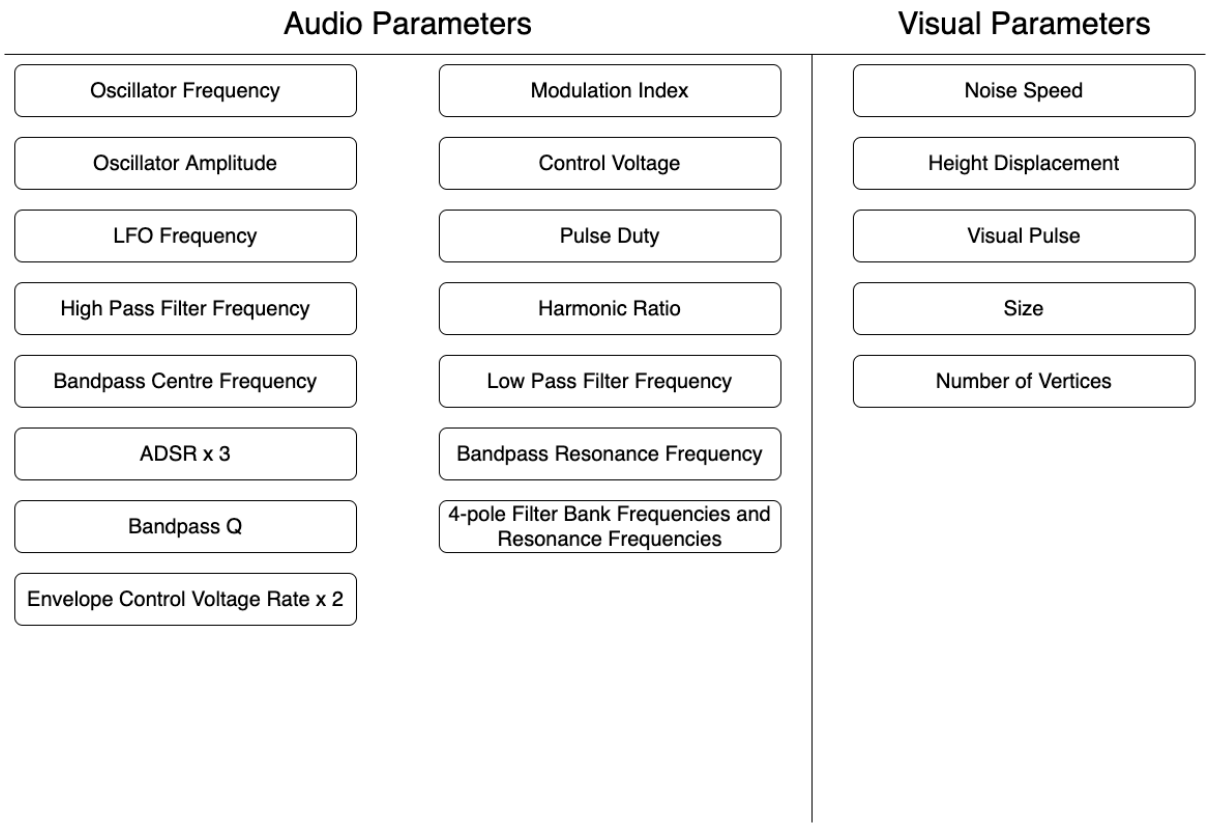

Figure 6.3 Parameters used as outputs from neural networks.

These parameters represent the output values of the neural network. The inputs to the network consist of the x, y and z cartesian coordinates of the PSMove controller. The controller will be discussed in

more detail in the next section. The rotation of the PSMove controller is also mapped to the visual *blur amount* and the audio *delay amount*. As the rotation values of the controller increase, the blur amount increases, and the wet delay-signal increases. This results in a perceptual mapping between blurry visuals and delayed audio.This was achieved by taking the rotation values from the PSMove controller, and mapping them linearly, in the code, to the audio patch and fragment shader simultaneously. The transition from blurry/wet delay effect, to a sharp/dry signal can be observed in example video *vent1_2ndDraft_mod0_av0* from 0:13 to 0:17 seconds. This file can be found in the media pack under *mediaFiles/ventriloquy/ventriloquy1/2ndDraftMaterial/model0*. It was felt that this correspondence made sense conceptually: the more blurry a visual object, the less precise it is. Similarly, in sonic terms, the more delay that is heard, the less crisp an audio source is. However, in practice it was not as effective as hoped. This point will be discussed further in the feedback section below.

Further manual mapping was necessary due to the fact that 3D shapes were being utilised. It was felt that localising the audio source within the scene would be beneficial to give the audience a sense that the shapes were, in fact, sound-generating objects. There was a desire to give the impression that the sound and images were unified. In pursuit of this goal simple distance and panning functions were implemented that attempted to localise the sound in terms of distance from the camera and stereo pan. This technique worked surprisingly well and gave a sense that the sound source belonged to the visual shape as it moved through the 3D space. The effect can be observed in the example video, *vent1_2ndDraft_mod1_av3.mov,* from 0:26 seconds to the end of the clip. This file can be found in the media pack under *mediaFiles/ventriloquy/ventriloquy1/2ndDraftMaterial/model1*. As the object moves from side to side, the sound follows. The volume of the sound also decreases as the object moves further back into the scene. The panning function can be seen in Ex. 6.4.

```
00 double * avObject::panner()

01 {

02     objPos_camSpace =  camModelViewMat * ofVec4f(xTrans, yTrans, zTrans,
1);

03     objPos_asPercentage_ofX_axis = ofMap(objPos_camSpace.x, -1440, 1440,
0.f, 1.f);

04     panOut = &objPos_asPercentage_ofX_axis;

05     return panOut;

06 }
```

Example 6.4 Audio source panning.

The object position is calculated in camera space using *camModelViewMat*. The *ofVec4f* vector, made up of *xTrans*, *yTrans* and *zTrans*, is implemented to account for the fact that the object may not be at the origin. The *x* value of *objPos_camSpace* is then mapped from the screen dimensions to the range, 0.0 to 1.0. This value is treated as a percentage of the x axis taken from left to right on the screen. The left side of the screen is 0.0 and the right side is 1.0. This value is then returned from the function as a pointer of type *double*. This return value is then used in *ofApp::audioOut()* to pan the sound source using a *maxiMix* object. The distance calculation is performed in *avObject::visual()* and is shown in Ex. 6.5.

```
00 //***** object sound level related to dist from camera and size *****
01 const double soundPowerThresh = 0.000000000001;
02 soundPower = ofMap(shapeSize, 50, 1050, soundPowerThresh, 0.01);
03 objPos = ofVec3f(xTrans, yTrans, zTrans) * camModelViewMat;
04 ofVec3f absCamPos = ofVec3f(abs(camPos.x), abs(camPos.y), abs(camPos.z));
05
06 //***** distance between object and camera*****
07 dist = absCamPos.distance(objPos);
08 if(dist <= shapeSize){
09   dist = shapeSize;
10 }
11 soundArea = 4*PI*pow(dist, 2);
12 soundIntensity = soundPower/soundArea;
13 logCalc = soundIntensity / soundPowerThresh;
14 dbVal = abs(10*log10f(logCalc));
15 mappedVol = ofMap(dbVal, 0.f, 100.f, 0.f, 1.f);
```

Example 6.5 Distance to volume mapping.

Here, the value of *soundPower* is calculated based on the *shapeSize*. There was a desire to give larger shapes more volume to align with expected cross-modal perceptual correspondences. The position of the object is then calculated relative to the view matrix. The distance between the camera and the object is calculated using the *distance()* function. A sound area is specified within which the decibel level will be calculated. The sound area is based on the value of *dist* which is the distance between the camera and the object. The camera is at the centre of this area which means the larger the area the smaller the value of *soundIntensity* at line 12. The value of *soundIntensity* is divided by the constant sound threshold value at line 13. The decibel level is calculated at line 14 according to the decibel formula ‘dB = 10log(W / Wref)’ (Siemens 2019). Finally, this value is mapped to a value between 0.0 and 1.0. This value is then used to attenuate the audio signals as shown in Ex. 6.3, line 04.

Controller

The method of control was an important element in the realisation of this piece. There was a desire to move away from the two dimensional control used in the previous studies. It was felt that extending the parameter control space into three dimensions would allow for more variation in the material and would hopefully lead to a more interesting parameter space.

There are many popular methods available for real time control of generative material. Some of these include the use of sensors like the Microsoft Kinect to control musical rhythm and pitches using gestures (Şentürk et al. 2012), or the electromyogram (EMG) to control both music and visuals using muscle tension[52]. Atau Tanaka has used EMG sensors extensively in his research and live performance practice (Tanaka and Ortiz 2017). Other methods employ physical controllers such as the Gametrak controller (Freed et al. 2009) which was originally released as a novel controller for the XBox and Playstation2. This controller was utilised in ensembles such as PLOrk (Princeton Laptop Orchestra) and DubLOrk[53] (Dublin Laptop Orchestra). The PSMove Motion Controller[54] is a device that was released with the Playstation3 and contains several sensors that can be used to measure orientation and acceleration. It can also be used with a camera to give location data and is freed from the constraints of the Playstation3 console by way of the PSMove API (Perl, Venditti and Kaufmann 2013). Each of these options were considered when approaching the real time control of *Ventriloquy I*.

Ultimately, the PSMove controller was chosen because it provides access to three dimensional positional data and also orientation data. Finally, it has a certain amount of tactile qualities that may be appropriate for live performance. These include physical buttons and an analog trigger. The controller also fulfils the criteria outlined by Correia and Tanaka for live performance systems: expressivity, flexibility and ease-of-use (Correia and Tanaka 2014). Once the controller had been decided upon, it needed to be integrated into the Neural AV Mapper. OSC communication was used to access the sensor data and button events. The buttons and motion data were then mapped as shown in Fig. 6.4.

[52] https://youtu.be/0WE-omAUxpw (accessed 07/06/2020)
[53] https://www.youtube.com/watch?v=Mk82Ka4eUjI (accessed 07/06/2020)
[54] https://thp.io/2010/psmove/ (accessed 07/06/2020)

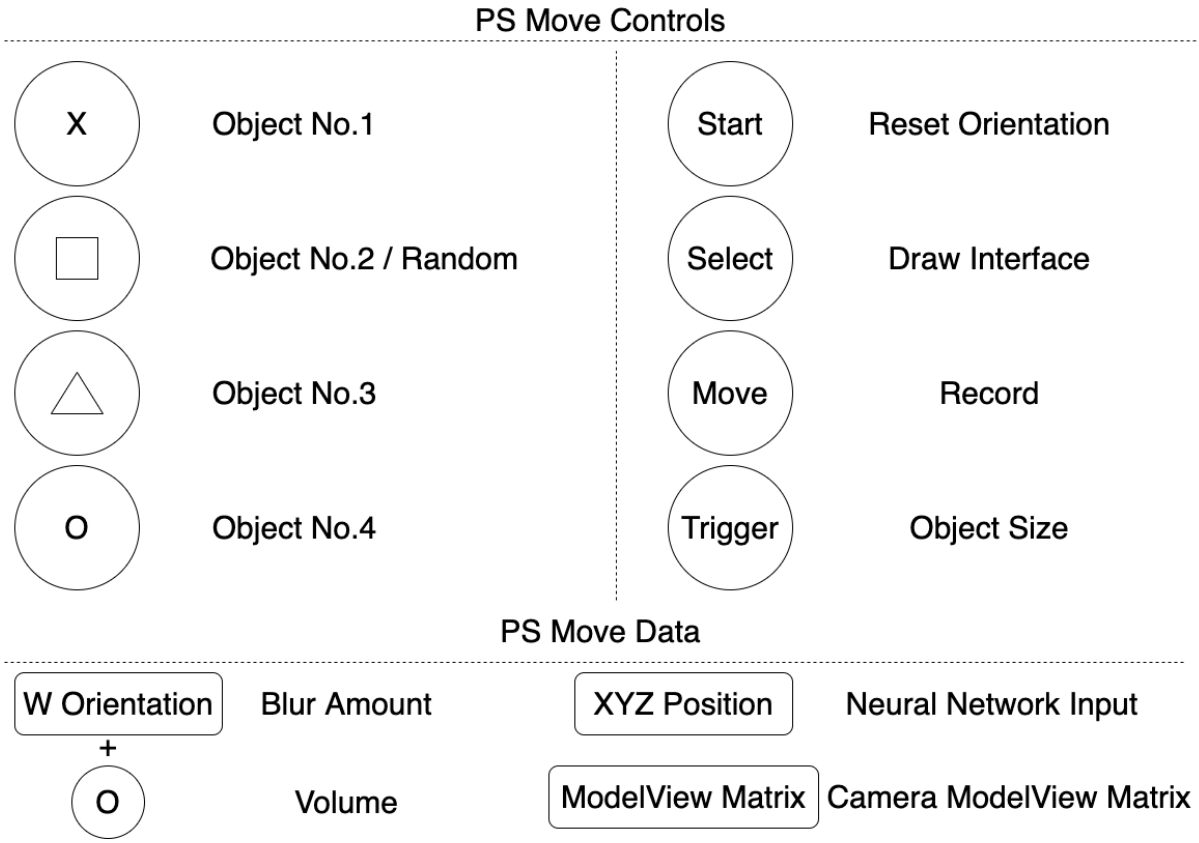

Figure 6.4 PSMove controls.

The cross, square, triangle and circle buttons select an audiovisual object. These buttons were used to transition between sections in the performance. The start button resets the orientation of the object. This was useful as the tracking of the camera would sometimes drift, and using this functionality helped to centre the object. The select button turns the GUI on and off. The move button records the position of the controller to add to the training examples during preparation. The trigger button is mapped to an additional object size. This had the effect of briefly enlarging the object on screen. Due to the mapping of size to sound power as discussed above, this would also create a brief swell in the audio. The *w* orientation of the controller was mapped to the blur effect. This meant that the rotation of the controller could be used to adjust the level of visual blur on screen. This visual blur is also associated with delay in the audio. When the circle button was pressed, the *w* orientation would become a volume control. This was necessary as a practical measure during performance to ensure there were no loud pops at the system startup. The *xyz* position of the controller is determined using the camera on the Macbook Pro and the PSMove API. This data is used as input to the neural network. Finally, the mode view matrix of the controller is mapped to the model view matrix of the OpenGL camera. This provided control of the position of the object on screen.

6.2.3 Compositional Methodology and Structure

The algorithms were built first to generate the audio and visual material. The controller was then chosen and integrated into the system using the PSMove API. After building this foundation the composition of the piece was undertaken.

It was felt that the approach taken with the four studies showed promise of forming the basis of a full composition. Due to this, a similar process was followed with *Ventriloquy I*. The process began by exploring the parameter space and recording audiovisual objects that were perceived to exhibit a strong audiovisual bond. The intention was to use these objects as structural points within the composition. The method used to find audiovisual objects was a mixture of manual adjustment and randomisation of parameters. This was followed by observation and assessment of the quality of the cross-modal correspondences.

First Draft

The first draft of the material was completed in the following way. The first draft version contained four distinct shapes. They consisted of an orange cube, an orange sphere, a white cube and a white sphere. Each of these basic shapes were conceptualised as being representative of a distinct parameter space. Within each parameter space it was decided that there should be six anchoring audiovisual objects laid out, as in Fig. 6.5. This arrangement seemed to make the best use of the available space. Note that the orange crosses denote that the objects are on the same vertical plane, whereas the green and red crosses are closer to the camera on the *z* plane.

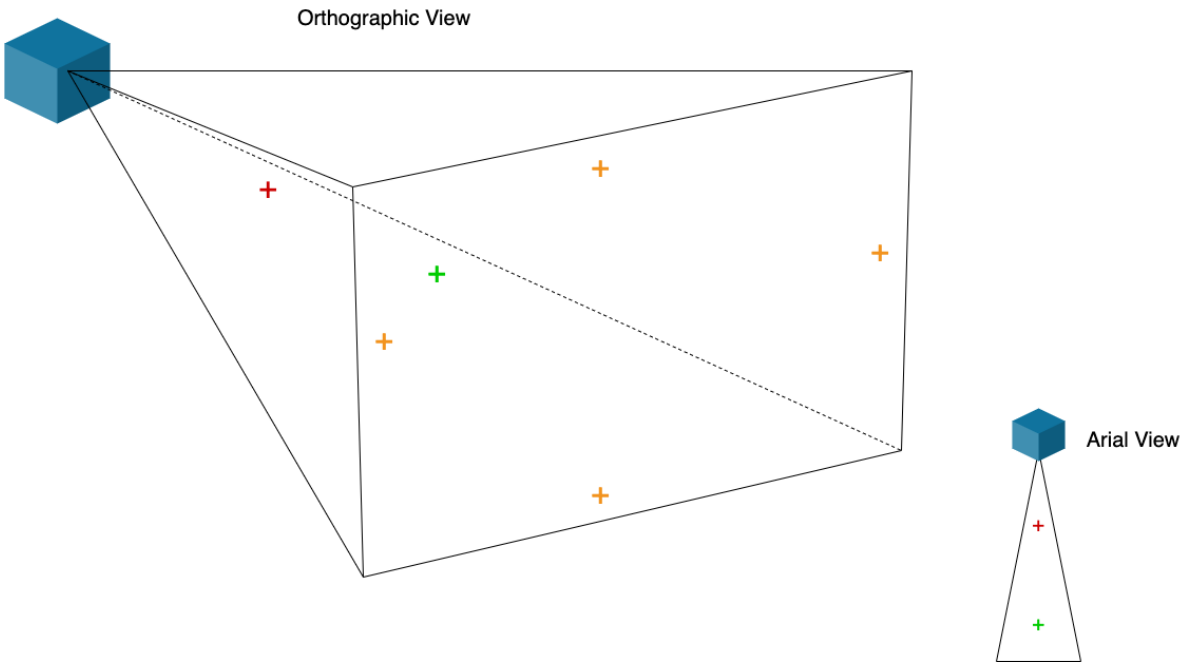

Figure 6.5 Spatial layout of audiovisual objects for first draft of neural network training.

Audio and visual parameters were then found that exhibited strong cross-modal correspondences. These initial audiovisual objects can be seen in the accompanying media pack under *mediaFiles/ventriloquy/ventriloquy1/1stDraftMaterial*. The files are named according to the regression model and object number within that model. For example, the third audiovisual object, belonging to the second model, is named *vent1_1stDraft_mod2_av3.mov*. There are twenty-four in total. The *model0* directory contains videos of the orange sphere; the *model1* directory contains videos of the white cube; the *model2* directory contains videos of the orange cube; the *model3* directory contains videos of the white sphere. Although the appearance of the audiovisual objects were satisfactory in and of themselves, it was felt that including six within each parameter space resulted in over-crowding. It was also felt that four separate spaces became a bit too similar in terms of variation. After this reflection, it was decided to construct a separate set of objects.

Second Draft

The second draft of the composition material was conducted in the exact same way as the first draft. It was decided to search the parameter space for suitable audiovisual objects and save them. These objects can be viewed in the accompanying media pack under *mediaFiles/ventriloquy/ventriloquy1/2ndDraftMaterial*. There are fifteen objects in total. They are named in the same manner as the first-draft objects except with *vent1_2ndDraft* at the start of each file name. It was decided to reduce the number of trained parameter spaces from four to three. The

number of objects was also reduced within each space from six to five. It was felt that this gave the performance area a bit more space, and also cut down on repetition of material.

After deciding on the objects, they needed to be arranged within each parameter space. They were laid out according to intuition as to how well they interacted with each other. One similar placement for each space, was that the closest object to the camera should be chaotic. During practice with the system, it was discovered that it was easiest to introduce another shape visually when the shape was behaving chaotically. This allowed for smooth transitions during performance. An example of such a transition, from the first section to the second section, can be seen between 3:00 and 3:09 in the performance video.[55] This video can also be found in the media pack[56].

After deciding on the placement of each audiovisual object, the neural network was trained using the cartesian coordinates of the PSMove as input data and the parameters of the audiovisual object as output data. After training each network some time would be taken to play with the system and explore the areas between the example objects. If the neural network exhibited interesting areas, the model would be saved. If the results were unsatisfactory, the network would be re-trained, with the objects associated with different locations in space. Once each of the neural networks exhibited satisfactory behaviour, three separate audiovisual parameter spaces were then available to act as performance interfaces. See Fig. 6.6 for the layout of the audiovisual objects within the space. As above, note that the numbering in the diagram below does not correlate to the numbering of the audiovisual objects in the example videos.

[55] https://youtu.be/suYmuV8lOPg (accessed 11/09/2020)
[56] *mediaFiles/ventriloquy/ventriloquy1/Ventriloquy_SeeingSound5_1080H264PCM256_PL_MedSpeed.mov*

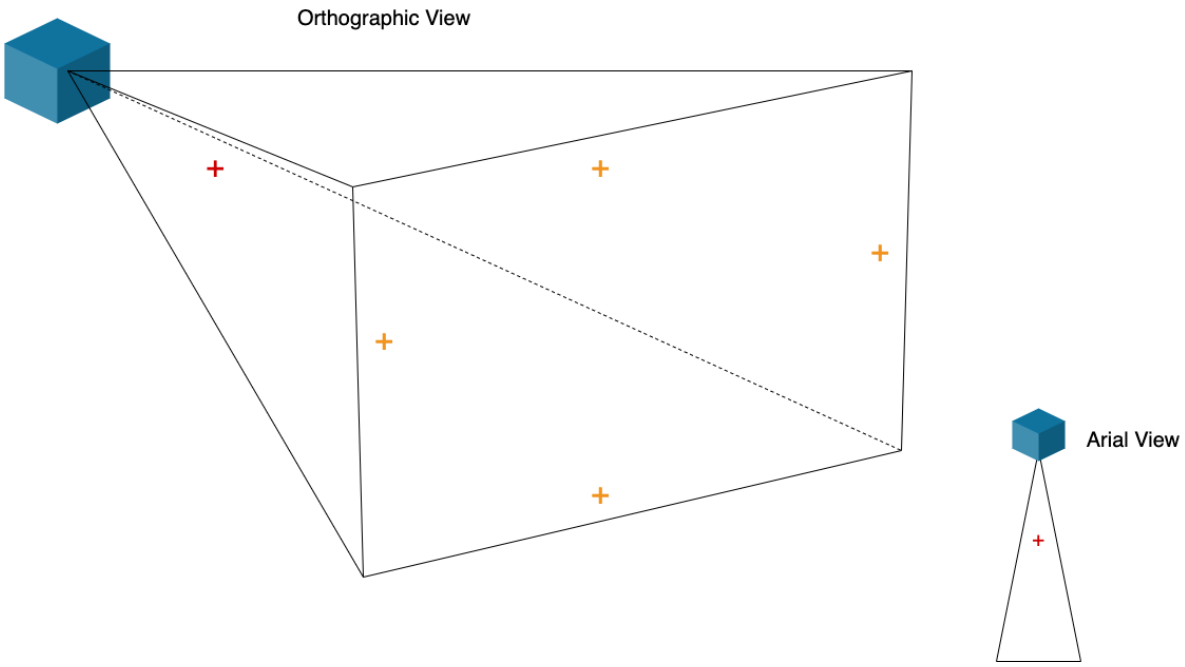

Figure 6.6 Spatial locations of audiovisual object training examples for the second draft.

The final arrangement of material then involved thinking about the composition as a whole in terms of sections. Smooth transitions between parameter spaces were desirable for the performance. As mentioned above, chaotic audiovisual objects were placed at the same spatial location within each neural network in order to achieve this transition. The combination of networks was also an interesting possibility. As discussed in Chapter 5, the parallel combination of parameter spaces could lead to interesting results. Some of the combined output from the parallel use of models became focal points themselves during performance. An example of this is the opening figure in the example video *Ventriloquy (excerpt2)*[57]. The video file can also be found in the media pack under *mediaFiles/ventriloquy/ventriloquy1/ventriloquy_ex2.mov*. This combination emerged through exploration of the combined cube-based neural networks. The visuals and audio demonstrate strong cross-modal perceptual binding and became an important part of the final piece. This section emerges in an altered form at 7:33 in the performance video. In this way, the actual performance of the piece became a mixture of the initial audiovisual objects, and also the combined, in-between formations.

[57] https://vimeo.com/251893597 (accessed 11/09/2020)

The piece has the following structure:

- Section 1 (orange cube) (1:00 - 3:08)
- Section 2 (orange sphere) (3:09 - 4:32)
- Section 3 (orange sphere + orange cube alternating) (4:33 - 6:16)
- Section 4 (orange cube + white cube) (6:17 - 9:00)
- Section 5 (white cube) (9:01 - 11:08)

The sections here overlap into each other, but an approximate delineation between the sections is represented through the timestamps indicated above. The colour scheme of the piece begins with vibrant orange and yellow and gradually fades to grey and white in the end. This was a conscious decision as it was felt that it provided a suitable visual arc.

6.2.4 Performance

The piece premiered at Seeing Sound 5, Bath Spa University, on the 23rd of March, 2018. It was performed on one screen with stereo sound. The screen itself took centre stage. The performer was positioned to the side but within view of the audience. An initial consideration was to place the performer behind the screen. This was intended to allow the audience to fully focus on the interaction between the audio and the visuals on the screen. There was some concern that the presence of the performer would distract the audience from the focus of the piece. However, after some discussion, it was realised that removing the performer from the performance would negatively impact the audience's perception of the liveness of the piece. In order for the audience to perceive a sense of liveness, they needed to see the performer. This also aligns with the study carried out by Correia and Tanaka (2014) in that they identify audience involvement as one of the important considerations an artist must make when performing live.

The performance was arranged within a loose structure. Repeated associations were utilised in order to provide some recognisable structure and a sense of resolution for the audience. This was informed by Chion's suggestion that synchresis is Pavlovian. Repetition is also a fundamental compositional device that has been used throughout the history of music composition. There is evidence to suggest that repetition is important to an audience on an emotional level (Livingstone, Palmer and Schubert 2012). Improvisation with the material was also an important element of the performance. For this reason, a precise score was not created. This also holds true for *Ventriloquy II* below. It was felt that an exciting characteristic of this approach to audiovisual composition is in discovery of new correspondences, even during the performance. This is one of the advantages of using a non-linear

mapping approach such as a neural network. Although the initial audiovisual objects provide cadential points, during a performance these can be used sparingly.

6.2.5 Theoretical Observations

The initial audiovisual objects, chosen as the core material of the piece, demonstrated strong perceptual binding. This was usually due to the effect of audiovisual ventriloquism and balanced relative-temporal-motion. It appears to be quite strong and constant in certain examples such as *vent1_2ndDraft_mod2_av3.mov*[58]. However, there are other objects where it is more variable, in that the audio and visual movement drifts away and back towards each other. For an example of this behaviour please see *vent1_2ndDraft_mod0_av1.mov*[59]. The audio and visual motion are not quite in sync until 00:30, the visuals become blurred, and some delay is applied to the audio. There is also an underlying, low frequency texture apparent in the audio that binds with the motion of the chunky parts of the visual object. The shape then comes back into focus, and the dry audio signal returns at 00:39. At both of these points there is a perceptual joining of the audio and visual streams that produces a sense of satisfaction. As the pitch of the audio rises the visual movement seems to speed up. There is no speed difference specified within the code, so this may be an example of audiovisual distortion where one modality will perceptually affect the other. In other words, this may be added-value. In this case the sound is affecting the perceptual movement of the visual shape. Whereas vision is usually the dominant sense, and will affect hearing, sound has been shown to affect vision in certain circumstances (Shams, Kamitani and Shimojo 2011).

This behaviour is interesting in that it draws the audio-spectator's attention to the fact that there are separate media streams. When the movement of the visuals and the movement of the audio are locked in sync it gives the impression that the audiovisual object on the screen is whole. This drifting characteristic provides a sense of satisfaction, or resolution, when the two streams come back together. It is interesting to note that the resolution experienced here depends on the media streams being separate and then coming together. So, without the separation, there would be no resolution and perhaps the effect of added-value would not be as amplified. In terms of audiovisual balance, this process can be conceptualised as the media stream becoming unbalanced, and then balanced again. When the audio and visuals are balanced they appear as a single audiovisual unit. When they are unbalanced they are separated and can be distinguished as two separate entities.

[58] mediaPack/ventriloquy/ventriloquy1/2ndDraftMaterial/model2
[59] mediaPack/ventriloquy/ventriloquy1/2ndDraftMaterial/model0

This reveals an extra dimension to the concept of audiovisual balance, as here, it is possible for the streams to be perceived as separate but not unbalanced in terms of perceptual domination. For an example of this see *vent1_1stDraft_mod1_av0.mov*[60]. In this clip it is possible to observe two separate streams and also perceive them simultaneously. In this case they do not seem to be interacting. Rather they exist alongside each other. This suggests that whilst audiovisual balance can manifest itself by unifying the sensory media stream, there are also situations where the streams may be balanced but appear separate. However, it could be argued that they are unified aesthetically rather than rhythmically. The foundations of *Ventriloquy I* have been built on the idea of rhythmic similarity between the audio and visuals. This has been the main method used to observe audiovisual balance and added-value. This example demonstrates other avenues of exploration based on characteristics other than rhythm.

6.2.6 Feedback and Improvements

Performing the piece live provided an excellent opportunity to get some feedback about the experience of the audience and how they perceived the piece. This feedback was received informally through conversation with symposium attendees throughout the weekend. Several comments were received from people who particularly enjoyed the emergence of textural detail on the surface of the sphere at 3:42 in the performance video. Another comment received was that the piece seemed very academic. After thinking about this remark and reviewing the recording of the performance, it is clear that the piece may have come across as quite clinical. In the performance video the motionless appearance of the performer (the author) was quite surprising. This is in stark contrast to the subjective feeling of performing with the system. During performance an attempt was made to use large gestures in an attempt to convey expressivity to the audience. This disconnect prompted a re-evaluation of the method of control. As a musician, there was an expectation of tactile resistance when trying to perform with expression. The PSMove offered no resistance during exploration of the parametric space. This was where the embodied expressivity of the composition was located. However, the lack of resistance in the controlling device meant that the gestures lacked a sense of expression and emotion. This will be addressed further in *Ventriloquy II*.

During soundcheck the sound-engineer was instructed to reduce some of the high frequencies of the audio, as there was a concern that the strength of certain high frequencies in the piece would prove

[60] mediaFiles/ventriloquy/ventriloquy1/1stDraftMaterial/model1

uncomfortable for the audience. After reviewing the recording of the piece, this appears to have been a mistake. There are certain parts of the piece where the object is close to the camera and chaotic. However, the sound seems to almost disappear at 3:53, 4:08, 5:05 and 8:26. This happened for several reasons. Firstly some of the high frequencies were filtered out of the master signal. Secondly, the use of delay further clouded the perception of the reduced signal giving an overly ethereal quality to the sound that did not match the visual movement. This led to some feedback that the sense of scale was off. When the object gets larger the volume of the audio should increase and become tonally deeper. On reflection, it could be interesting to maybe harness this expectation and subvert it as a compositional device. In a similar fashion, Chikashi Miyama subverted synchretic expectations in his piece *Quicksilver* (2010). Here Miyama plays with the audience's expectation of what a liquid drop would sound like. It could be interesting to see how the expectation of size and sound could be subverted.

6.3 *Ventriloquy II*

Ventriloquy II was developed using the same system as *Ventriloquy I*. The source code can be viewed at its GitHub repository[61]. It can also be found in the accompanying media pack[62]. This piece can be thought of as a further exploration of the techniques and compositional devices employed in *Ventriloquy I*. It can also be seen as an attempt to expand into a more immersive practice. Finally, it can also be seen as a demonstration of the flexibility of the Neural AV Mapper. Whilst the underlying architecture is exactly the same as that of *Ventriloquy I*, the aesthetic result is quite different.

6.3.1 Motivation

This piece arose out of a combination of factors. Firstly, there was a desire to develop some of the techniques and ideas that had been explored in *Ventriloquy I*. It was felt that there was room for expansion regarding the audiovisual aesthetic. An exploration of more textured and varied colours was undertaken. In addition to this, the visual structures were expanded from those used in the *Ventriloquy I*, which consisted of simple cubes and spheres. Regarding the audio aesthetic, it was felt that there was room for more textural depth and richness. The controller was also considered for experimentation in the hope of achieving a more expressive performance. The piece was performed in the SIML Space[63] as part of the Immersive Pipeline[64] project led by Atau Tanaka. This venue is an

[61] https://github.com/bDunph/ventriloquy2 (accessed 22/09/2020)
[62] sourceCode/ventriloquy2
[63] http://sonics.goldsmithsdigital.com/the-siml-facility/ (accessed 11/06/2020)
[64] http://sonics.goldsmithsdigital.com/immersive-pipeline/ (accessed 11/06/2020)

immersive space capable of 360 degree projection with a 12.2 surround sound system. In this space, the IML approach to controlling audiovisual material could be implemented in an immersive context for the first time. In this way, *Ventriloquy II* evolves from the single screen and stereo sound of *Ventriloquy I* into a shared immersive experience for the audience. The layout of the space and technical specifications informed the aesthetic development of the piece to a certain extent.

6.3.2 Development of System and Aesthetic Elements

The underlying system itself is identical to the one used for *Ventriloquy I*. The elements that differ are the aesthetic material and the controller used. Further, for the performance, some commercial software was used to map the visuals to the six projectors that were used in the space.

Visual Aesthetic

For this piece, there is an emphasis on using more complex shapes than the cube and sphere of *Ventriloquy I*. It was decided to focus on a single shape, but vary the colour and texture across three iterations of the same form. These are shown in Fig. 6.7.

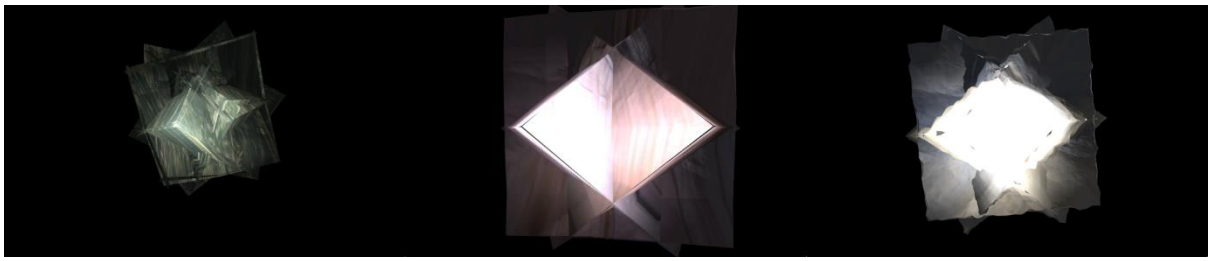

Figure 6.7 Ventriloquy II visual structures.

The shape itself was created by combining three cubes, rotating one of them 45 degrees around the positive *y* axis, and another cube 45 degrees around the negative *y* axis. The code is shown in Ex. 6.6.

```
00 ofVec3f axis = ofVec3f(0, 1, 0);
01 ofVec3f oppositeDirectionAxis = ofVec3f(0, -1, 0);
02 box2.tilt(45);
03 box2.rotate(45, axis);
04 box3.tilt(45);
05 box3.rotate(45, oppositeDirectionAxis);
```

Example 6.6 Construction of Ventriloquy II visual shape.

The textures and colours consist of different types of rock and stone. They are loaded and bound to the objects in the same manner as *Ventriloquy I*. The colours appeared to work together aesthetically, and the detail in the textures provided a certain amount of depth and richness than was achieved using the more uniform textures of *Ventriloquy I*.

Audio Aesthetic

The audio processing is very similar to *Ventriloquy I*. The only differences are the use of less delay and the use of frequencies in lower ranges, as indicated in Ex. 6.7.

```
00 switch (waveform){
01           case 0 :{
02           sine1 = sineOsc1.sinewave(oscilFreq + modulator);
03           sine2 = sineOsc2.sinewave((oscilFreq * 0.2) + lfo) + sine1;
04           sine3 = sineOsc3.sinewave(oscilFreq * 0.1) + sine2;
05           sine4 = sineOsc4.sinewave(oscilFreq * 0.78) + sine3;
06           subSig1 = (((sine1 + sine2 + sine3 + sine4) * 0.25)  *
mappedVol);
07           break;
08           }
09          case 1 :{
10          pulse1 = pulseGen1.pulse(oscilFreq + modulator, pulseDuty);
11          pulse2 = pulseGen2.pulse((oscilFreq * 0.37) + lfo, pulseDuty) +
pulse1;
12          pulse3 = pulseGen3.pulse(oscilFreq * 0.92, pulseDuty) + pulse2;
13           pulse4 = pulseGen4.pulse(oscilFreq * 0.31, pulseDuty) + pulse3;
14           subSig1 = (((pulse1 + pulse2 + pulse3 + pulse4) * 0.25) *
mappedVol);
15           break;
16           }
17           case 2 :{
18           square1 = squareOsc1.square(oscilFreq + modulator);
```

```
19            square2 = squareOsc2.square((oscilFreq * 0.17) + lfo) + square1;
20            square3 = squareOsc3.square(oscilFreq * 0.69) + square2;
21            square4 = squareOsc4.square(oscilFreq * 0.04) + square3;
22            subSig1 = (((square1 + square2 + square3 + square4) * 0.25) *
mappedVol);
23            break;
24            }
25    }
```

Example 6.7 Ventriloquy II audio generators.

Above, there are three separate waveform generators. The *case 0* processing block is built on sine waves, *case 1* is built on pulse generators and *case 2* is built on square waves. If you compare the square wave generator here to the same code displayed in Ex. 6.3, you can see that the partials at lines 19 to 22 are processed at a lower frequency. Finally, each of the generators was utilised for a separate regression model. Model 0 utilised *case 0*, model 1 utilised *case 1* and model 2 utilised *case 2*. See the video files in the accompanying media pack under *mediaFiles/ventriloquy/ventriloquy2/2ndDraftMaterial*.

Mapping

A similar mapping approach was undertaken in this piece as in *Ventriloquy I*. A neural network is used to map real-time input from the controller to audio and visual parameters. The audio and visual parameters output from the neural network are the same as *Ventriloquy I*. Some parameters are also mapped manually to create certain effects that were felt to enhance the performance and perception of the audiovisual objects. As in *Ventriloquy I*, the horizontal position of the objects are mapped to the left and right speakers depending on where they are on screen. The simple distance function, that made the volume of the object decrease as it moved away from the camera, was also retained. It was decided to change the blur mapping from *Ventriloquy I* as the results of the mapping were deemed unsatisfactory. Here, the blur amount is mapped to a low-pass filter that cuts high frequencies as the blur amount increases. This mapping seems to work quite well and creates a strong cross-modal perceptual bond. This effect can be seen at 5:26 and 6:20 in the performance footage[65]. The video file of the performance can also be found in the media pack[66]. A further general mapping was also made between the rough frequency range of the audio to textural colour. This is illustrated in Fig. 6.8. The

[65] https://www.youtube.com/watch?v=BafCWoU8aV8&t=8s (accessed 23/09/2020)
[66] mediaFiles/ventriloquy/ventriloquy2

dark green shape corresponds to the lower frequency range (*oscilFreq* = 0 - 1000Hz). The white texture was mapped to a mid frequency range (*oscilFreq = 1000 - 3000 Hz)*. Finally, the pink texture was mapped to a higher frequency range (*oscilFreq = 3000 - 4000Hz)*. The *oscilFreq* parameter was adjusted manually as the material was being generated. These values can be heard in the *2ndDraftMaterial* example videos[67].

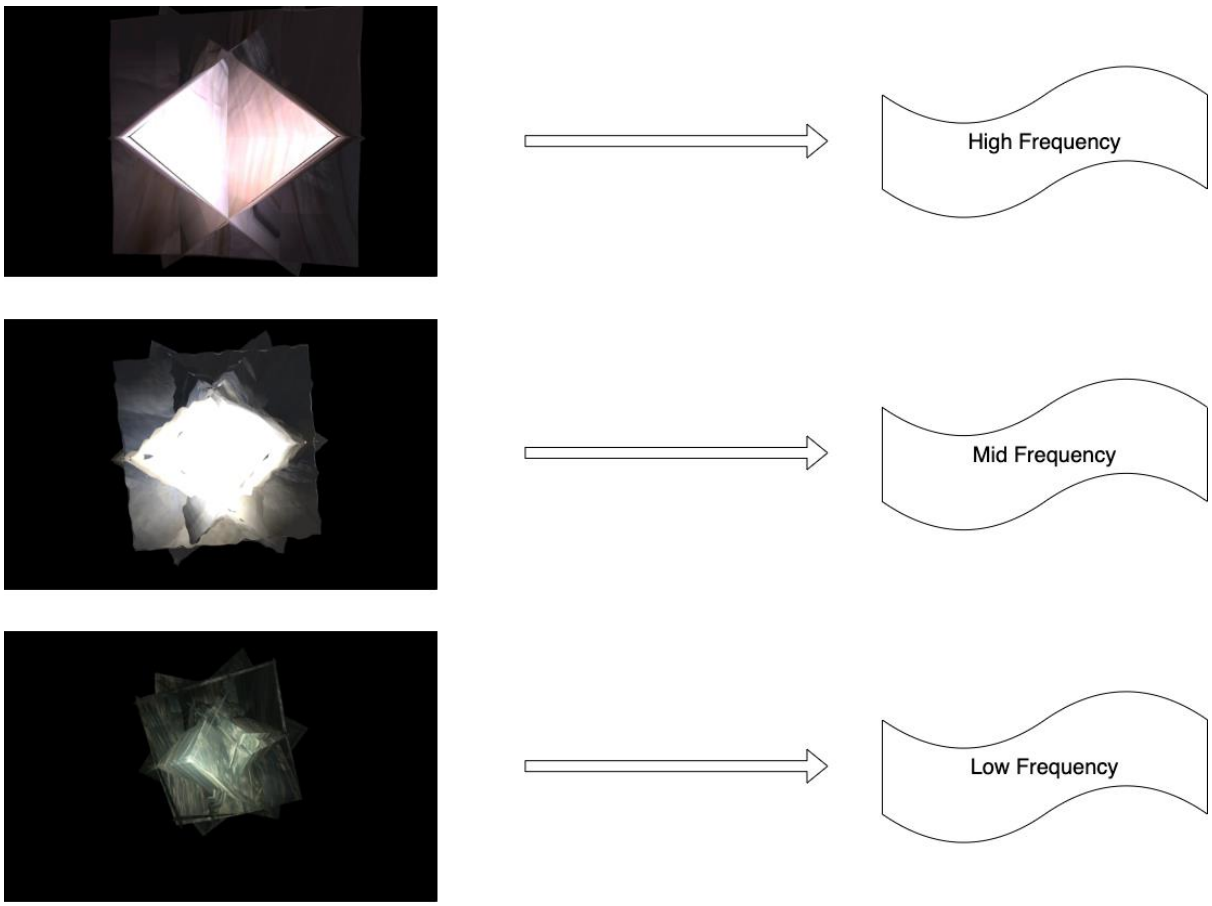

Figure 6.8 Approximate frequency range of models.

These mappings were made loosely in accordance with some of the automatic cross-modal correspondences discussed in Chapter 5. It has been shown that darker surfaces correspond well with lower-pitched sounds and lighter surfaces correspond with higher-pitched sounds (Evans and Treisman 2010). Although the sound in *Ventriloquy II* is unpitched, the general frequency range is intended to fill the same role here. The opportunity was also taken to introduce some dissonance, in that the white model is mapped to the mid range and the pink model to the higher-frequency range. Intuitively, the white model would seem to match more readily with the highest frequency range. This intentional dissonance may not be immediately apparent, however the intentional subversion of

[67] mediaFiles/ventriloquy/ventriloquy2/2ndDraftMaterial

automatic cross-modal correspondences could be an interesting method for creating tension in an audiovisual piece.

Controller

A different controller was employed for the performance of *Ventriloquy II* than had been used for *Ventriloquy I*. This was due to several reasons. Firstly, the lack of resistance that the PSMove offered was a concern. The lack of tactile resistance was noticeable in the performance of *Ventriloquy I*. This suggested that it is an essential part of the act of performing for the author. Secondly, the performance context did not lend itself well to using a camera to track the PSMove. The PSMove API relies on a dark environment so that the camera can track the illuminated bulb on the controller. For *Ventriloquy I*, it was possible to position the performer off to the side of the screen, in a darker part of the space. The SIML space is an immersive venue with 360 degree projection. This meant that it would not be possible to position the camera towards a dark part of the space.

Some experimentation with the touchOSC[68] app was undertaken, as it was thought that it would be inexpensive, convenient and flexible. The app does have each of those qualities. However, the tactile response of the hard screen did not provide enough flexibility to encourage expressive control of the material. There was a sense of simply tapping controls on a phone, with no real connection to what was happening on the screen and through the speakers. This suggested a controller that provided a sense of flexible resistance during performance.

The Keith McMillan QuNeo[69] offers multiple pads, faders and switches that are all capable of sensing pressure and position. Every control on the device is made from the same material, which offers tactile resistance when interacted with. This allows the user to press into the pad or fader and feel it give way. This characteristic of the controller gives the impression of expressivity that was necessary for this piece. Each of the controls also light up with embedded LEDs that give useful feedback during performance. The QuNeo can communicate with the computer via MIDI or OSC. OSC was used in this performance to send streaming position data from the faders. Some of the pads were also used as simple triggers.

6.3.3 Compositional Methodology and Structure

A very similar approach to composing this piece was followed as in *Ventriloquy I*. With *Ventriloquy II* it was felt that there was an opportunity to refine the compositional methodology. It was felt that

[68] https://hexler.net/products/touchosc (accessed 13/06/2020)
[69] https://www.keithmcmillen.com/products/quneo/ (accessed 13/06/2020)

the steps required to create the material, train the networks and develop the performance, were better understood here. There was also a better idea of the types of audiovisual objects that would work well together. The reliance on delay was also addressed through adjustment of the audio aesthetic and relevant mappings. There was also an effort to be more conscious of scale and expressivity.

First Draft

Like the previous piece, material was gathered in the form of specific audiovisual objects that appeared interesting and that exhibited strong cross-modal bonds. The material was separated into three distinct categories, based initially on the visual textures that were used to cover the shapes. It was hoped that using the same underlying shape, but varying the texture, colour and sound associated with each, would help to create a unified composition. The number of training examples for each network in *Ventriloquy I* seemed appropriate, so five examples for each network were also implemented in this piece. The first draft of the material yielded fifteen audiovisual objects. These can be seen and heard in the accompanying media pack[70].

The correspondences again rely on cross-modal ventriloquism and balanced relative-temporal-motion. The preparation for this piece involved a strict observation process when choosing audiovisual objects that stayed in sync, rather than objects that drifted in and out of sync. For example *vent2_1stDraft_mod2_av0.mov*[71] stays tightly in sync. Another type of perceptual ventriloquism here is the movement of the textures with the sound. An example of this can be seen in the clip *vent2_1stDraft_mod0_av0.mov*[72]. Here, you can see the OpenGL texture is moving noisily across the surface of the object. In this example, the cross-modal correspondence is located in the movement of the texture rather than solely the movement of the vertices, although there is some movement in the vertices. Further, this is not the only correspondence that is at play here. Remember that the frequency range of the audio is also corresponding with the darker shade of the texture. Each of the three objects are composed in this way, with multiple correspondences being perceived simultaneously.

At the end of the previous clip, the visual blur effect can be perceived as being mapped to the low-pass filter. It happens at 0:16. Another example of this mapping can be seen in the clip *vent2_1stDraft_mod1_av1.mov*[73]. As the visual details become less defined the audio features also

[70] mediaFiles/ventriloquy/ventriloquy2/1stDraftMaterial
[71] mediaFiles/ventriloquy/ventriloquy2/1stDraftMaterial/model2
[72] mediaFiles/ventriloquy/ventriloquy2/1stDraftMaterial/model0
[73] mediaFiles/ventriloquy/ventriloquy2/1stDraftMaterial/model1

lose their edge. This is an example of a simple, static, one-to-one mapping as discussed in Chapter 5. Due to the direct nature of this mapping, the effect was used sparingly, at select moments in the performance. This will be discussed in more detail below. In several of the clips, the object will suddenly become larger. This is a performative mapping implemented within the piece. It works the same way as it did in *Ventriloquy I*. An example of this is in the clip *vent2_1stDraft_mod2_av4.mov*[74].

After settling on a range of audiovisual objects, they were arranged in three-dimensional space. This arrangement in three-dimensional space was the same approach used in *Ventriloquy I* above, and represented the performance control interface. At this point the PSMove controller was still being used. The spatial layout of the objects in *Ventriloquy I* was deemed satisfactory, so it was decided to lay them out in a similar fashion, with an example object at each corner of a square, and one more in the centre towards the computer. A process of playing with each of the parameter spaces was then undertaken, followed by reflection on the results. At this point it was decided to return to the audiovisual objects, and attempt to find some more variety in their appearance and sound.

Second Draft

The second draft of the piece was primarily focused on adjusting the appearance and behaviour of the audiovisual objects, so that they appear more unified across the different models. This was achieved by creating similar figures within each model. These can be viewed in the accompanying media pack[75]. After generating the audiovisual objects, it was time to decide how they would be arranged to form the parametric spaces for each object. Using the QuNeo, it was necessary to adjust the way in which the audiovisual objects were arranged in the performance space. With the PSMove, it was possible to make an intuitive correlation between physical space and parametric space. With the QuNeo this was not possible. To overcome this challenge it was decided to map physical coordinates to fader values on the controller. Cartesian coordinates are represented using three separate numbers. These are the *x*, *y* and *z* positions in 3D space. These values were mapped to three separate faders on the controller. This meant that instead of providing the neural network with input directly from a spatial position, the neural network was provided with the three values from the faders.

[74] mediaFiles/ventriloquy/ventriloquy2/1stDraftMaterial/model2
[75] mediaFiles/ventriloquy/ventriloquy2/2ndDraftMaterial

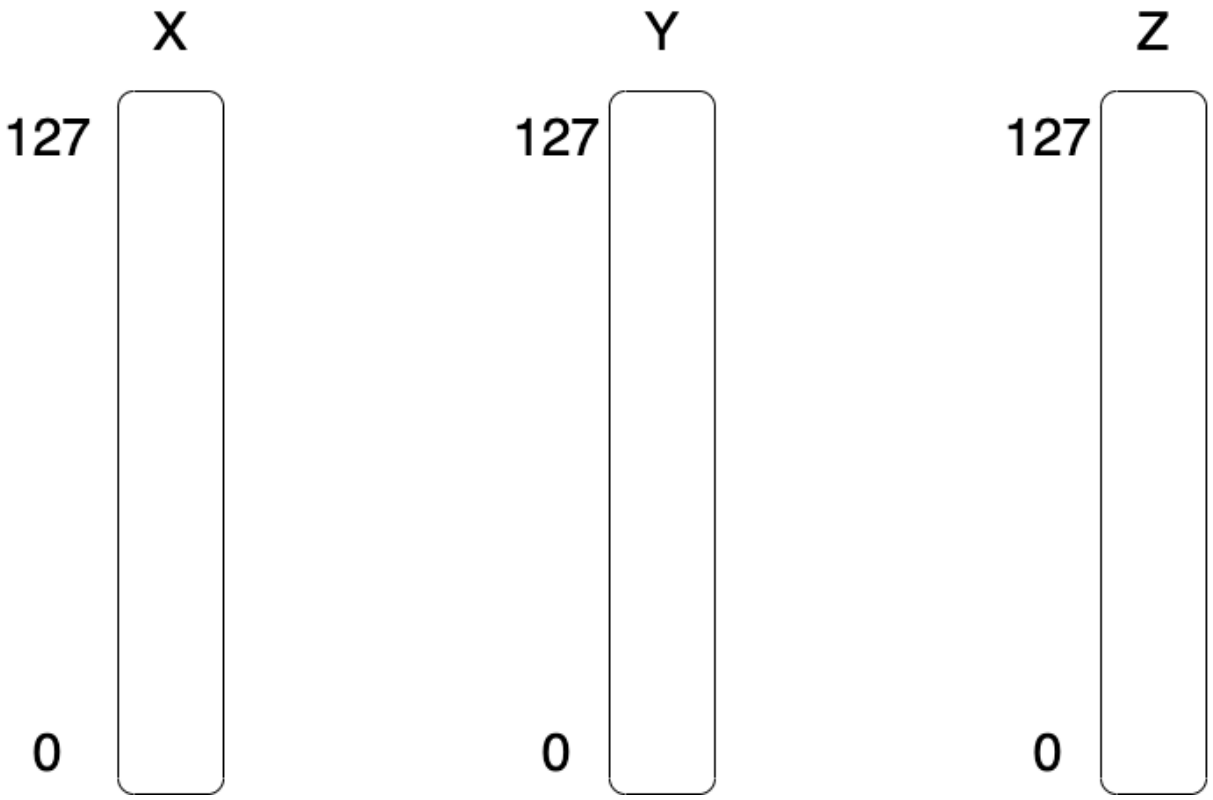

Figure 6.9 QuNeo faders as input control.

Fig. 6.9 shows the faders of the quNeo as they are mapped to what were previously *x*, *y* and *z* parameters. The ranges of the cartesian parameters were mapped to the midi values of the controller. In addition to this, it was decided to focus on keeping similar objects at similar locations in the parameter space. For *Ventriloquy I*, chaotic objects were consciously placed at the same point in space to allow for smooth transitions between objects. This concept was extended here by making sure each neural network was capable of reproducing a star-like object, two solid objects and two chaotic objects for each object. These were consciously placed at roughly the same point in the parameter space for each neural network. It was hoped that this would create a sense of continuity between the three parametric areas and would reinforce the objects as cadential points in the piece.

After deciding on where to place each audiovisual object, each network was trained and the resulting model was saved. Once this was done, each parameter space was explored, combining them to see if any interesting behaviour emerged. Through this exploration and play, a structure emerged. As the spaces and combinations were explored, there was a noticeable reduction in awareness of the location of the audiovisual objects used as training examples. Using the QuNeo, the coordinates of the initial objects were not intuitive in a spatial way. The separation of the *x*, *y* and *z* values into three separate

channels removed the sense of a three-dimensional arrangement of the material. This observation will be discussed further in section 6.3.5.

The piece is in a rough *ABCBA'* form. Section A consists of three star-like objects that oscillate at different frequencies (0:22 - 2:35 in the performance video). At the end of section A there is a brief bridge section (2:35 - 3:51) that introduces section B. Section B starts at 3:52 and ends at 5:37. This section emerged through experimentation with combining objects. The camera is positioned inside one of the objects and a second object is seen at the centre of the screen. The objects that make up the core of this section are the low-frequency/dark object and the mid-frequency/white object. This gives the audio a low-end and mid-frequency sound with space in the high-end. The objects move to either side of the screen as section C begins (5:38). This section features the same two objects except from outside. This section utilises the high-range model in a staccato manner. The section is punctuated by short irregular bursts of the high-frequency model. This can be seen starting at 6:46. This punctuating figure was introduced in section B. The mid-range and low-range models traverse the screen and move in and out of focus before combining again at (7:27). Chaotic material again acts as a transition into a repeat of section B at 7:28. This section is shorter than the last time and gradually moves to section A' at 8:43. Here, a gong-like event in the audio material announces the final section, an inversion of the opening section, in which the objects separate from each other and disappear in turn.

6.3.4 Performances

The performance of *Ventriloquy II* took place in the SIML space at Goldsmiths, University of London, on the 12th of April, 2018 (Fig. 6.10). The venue is in a rectangular shape with a total of six screens. The sound system comprises twelve speakers arranged throughout the venue. Four speakers at each top corner and four speakers at each bottom corner. Two speakers, top and bottom, at the midpoint of each long side. The position of the performer in the performance space was dictated by the location of the computer that was being used for the event. The QuNeo is a wired controller so it needed to stay close to the computer. It was also useful to have the monitor close-by to set up the piece, and briefly glance at it during the performance. However, for the majority of the performance, the performer was looking at the screen directly in front of them. This fostered a more immersive experience for the performer during the performance, rather than just looking at the computer screen.

In the weeks approaching the performance, Pierre Tardif and Dr. Blanca Regina provided assistance in the venue. The visual aspect of the piece was initially conceptualised as a single texture that would wrap around the space covering all six screens. However, during preparations for the event a bug was

discovered in the openFrameworks code that would not allow the application to output the correct dimensions. MadMapper[76] was used to perform projection mapping, and *ofxSyphon*[77] was used to send the video data from the Neural AV Mapper to MadMapper. Unfortunately the bug was not located in time for the performance and a decision had to be made to output six versions of the texture and map them to each projector. With the help of Dr. Blanca Regina the textures were inverted on the long sides of the venue and at each end of the space. The edges were then blended together to create a sense of continuation between the screens.

The audio was output from the application in stereo. The left signal was mapped to all the speakers on one side and the right signal was mapped to all the speakers on the opposite side. The performance followed a loose structure but was improvisational within this. In a similar way to *Ventriloquy I*, it was felt that, as it was a live performance, some improvisation would communicate the liveness to the audience. The piece was also performed in an edited version on a miniature model of the SIML space at Rich Mix, London on the 11th of May 2018, as part of the Splice festival. This performance followed a talk given by Prof. Atau Tanaka on the Immersive Pipeline Project[78].

6.3.5 Theoretical Observations

Ventriloquy II was concerned with similar perceptual explorations as *Ventriloquy I*. At the forefront of the piece is the relative-temporal-motion of audio and visual material and how this shifts in and out of balance. The Gestalt principle of common fate also plays an important role in creating an illusion of synchronicity.

[76] https://madmapper.com/ (accessed 14/06/2020)
[77] https://github.com/astellato/ofxSyphon (accessed 14/06/2020)
[78] http://2018.splicefestival.com/line-up/immersive-pipeline/ (accessed 23/09/2020)

Figure 6.10 Performance of Ventriloquy II at SIML, Goldsmiths, University of London.

There was an attempt to create a sense of isolated-structural-incoherence by using harsh audio textures and ambiguous visual shapes. There is often no discernible structure or progression with expected resolution in the audio material. This is the advantage of the noise aesthetic in this context. The visuals alternate between agitated semi-visible shapes and chaos at different points in the performance. At 2:38 the visuals transition from the opening star-type figures to chaotic, ambiguous colours before consolidating again into a semi-regular form at 3:03. During this chaotic passage, the audio and visuals lose their tight bond. This suggests that if both audio and visuals are simply chaotic noise there has to be some sort of connecting tissue to keep them together. Again we come back to an issue of balance. If the whole piece were to consist of formless noise then it would lose the interest of the audio-spectator quite quickly. At 3:03 a semi-formed object returns and reasserts the audiovisual bond. However, if either of the media streams are too well formed, there is a risk that one specific sensory mode of the material will dominate the material in the other sensory mode. The section beginning at 5:36 succumbs to this. Two solid shapes emerge from the previous section and dominate the perception. Audiovisual balance here is skewed towards the visual. There is an attempt to break this by using a sharp punctuating figure at 6:46 which aims to reassert the position of the audio. On reflection, throughout the piece, there seems to be an imbalance in the relative-expressive-range

between the audio and visual material. The visual palette may be perceived as being much richer in variation and more complex than the audio palette.

On reflection, the deconstruction of the spatial parameter space was undesirable in that it obscured the careful selection of audiovisual objects that were used to train the neural networks. As a result of this deconstruction, the sense of returning to predefined cadential points, that were a feature of the studies and *Ventriloquy I,* was lost. On the other hand, perhaps the obfuscation of the original audiovisual objects allowed a more open-minded exploration of the rest of the parameter space. Ultimately, the spatial-control metaphor that had been previously explored, allows for a more intuitive paradigm in which to arrange and perform with this material. However, deconstructing the control parameters in this way provided some valuable insight into the importance of this spatial approach, and also into the importance of the relationships that emerge through the exploration of the space in-between the audiovisual objects. It was realised that it is important to be aware of the cadential points, but not to rely too heavily on them.

6.3.6 Feedback and Improvements

Ventriloquy II was received well by the audience, with encouraging comments about the sense of unity between the audio and visual material. It was also described as ‘intense’ by another member of the audience, which could be attributed to the relentless character of the audio within the immersive environment. Even though the spatial capabilities of the sound system were under-utilised, the twelve speakers and two subwoofers created an imposing sonic space. It was remarked by several people that there was not enough variation in the sound world. It was felt that the extent of the sonic palette did not match the complexity of the visual elements. This may be due to a reliance on traditional methods of FM, subtractive and additive synthesis. There are many other forms of synthesis that may provide sufficient variation including granular synthesis, physical modelling and spectral techniques. These will be explored in future compositions. Perhaps there are also other musical parameters that could be exploited to create variation. These could include variation of rhythm, dynamics and silence.

The piece was originally intended to consist of a single texture mapped across all six screens. Unfortunately due to a bug in the code, the correct dimensions could not be sent to MadMapper. The cause of the bug was never ascertained. The obfuscation of lower-level functionality in the openFrameworks codebase may have contributed to the difficulty in finding the bug. This lack of clarity is one of the reasons it was decided to move away from openFrameworks. This decision will

be discussed in more detail in Chapter 7, where the *ImmersAV Toolkit* is presented. Due to this bug, one texture was output and mirrored across the six screens. This created six identical images. With the help of Dr. Blanca Regina, the images were rotated to create composite images. This technique worked quite well as a last minute adjustment. The images may have also looked distorted if they were stretched throughout the screens, so perhaps this approach may have been best for the material anyway. Members of the audience asked several times if the visuals were mirrored across the screens. This is confirmation that the visual possibilities of the venue were not fully exploited. It also suggests that some variation in the visuals across screens would have worked well. This also may have made better use of the space. Six separate textures could have been output, as opposed to just one. Small random variation of some of the parameters within each texture could also have been introduced. This would have created interesting movement across the screens. Subtle differences between the textures that may have added some complexity. This approach may provide interesting results in the future.

6.4 Conclusion

Ventriloquy I and *Ventriloquy II* are the culmination of initial investigations into using an IML approach to controlling audiovisual material. The two pieces, whilst aesthetically distinct, were built on the same system. They were built upon the use of a neural network as an intermediate layer that allowed non-linear mapping between a source of data and the audiovisual material. An advantage of this approach meant that the bulk of the material was controlled simultaneously with some audio-to-visual direct mapping. Regarding the idea of kinetic agency discussed in Chapter 3, this approach seemed well-suited to address any potential imbalance in relative-temporal-motion. The system was designed according to the theoretical principles outlined in Chapter 3. They focus on the assertion that audiovisual balance can affect the audience's perception of a piece and allow the relationships between the audio and visuals to emerge. Audiovisual balance is, in turn, affected by certain properties of the material including relative-temporal-motion, relative-expressive-range and isolated-structural-incoherence. These principles originated in the assertion, throughout the literature, that audiovisual artists aim to give equal importance to both the audio and visual material. The concept of relative-expressive-range was discussed in Chapter 3 but its genesis was in the reflection and feedback from the *Ventriloquy* pieces.

One of the major compositional outcomes from the development of *Ventriloquy I* was the discovery that chaotic audiovisual movement allows for the smooth introduction of new material and transitions between sections. This structural discovery is a tangible technique that can be utilised in future

compositions to avoid overly block-like compositional structures. The choice of controller for *Ventriloquy II* was influenced by the lack of tactile resistance afforded by the PSMove. The importance of performing with this resistance had not been considered before, as it was something that had always been taken for granted when playing a traditional instrument such as the guitar or piano. However, to be clear, this preference for tactile resistance is mostly relevant within the specific performance context in which the above pieces are presented. When considering the effectiveness of the IML paradigm in controlling audiovisual material, it was found that the three-dimensional spatial approach, used in *Ventriloquy I*, was much more intuitive than using the faders and touchpad of the QuNeo. It was felt that the spatial method of control is more relevant to the core aims of the research. In a fully immersive context, the AV-participant will be more focused on exploring the parametric space, than with conveying expressivity to an audience. This spatial approach will be explored further in Chapters 8 and 9.

The dramatic effectiveness of the IML control paradigm has been explored through live performance pieces. The next step to achieving the core research goal is to implement the control paradigm into a fully immersive virtual environment. Chapter 7 will present a software toolkit that will allow for the exploration of audiovisual compositions, using an IML control paradigm, in VR.

Chapter 7 The ImmersAV Toolkit

This chapter details the development of the *ImmersAV* toolkit. The source code for the toolkit can be found at its GitHub repository[79] and in the accompanying media pack under *sourceCode/immersAV*. The example videos are linked to YouTube throughout the text and are included in the accompanying media pack under *mediaFiles/immersAV*.

Approaching the exploration of immersive audiovisual practice, the technical approach to creating audiovisual material was re-assessed. It was felt that more control was needed over the tools being used to generate material. There was a desire to work at a level of abstraction that provided a sense of interaction with the hardware. The motivation for creating a bespoke environment was to address this sense of abstraction and create a tool that fit the desired compositional workflow.

7.1 Contemporary Tools

There are a multitude of programs, environments and tools focused on the creation of computational art. Similarly there are several ways to create immersive work, whether that is for augmented reality or full-scale VR. Regardless of the software used, there is a belief that the tool should become invisible in the end result. It is easy to get lost amongst all the contemporary options. Some of the tools considered before deciding on the approach to creating immersive audiovisual art, will be discussed.

7.1.1 Game Engines

General purpose game engines provide an integrated environment within which fully-realised commercial computer games can be developed. The main advantages of these environments is that they contain well established code bases that include physics engines, primitive objects, scene managers, camera rigs and large user bases with many tutorials. The documentation for the main game engines are also very well maintained. This lowers the knowledge threshold for entry into these environments and allows for fast development. The two main game engines are Unity[80] and Unreal Engine[81]. These environments are free to use for non-commercial projects. However, if the project is a commercial development, there is a threshold after which the developer or studio must pay for a licence.

[79] https://github.com/bDunph/ImmersAV (accessed 29/09/2020).
[80] https://unity.com/ (accessed 21/04/2020).
[81] https://www.unrealengine.com/en-US/ (accessed 21/04/2020).

On deciding which route to take with regard to creating immersive audiovisual work, Unity was considered due to the author's limited experience with it in the past. It was also a straightforward way to build an application for VR. However, as the environment was explored, a realisation emerged, that the integration of audio was a secondary concern. Efforts are being made to address this issue, an example being the Chunity programming environment for Unity (Atherton and Wang 2018).

The scripting language that Unity uses is C# and the shader language is HLSL. The author's preferred shader language is GLSL. The Unreal engine scripting language is C++ which is the author's preferred coding language. However, attempting to learn how to use an entirely new environment was not practical. On reflection, it seemed that the established game engines, whilst very powerful, were too big and cumbersome. There was a realisation that a more streamlined approach was desirable. Also, in the author's experience, using software such as this sometimes takes the individuality out of the experience. There is a tendency to begin learning how to use Unity, or Unreal as opposed to creating audiovisual art. Further, it was felt that the large ecosystem of game engines such as these, creates a distance between the hardware itself and the user. As mentioned previously, the aim was to get closer to the hardware to achieve a greater sense of control over the computer itself.

7.1.2 Creative Coding Environments

There are several creative coding environments that are very popular with artists working in the computational art field. These include Processing[82], openFrameworks and Cinder[83]. Processing is based on Java and, as such, allows for quick development of visuals. It was traditionally developed for visual artists rather than audio developers. However, it does have generative audio capability using the *processing.sound* library[84].

OpenFrameworks is a staple of the creative coding scene and has a large community built around it. It has been used extensively in the previous work presented in this thesis, as discussed in Chapters 5 and 6. Throughout the author's previous work with openFrameworks, some issues arose related to the size of the codebase. This manifests itself in quite lengthy compile times during development. Further, openFrameworks is very visually oriented. Audio synthesis has been implement previously

[82] https://processing.org/ (accessed 26/04/2020).
[83] https://libcinder.org/ (accessed 26/04/2020).
[84] https://processing.org/tutorials/sound/ (accessed 26/04/2020).

with the Maximilian audio library[85], which worked very well. However, at the initial stage of development, there were few options for creating localised sound sources that were built into the library. There is an add-on for VR integration with openFrameworks called ofxOpenVR. This seems to be quite useful and easy to use. However, due to some previous experiences trying to debug openFrameworks there was a desire to explore other options.

Cinder is another creative coding framework built on C++, like openFrameworks. However, as the author has not used it before, there was a reluctance to learn another framework with its own syntax. Overall, whilst these creative coding frameworks are powerful and flexible, they can obscure some of the lower-level workings of the code. Again, there was a desire to gain more control over the computer and it was felt that knowledge of lower-level graphics APIs, in particular, would provide this.

7.1.3 Audio Synthesis Environments

Max/MSP/Jitter[86] and PureData/GEM[87] are two environments that are traditionally audio focused. However, they also have graphical capabilities. They are visual-coding environments that utilise a graphical patching system, as opposed to text-based environments, where the coding is done through text editing. Max/MSP/Jitter is a commercial product, whereas PureData/GEM is open-source. It was preferable to use open-source or freely available software, so a decision was made to not use Max/MSP/Jitter. PureData had been used by the author in previous work and it is considered quite flexible. PureData has also spawned libpd[88], which is an embeddable library for sound synthesis. This is a very useful library that can be integrated with other applications or environments, for example, Niall Moody's LibPdIntegration[89].

CSound[90] is a long-established audio synthesis environment that was created in 1984 by Barry Vercoe. It was the first audio language written in C and is an evolution of audio synthesis languages that came before it (Csound 2020). Over the years, CSound has developed a huge library of opcodes including sound localisation opcodes. This makes it an attractive possibility for use in VR. The CSoundAPI allows the developer to build CSound into any C++ application and run it on its own

[85] https://github.com/micknoise/Maximilian (accessed 26/04/2020).
[86] https://cycling74.com/ (accessed 26/04/2020).
[87] https://puredata.info/ (accessed 26/04/2020).
[88] https://github.com/libpd/libpd (accessed 21/04/2020).
[89] https://github.com/LibPdIntegration/LibPdIntegration (accessed 26/04/2020).
[90] https://csound.com/ (accessed 21/04/2020).

thread. This flexibility is very useful as it means, just like libPd above, that CSound can be incorporated into other applications.

There are several other text-based audio synthesis environments that provide very powerful and flexible options for the audiovisual composer. These include SuperCollider[91], ChucK[92] and TidalCycles[93]. These languages are text-based, and each have their own features, workflows and syntax. However, due to the author's inexperience in these languages they were not utilised for the work in this thesis.

7.2 System Requirements

In order to compose immersive audiovisual work in line with the artistic approach presented in this thesis, there was a need to identify important functionality that would lead to a successful workflow. Once this functionality was identified, an informed decision could then be made on the environment and technologies that were needed to explore the practice. The main objectives for the system grew out of the artistic concepts outlined in Chapters 3 and 4 and also the implementation of the IML control paradigm discussed in Chapters 5 and 6. Five basic requirements were identified for the desired workflow. These are illustrated in Fig. 7.1.

[91] https://supercollider.github.io/ (accessed 26/04/2020).
[92] https://chuck.cs.princeton.edu/ (accessed 26/04/2020).
[93] https://tidalcycles.org/ (accessed 12/04/2022).

Figure 7.1 Five aims for the ImmersAV Toolkit.

7.2.1 GLSL shaders

This requirement was essential. The work in this thesis is mostly abstract and generative. The most powerful way to generate real-time, complex visuals is to use a GPU. Through the earlier work with openFrameworks, the GLSL shading language was used. Therefore, it was preferable to continue the graphics development using this language. It was also discovered that the use of raymarching would allow for the creation of entire visual scenes solely within the fragment shader. This would help to provide a well-defined, focused environment in which the visuals could be developed. Further, raymarching techniques allow for dynamic morphing of physical objects and the physical modelling of light characteristics. This ability would lend itself well to abstract artistic visuals. The work of Inigo Quilez was a major influence in this decision (Quilez 2020).

7.2.2 Audio Synthesis Capabilities

The system needed to be capable of producing complex generative audio. This capability needed to be on par with the ability to produce generative graphics. The concept of audiovisual balance arose out of the desire to treat the audio and visual material equally. Equality in this sense means that both the audio and visuals were to command a similar level of attention during development. To interpret

this statement in a compositional manner, the methods by which both the audio and visual material was to be created, needed to be as unrestrictive and as powerful as possible. This would allow for the creation of both audio and visual material with a sufficiently wide expressive range. This level of control over the generation of material is important in order to give the impression that both the audio and visuals inextricably belong together and would be diminished in isolation. This is related to the concept of isolated structural incoherence discussed in Chapter 3.

7.2.3 External Libraries

The use of IML techniques to map input data to audio and visual parameters is an element of the core research question of the thesis. In order to continue exploring the use of IML technologies to create rapid complex mapping, this functionality needed to be built into the toolkit. Therefore, one of the requirements for the immersive toolkit workflow was the ability to integrate external libraries such as the RapidLib library.

7.2.4 Omni-directional Mapping Capabilities

In Chapter 3, audiovisual balance was identified as an important concept within the work presented in the thesis. This concept posits that there is a balance between the audio and visual material that can be manipulated to maximise the potential for added-value experiences. The work of Sá was referenced, who demonstrated the natural sensory dominance of sight over sound in human perception (Sá 2016: 30). It was proposed that this dominance can be manipulated at two points in the audiovisual workflow. Firstly, and as Sá discussed, the arrangement and character of the audio and visual material can affect the audience's perception of sensory dominance. Secondly, as discussed by Callear, the mapping method can also affect the hierarchy of the material (Callear 2012: 31). It is important for the material to be treated equally. Therefore a hierarchy is undesirable. Following from this, the architecture of the audiovisual system may be a location where audiovisual balance can be affected.

If the mapping capabilities of the system are limited, then the composer's ability to treat the material equally may be restricted. There was a need to make sure that the system chosen to create audiovisual work provided easy control of parameter mapping, both from the audio stream to the visual stream, and vice versa. It should also provide a straightforward way to implement algorithms that would allow for the mapping of data to both the audio and visual streams simultaneously. This would allow for multiple layers of mapping that could create interest and complexity in the piece.

The omni-directional mapping capabilities of the ImmersAV toolkit are demonstrated using walkthrough examples included in the GitHub repository README file. There are also links to video

demonstrations of audio reactive mapping[94], simultaneous mapping of a central data source[95] and examples of cyclical mapping[96,97]. These video examples can also be found in the media pack in the folder called *mediaFiles/immersAV*. The audio reactive mapping file is called *immersAV_audioReactive.mp4*; the simultaneous mapping example is called *immersAV_simultaneousControlValEx.mp4*; the cyclical mapping examples are called *immersAV_cyclicalEx_1.mp4* and *immersAV_cyclicalEx_2.mp4*.

7.2.5 Single-purpose Environment

As discussed above, the audiovisual composer has a wealth of environments to pick from when creating their work. Whilst these environments are extremely powerful and flexible, they are not tailored specifically for audiovisual composition. The game engines, Unity and Unreal Engine, cater for the gaming community and, as such, are primarily visual environments focused on the development of game mechanics and graphics. Only recently are they beginning to integrate tools and libraries for generative audio creation in the form of libpd and Chuck for Unity and FMOD for both Unity and Unreal. The all encompassing nature of these environments also means that a generative audiovisual composer would only be using a very small part of their functionality. Ultimately, a single-purpose environment, built specifically for creating abstract audiovisual art with minimal overhead was required. The ImmersAV toolkit has provided that.

7.3 Development

Considering all of the options and requirements discussed in the previous sections, a toolkit, built in C/C++, was created. This was done for the following reasons:

- A minimal application that focuses on both audio and visuals, and the mapping of data in any way between them was required.
- There was a desire to simplify the creation of material to take place directly in specific, well-defined locations.
- There was a desire to gain more comprehensive mastery of the computer as an audiovisual instrument.

[94] https://www.youtube.com/watch?v=nAH7aKPWTZw (accessed 29/09/2020).
[95] https://www.youtube.com/watch?v=ot0BNak_W6g (accessed 29/09/2020).
[96] https://www.youtube.com/watch?v=7E4uOEJfCEg (accessed 29/09/2020).
[97] https://www.youtube.com/watch?v=Zm7Ipx0HCCg (accessed 29/09/2020).

In the author's experience, when creating music, the musician should feel a connection between their actions and the instrument. Previously, when using any of the software discussed above, a disconnect was felt between the act of creating and the end result. There was always a level of abstraction that created a distance between the artist and the computer. Perhaps this comes from embodied experience playing music. When playing an instrument, the musician affects sound waves in a very direct and tactile way. There was a desire to use the computer as a musician would use an acoustic instrument. Therefore, there was a motivation to strip away any and all extra layers of abstraction. C++, OpenGL and the libraries discussed throughout this chapter are all considered to be high-level abstractions in terms of programming languages. However, even working with slightly lower-level code gave the illusion of more control, and the sense that the computer was being used as an instrument at a deeper level. This process came to be an extremely empowering and important stage in the development of the subsequent artworks. In the following sections, some of the features of the ImmersAV toolkit will be discussed.

7.3.1 Architecture

As discussed above, a clear idea of the desired workflow was formulated before development began. This workflow existed in a proto form for *Ventriloquy I* and *II* discussed in Chapter 6. The basic workflow is as follows:

- Create an area in the code that processes audio.
- Create an area in the code that processes the visuals.
- Map between these two processing blocks by sending data to parameters simultaneously.

There was a desire to further consolidate this workflow into more well-defined environments within which each element of a piece could be developed. It was decided that the ideal system would allow for the creation of all the visuals in one class, all the audio in another class and finally that data could be mapped to any part of the system from any other part of the system. Fig. 7.2 shows the conceptual structure of the system.

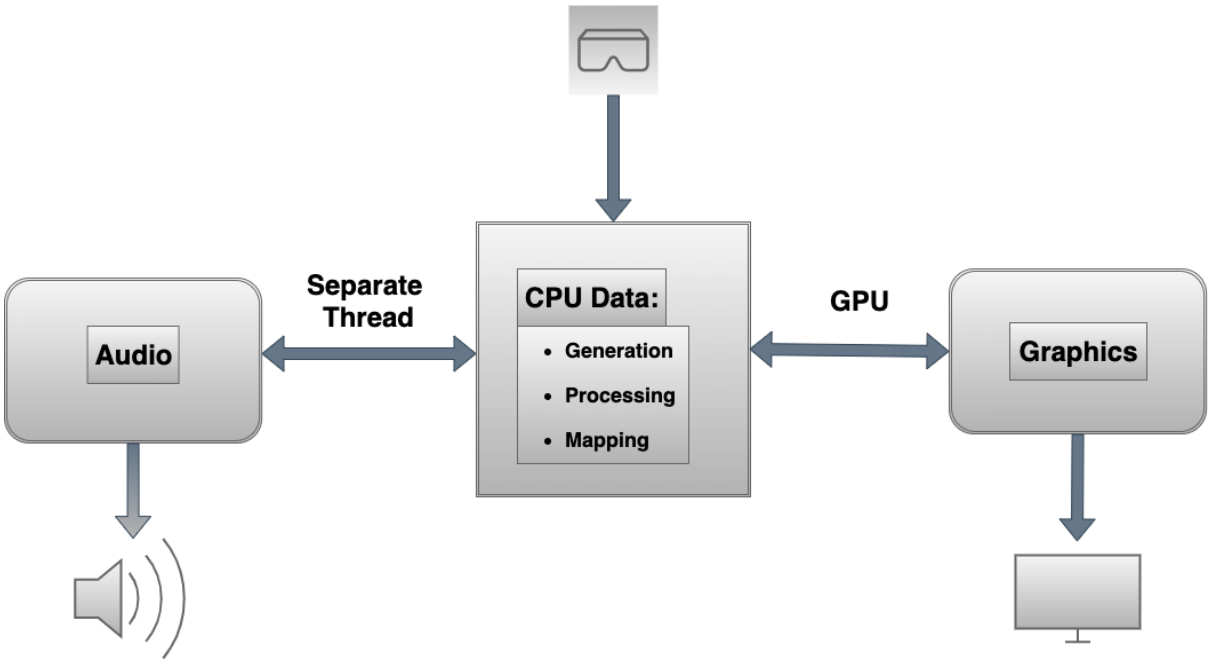

Figure 7.2 Diagram of system structure.

Hardware sensors output data from the headset and controllers to the central CPU data module. The audio and graphics modules can also send data to the central module. Here data is processed from various sources. New data can also be generated and mapped simultaneously to both the audio and graphics modules. The audio module operates on its own thread on the CPU. All audio is generated here. The graphics are generated entirely on the GPU. Any data going to or coming from the graphics module must be uploaded to or downloaded from the GPU. Audio parameters can be sent to the graphics module by way of the central module. Data from the graphics module can also be sent to the audio module via the central module. Audio is then output to the speakers and graphics are output to the visual display.

This structure complements the concepts of audiovisual balance that form the basis of the approach to audiovisual composition in this thesis. Omni-directional mapping in this form would provide the desired freedom for manipulating the material in the desired way.

7.3.2 Technologies and Techniques

In order to implement the structure of the system, there was a need to decide on a number of factors. There were decisions to be made on the graphics API, the audio engine, the VR SDK and the machine learning library. Here follows an account of the technologies and libraries that were used to create the ImmersAV toolkit.

OpenGL

OpenGL was chosen as the graphics API. This decision was made due to the fact that GLSL shaders had been used for *Ventriloquy I* and *II*. OpenFrameworks is built on top of OpenGL and one of the reasons for developing a native application was to work with code at a lower-level. Although OpenGL is still a relatively high-level API it still afforded a clearer route for interaction with the GPU, than using any of the above options. Further, this was treated as a learning opportunity in artistic terms. It was posited that a deeper knowledge of one of the main graphics APIs would foster more complete knowledge of graphics programming and the rendering pipeline. In terms of using the computer as an instrument, this deeper knowledge could only be beneficial to the practice.

However, OpenGL has its limitations. Sometimes it is cumbersome and it is notoriously difficult to debug. Also, macOS has stopped supporting it so it is unclear how much of a future it has as a truly cross-platform API. However, OpenGL 4.1 can currently be used on macOS 11 (Big Sur) and Windows 11.

Raymarching

GLSL shaders have been around for a long time and are small programs that run in parallel on each pixel fragment on the GPU. There was a desire to explore the technique of raymarching which was found to be a very expressive form of graphical programming that lends itself well to abstract audiovisual art.

Raymarching is a rendering technique that was made popular in the *demoscene*. The demoscene is a long running community of coders and artists that compete to create short, real-time audiovisual clips that are technically accomplished. They apply severe limitations to the size of the finished application which makes the complex pieces even more impressive. Due to hardware restrictions, demoscene coders had to come up with ingenious methods to create more impressive visuals. As GPUs began to develop into powerful hardware tools, some coders began to utilise the massive potential of parallel processing offered by GPUs. One way to harness this power was to use the fragment shader to execute an optimised form of ray-tracing.

While ray-tracing is extremely accurate and can create very realistic graphics, it is computationally expensive as the exact point of intersection of the ray needs to be calculated every time it hits an object in the scene. However, the raymarching algorithm estimates where the nearest object is by casting a ray into the scene. This is illustrated in Fig. 7.3. This means that the ray can be tested for intersection without having to calculate the exact point of intersection.

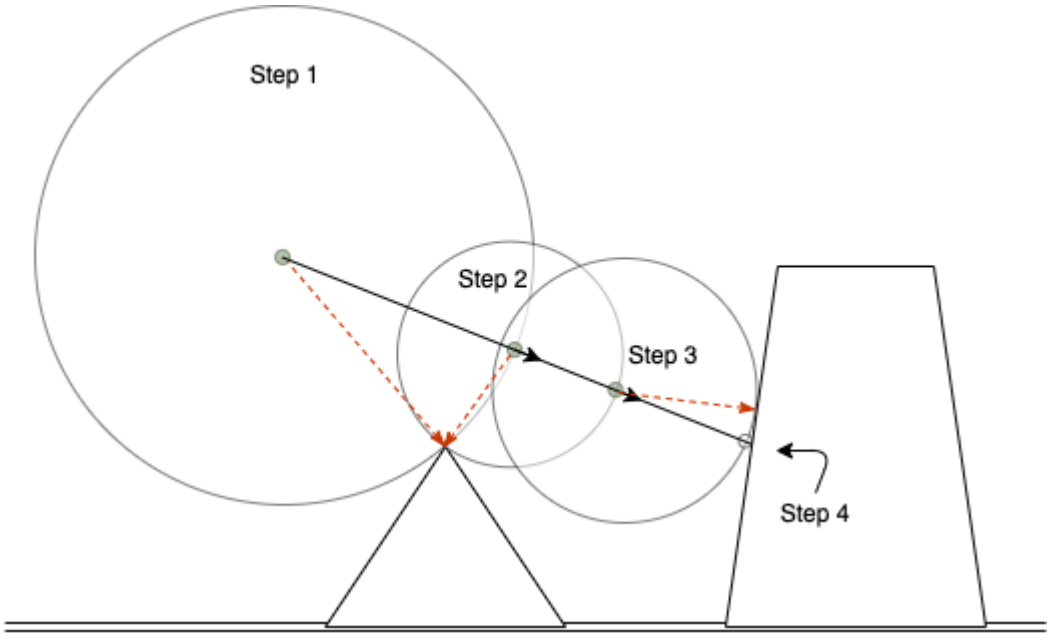

Figure 7.3 Raymarching process [diagram based on Pharr (2005: Fig 8-5)].

This approach was championed by Inigo Quilez who has written extensively, and given many tutorials on the subject. Quilez's work demonstrates, to a high level, the expressive and artistic possibilities of the raymarching technique. All the code needed for a complex scene can be written in a single fragment shader file. The technique does not rely on vertices and polygons to create objects. This means that you can create malleable shapes and terrains, capable of evolving and morphing continuously in time, that lend themselves well to dynamic abstract visualisations.

Asynchronous GPU read back

An important feature of the ImmersAV toolkit is that it provides a mapping class that allows the artist to map freely from any part of the application to any other part. In order to achieve this functionality, asynchronous read-back from the GPU to the CPU needed to be implemented. This technique involves the use of a framebuffer object (FBO) and a pixel buffer object (PBO). The PBO is a distinct OpenGL object that is specifically intended to allow for transferring data between the CPU and GPU[98]. The pixel information from the shader is rendered as a texture to a temporary FBO. The data is then copied to the PBO which allows for asynchronous mapping of data to a memory location on the CPU. This has to be done asynchronously, as to synchronise it with the frame rate would cause severe delays in the rendering process, causing the graphics to freeze.

[98] https://www.khronos.org/opengl/wiki/Pixel_Buffer_Object (accessed 12/10/21).

CSound

As discussed above, CSound is a well-established environment for creating synthesised audio. The reasons for choosing CSound as the audio engine are as follows:

- The author had previous experience using it, so there was no need to learn a new language from scratch.
- The CSound API and cSoundPerformanceThread[99] provide an easy way to build Csound into the toolkit and run the audio on its own thread.
- The CSound workflow lent itself well to the overarching conceptual and practical aims of the ImmersAV toolkit.
- CSound has an enormous library of opcodes that are ideal for creating generative audio.
- There are several HRTF opcodes that allow for sound placement in a virtual environment.
- There is a very active community that corresponds through the CSound mailing list.

With regard to the above points, it was felt that CSound was more than capable of providing a scalable and modular audio environment that could be used as the main audio engine for the ImmersAV toolkit. Although it was previously stated that there was a desire to work with code at as low a level as possible, it seemed that it was not necessary to implement a lower level engine than CSound.

The Maximilian library, used in *Ventriloquy I* and *II* was also a viable option as it could be easily integrated into a C++ application. However, at the time of development, there were no HRTF functions built into the library. Further, the CSound csd file, which represents the consolidated orchestra and score files, provided an attractive audio counterpart to the fragment shader. The idea of building all of the audio processing in the csd file, all of the visuals in the fragment shader, and finally, all of the mappings in another central file or class drove the conceptualisation and development of the toolkit. This structure was logical, and also aligned with the artistic principles outlined earlier in the thesis, in that it would allow for individual control of all the elements needed to create an audiovisual work. However, at the same time each individual element would be easily accessible to each other.

[99] https://csound.com/docs/api/index.html (accessed 13/04/2022).

Interactive Machine Learning

IML was a significant element of the workflow that led to *Ventriloquy I* and *II* and is a central part of the research in this thesis. Here are some reasons for including IML capabilities and more explicitly, the RapidLib library:

- IML is a relatively new area of interest within the creative arts and there is much yet to discover and explore. Allowing easy access to this functionality in an immersive audiovisual toolkit would hopefully lead to interesting work.
- The RapidLib library is written in C++ and can be easily integrated into the wider application.
- Previous experience was gained working with RapidLib and the author was familiar with its workflow.
- As discussed in Chapter 5, using a regression algorithm to quickly map nonlinear parameters is very effective. This characteristic fits well with the main aim of the toolkit in that it will provide powerful mapping functionality.

As mentioned above, the use of IML in creative work is relatively new. However, it is also extremely popular and is growing in popularity. This newness makes the technology very exciting due to the fact that there are so many possibilities for its use.

OpenVR

When development started on the ImmersAV toolkit, the target headset needed to be specified. The mobile headsets were discounted as there was a need for the full power of a PC GPU. Following from this, there were two main headsets in the PC market. These were the Oculus Rift and HTC Vive Pro. Since October 2018 the market has changed and there are several new headsets available with better specs than either of these products. Given the choice between Oculus and HTC, it was decided to target the Vive Pro for several reasons.

- It had better resolution than the Oculus Rift at the time.
- The OpenVR SDK could be used to communicate with it.

OpenVR was more attractive than the Oculus SDK as OpenVR has the potential to work with several headsets whilst the Oculus SDK can only be used with Oculus products. This potential for OpenVR to work with several headsets means that the ImmersAV toolkit does not have to necessarily be tied to the Vive Pro in future.

7.3.3 Workflow

The ImmersAV toolkit was built from the very start with a specific workflow in mind. As stated above, a minimal environment was needed, where audio and visuals could be developed in their own contexts and then parameters and data could be easily mapped from either processing context at any time. With the libraries, SDKs and API discussed above, this goal could be achieved. Here, the workflow suggested by the toolkit will be discussed. Some sections of code will be presented that are important to the flow of the toolkit. The GitHub repository, README.md file, contains further detailed information.

Studio

The *Studio()* class is conceptualised as the central workspace where the audiovisual artist binds the audio and visual material together. This is where the audiovisual composition happens. It was important that this workspace is as clean as possible with a minimum of boilerplate code needed to link various parts of the toolkit. It contains three functions that are modelled on the openFrameworks *ofApp()* class. The three functions are:

- Studio::Setup()
- Studio::Update()
- Studio::Draw()

Just as in the openFrameworks runtime, *Setup()* runs once before the first frame. Then *Update()* and *Draw()* each run once per frame in that order.

Setup

As can be seen in Ex. 7.1, the *Studio::Setup()* function contains the CSound performance thread initialisation. A new pointer is declared at line 00 that gives access to the *StudioTools()* class. This is the class that contains all the functionality needed to communicate with Csound and the OpenGL draw commands. The function, *PCsoundSetup()*, returns a pointer called *csSession* which can be used to communicate with the CSound instance from the *Studio()* class. It also receives the name of the csd through the function arguments. CSound is now running completely within its own thread which is desirable so as not to interfere with other processes.

```
00 m_pStTools = new StudioTools();
01
02 //audio setup
03 CsoundSession* csSession = m_pStTools->PCsoundSetup(csd);
04
05 if(!m_pStTools->BSoundSourceSetup(csSession, NUM_SOUND_SOURCES))
06 {
07     std::cout << "Studio::setup sound sources not set up" << std::endl;
08     return false;
09 }
```

Example 7.1 Csound thread initialisation.

Now that there is a pointer to the CSound instance, it is possible to set up the sound sources for the scene. This is done by calling *BSoundSourceSetup()* and passing the *csSession* pointer and the number of sound sources to be included in the scene. After setting up the sound sources, it is now possible to send data to, and receive data from, CSound. This is achieved through the *BCSoundSend()* and *BCSoundReturn()* functions. These functions take the *csSession* pointer, a vector of type *const char**, for channel names, and a vector of type *MYFLT**, for the float data. The CSound instance is now set up with straightforward methods to send and receive data. The code for this functionality is shown in Ex. 7.2.

```
00 //setup sends to csound
01 std::vector<const char*> sendNames;
02
03 sendNames.push_back("sineControlVal");
04 m_vSendVals.push_back(m_cspSineControlVal);
05
06 sendNames.push_back("randomVal");
07 m_vSendVals.push_back(m_cspRandVal);
08
09 m_pStTools->BCsoundSend(csSession, sendNames, m_vSendVals);
10
11 //setup returns from csound
12 std::vector<const char*> returnNames;
13
14 returnNames.push_back("pitchOut");
15 m_vReturnVals.push_back(m_pPitchOut);
16
17 returnNames.push_back("freqOut");
18 m_vReturnVals.push_back(m_pFreqOut);
19
20 m_pStTools->BCsoundReturn(csSession, returnNames, m_vReturnVals);
```

Example 7.2 Csound send and return setup.

The *Setup()* function is also responsible for initialising the code needed to communicate with the GPU. The function *RaymarchQuadSetup()* is responsible for creating a quad for the fragment shader to raymarch onto. This is shown in Ex. 7.3. *RaymarchQuadSetup()* is passed a pointer to the shader program, which is accessed through the *Setup()* function arguments. Once this function is called, the necessary OpenGL initialisation of the quad is completed. The artist is now free to send data to the shader through any uniforms they may need. This is done by calling the OpenGL function *glGetUniformLocation()*, passing the shader program pointer and specifying the uniform name. This call returns a handle for the uniform that can be used as a way to send data to the shader.

```
00 //setup quad to use for raymarching
01 m_pStTools->RaymarchQuadSetup(shaderProg);
02
03 //shader uniforms
04 m_gliSineControlValLoc = glGetUniformLocation(shaderProg,
"sineControlVal");
05 m_gliPitchOutLoc = glGetUniformLocation(shaderProg, "pitchOut");
06 m_gliFreqOutLoc = glGetUniformLocation(shaderProg, "freqOut");
07
08 //machine learning setup
09 MLRegressionSetup();
```

Example 7.3 Raymarching and machine learning setup.

As part of the cyclical mapping example[100], a video demonstration was created, of how an IML regression algorithm can be used. The video file can also be found in the media pack[101]. This example is also discussed in the section entitled *Cyclical Mapping Example* on the ImmersAV Github README[102]. The machine learning code is initialised by calling *MLRegressionSetup()*. This simply initialises a set of bool types that enable the IML workflow to be controlled either from a laptop keyboard or a Vive controller.

Update

The *Update()* function runs before *Draw()* once every frame. This is where any data is updated before being drawn to the screen. The function arguments, *controllerWorldPos_0*, *controllerWorldPos_1*, *controllerQuat_0* and *controllerQuat_1*, provide access to the positions and rotation quaternions of the Vive controllers. This data can be used as control data for audio or graphical processes.

[100] https://www.youtube.com/watch?v=7E4uOEJfCEg&t=1s (accessed 30/09/2020).
[101] mediaFiles/immersAV/immersAV_cyclicalEx_*.mp4.
[102] https://github.com/bDunph/ImmersAV (accessed 18/10/2021).

The struct, *SoundSourceData*, is used to specify the position in world space, position in camera space, azimuth, elevation and distance in camera space, of a sound source. This data is used by *SoundSourceUpdate()* to place sound sources in the virtual environment. *SoundSourceUpdate()* accepts the view matrix and a vector of type *SoundSourceData* as arguments to create the sound sources. The position of the sound sources can then be manipulated by changing the *SoundSourceData.position* member from the *Update()* function. This can be seen in Ex. 7.4.

```
00 // example sound source at origin
01 StudioTools::SoundSourceData soundSource1;
02 glm::vec4 sourcePosWorldSpace = glm::vec4(0.0f, 0.0f, 0.0f, 1.0f);
03 soundSource1.position = sourcePosWorldSpace;
04 std::vector<StudioTools::SoundSourceData> soundSources;
05 soundSources.push_back(soundSource1);
06
07 m_pStTools->SoundSourceUpdate(soundSources, viewMat);
```

Example 7.4 Sound source creation.

In the example, *Simultaneous Data Mapping*[103], a description of which is found on the GitHub README page, there is a control signal being generated by *sin()*. This is shown in Ex. 7.5. A video of this example can also be found in the media pack[104]. This is a simple sine function that returns a different float each frame that approximates oscillating motion. This is an example of using the *Update()* function to generate a source of data that can then be sent simultaneously to the audio and visual processes. The return float is assigned to the *float* variable, *sineControlVal*. This will be used directly in the *Draw()* function to send the data to the fragment shader. On the next line, *sineControlVal* is cast to type *MYFLT*, which is a CSound type. This is then assigned to the first index of the *m_vSendVals* vector. This value can then be accessed from the CSound *csd* file and used to control audio parameters.

```
00 //example control signal - sine function
01 //sent to shader and csound
02 m_fSineControlVal = sin(glfwGetTime() * 0.33f);
03 *m_vSendVals[0] = (MYFLT)m_fSineControlVal;
```

Example 7.5 Data source controlling parameters.

The *Update()* function is also where the artist would update any values that need to be sent to the machine learning algorithm. The function *MLRegressionUpdate()* is called and passes three

[103] https://www.youtube.com/watch?v=ot0BNak_W6g (accessed 30/09/2020).
[104] mediaFiles/immersAV/immersAV_simultaneousControlValEx.mp4.

arguments. The first argument is a reference to a struct of type *MachineLearning*. This struct holds a number of *bools* that allow the artist to control the IML workflow. It is passed through the *Update()* function arguments. The second argument to *MLRegressionUpdate()* is a reference to the struct *PBOInfo*. This gives the machine learning algorithm access to data returned from the shader using the *PBOInfo.pboPtr** member. This is also passed through the *Update()* function arguments. The third argument is a vector of type *MLAudioParameter*. This is a vector that specifies input and output parameters to be processed by the regression algorithm. This is specified in *Update()* as shown in Ex. 7.6.

```
00 //run machine learning
01 MLAudioParameter paramData;
02 paramData.distributionLow = 400.0f;
03 paramData.distributionHigh = 1000.0f;
04 paramData.sendVecPosition = 1;
05 std::vector<MLAudioParameter> paramVec;
06 paramVec.push_back(paramData);
07 MLRegressionUpdate(machineLearning, pboInfo, paramVec);
```

Example 7.6 Defining parameters for machine learning functionality.

As mentioned above, the *Update()* function gives the artist access to data that is being returned from the fragment shader. This data can be accessed through the use of a pointer. The size of the buffer can be obtained using *PBOInfo.pboSize*. This gives the size of the Pixel Buffer Object (PBO) which is the buffer that OpenGL writes to. The individual pixel data can then be accessed by dereferencing the pointer. There are four values for each fragment, accessed in RGBA order. These values can then be used as control data for audio or machine learning processes.

Draw

The *Draw()* function draws the graphics to the screen. This function is called after *Update()* once every frame. The function arguments give the artist access to the projection matrix, view matrix and eye matrix, for headset eye-displacement. The camera translation vector is also provided for locomotion, and finally, the shader program handle is provided.

To begin the drawing process, *DrawStart()* is called. Each of the matrices are passed along with the shader program handle and the translation vector. This sets up the OpenGL context for drawing. To finish drawing, *DrawEnd()* is called. This cleans up the OpenGL context and signifies that all elements have been drawn for that frame. This is shown in Ex. 7.7.

```
00 m_pStTools->DrawStart(projMat, eyeMat, viewMat, shaderProg,
translateVec);
01
02 glUniform1f(m_gliSineControlValLoc, m_fSineControlVal);
03 glUniform1f(m_gliPitchOutLoc, m_fPitch);
04 glUniform1f(m_gliFreqOutLoc, *m_vReturnVals[1]);
05
06 m_pStTools->DrawEnd();
```

Example 7.7 Draw calls.

To send data to the shader, the artist calls the OpenGL function, *glUniform1f*, between *DrawStart()* and *DrawEnd()*. The uniform handles retrieved in *Setup()* are used here. In the *Simultaneous Mapping* example, the *sineControlVal* data signal is being sent to the GPU using the variable from *Update()*. In the *Audio Reactive* example, the *m_vReturnVals* vector is being directly accessed to retrieve an RMS value from CSound before being sent to the shader.

As discussed previously, all of the audio and visual processing takes place in the *csd* and *fragment shader* files respectively. The *Studio()* class discussed here facilitates easy mapping of data between these two processing locations. The *Studio()* class is the main hub where the artist has access to sends, returns, machine learning and controller information. This provides a unified workspace where the artist can concentrate on creating interesting and dynamic mapping strategies. The examples folder in the GitHub repository include example vertex and fragment shader files. They also contain the example CSound files. These are setup to work specifically with the ImmersAV toolkit and can be used as templates.

7.4 Future developments

There is much work to be done to refine and expand the functionality of the toolkit. It is still in an early stage of development and there are many improvements that it could benefit from.

In terms of the organisation of the code, a cleaner, more modular code base would be desirable. As it stands, the code does not make enough use of modern C++ and object-oriented features. It has been described as C-style C++ in personal correspondence with a professional software engineer. In terms of features, there is a desire to implement an option to run on the Vulkan API which will become more important as OpenGL loses support. It would also be beneficial to include an option to use the Oculus SDK to provide a wider range of hardware compatibility. Finally, it may be beneficial to implement a simple GUI using JUCE or Qt, that would allow the user to use the *Studio()* class in a

more visual way. It may be beneficial to keep the text-based audio and visual processing whilst implementing a graphical way of mapping data.

7.5 Conclusion

The development of the ImmersAV toolkit evolved in response to the core research aims of the thesis. The tools explore the use of IML techniques to control audiovisual compositions in VR were separated into several different environments. This toolkit provides a centralised environment where the core research aims can be realised.

Part of the motivation to specifically shape the workflow of the toolkit, in the way discussed above, was also to place some constraints on the creative process. It is posited that constraints can foster creativity in areas such as digital media (Candy 2007) and poetry (Bauer 2018). This motivation manifested itself in the decision to focus on creating well-defined areas of work. It also manifested in the creative decision to constrain the visual process and concentrate solely on raymarching. It was hoped that the act of adhering to these constraints would allow for more focused work. Although the structure of the toolkit is straightforward, the creative possibilities afforded by raymarching, CSound and omni-directional mapping are vast. In creating the ImmersAV toolkit in this manner, the aim is to situate the audiovisual composer in the face of this vastness and provide a place to start.

The contribution of the ImmersAV toolkit to the field is the way in which it provides a minimal workspace for the audiovisual composer to create immersive work that incorporates the IML control paradigm. It is hoped that this toolkit can aid artists in their own work, and can provide an alternative route to creating immersive audiovisual work than what is currently available. The toolkit also provides the means by which the rest of the compositions in the portfolio were created. Chapter 8 will discuss the first of these pieces.

Chapter 8 Immersive Audiovisual Composition: *Obj_#3*

This chapter will discuss the piece *Obj_#3* that was created using the *ImmerAV* toolkit. The concepts behind the piece will be discussed, followed by some important compositional and technical features. The source code for this piece can be found at its GitHub repository[105] and in the accompanying media pack under *sourceCode/obj_3*. There is also a prebuilt Windows 64bit binary file in the *sourceCode/obj_3_win64Bin* directory. The example videos throughout this chapter are linked to YouTube and can be found in the accompanying media pack under *mediaFiles/obj_3*.

The primary topic of inquiry in this thesis, as posed in Chapter 1, is the question of how to employ machine learning techniques to control audiovisual compositions in the emerging, fully immersive medium of VR. *Obj_#3* is the first full realisation of the core research aims. This takes the form of a fully immersive and interactive, audiovisual sculpture. It is a culmination of several attempts to create a virtual environment in which there exists a sense of presence. It also builds on the exploration of audiovisual balance and considers the medium-specific question of how to incorporate presence into the compositional process. *Ventriloquy I* and *II* were constructed using basic three-dimensional cubes and spheres. *Obj_#3* marks a departure from these simple shapes and explores a generative visual form that does not exist in the physical world. The choice of audio material is similarly synthetic. This approach was motivated by Slater and Sanchez-Vives' concept of VR as "an *unreality simulator"* (Slater and Sanchez-Vives 2016: 6). The balance between the *real* and *unreal*, as presented by Latham et al. (2021) and discussed in Chapter 4, also plays a significant role in the composition. The material in this piece consists of *environmental* and *foreground* elements. The environmental material consists of elements that make up the surrounding environment, whereas foreground material consists of the elements that make up the audiovisual sculpture. The development of these elements will be detailed followed by a discussion of the mapping layers and interaction. The public presentation of the piece is then documented followed by a discussion around the AV-participant feedback and compositional aspects of the piece.

8.1 Visual Elements

This section will outline the development of the visual elements of the piece. The discussion will begin with the material that makes up the wider environment. This will be followed by a discussion of the foreground visual elements.

[105] https://github.com/bDunph/obj_3 (accessed 29/09/2020).

8.1.1 Environmental Material

Slater (2009: 3554) states that tactile response, 'correlated with vision', can be said to enhance place illusion (PI). A significant goal of this composition was to create a strong sense of PI and plausibility illusion (Psi) for the av-participant (Slater 2009). Therefore, the environmental material was carefully considered, as it could be an important contributing factor in whether the AV-participant would experience a sense of presence or not.

It was decided to situate the AV-participant on a horizontal surface, as the intention was to present the piece with the AV-participant in a standing position on a physical surface. It was hoped that because the AV-participant would proprioceptively feel that they are standing on a solid surface in the real world, the sensation of presence would be reinforced. However, there is a conceptual difficulty with this approach as it relates to abstract audiovisual composition. By creating an environment that situates the AV-participant, representation is being introduced into the piece. As soon as a surface is created for the participant to stand on, an element of mimesis is present. There is tension between creating the amount of physical realism necessary to foster presence and the exploration of abstraction. This dilemma is representative of the difficulties in moving from a screen-based practice to an immersive practice. It is an example of the inherent differences between the more stable and understood languages of screen-based art and the still unstable and evolving language of immersion.

As discussed in Chapter 4, the majority of audiovisual work has historically existed with some sort of screen or surface on which visual material is projected. The audioviewer is situated in the screening room, theatre or in front of their computer. The audiovisual composer does not have to build the environment for the audioviewer. This changes in an immersive context. The AV-participant must be situated somewhere in the virtual environment in order to experience a sense of presence. This is a characteristic of the medium and, as discussed in Chapter 4, is central to the experience of the AV-participant. In terms of screen-based abstract audiovisual art, everything on the screen is part of the piece. The audiovisual composer can also include sounds and imply they originate from off-screen artefacts. The audiovisual composer is responsible for integrating all the elements into a coherent whole. This is a central principle of the practice presented in this thesis, and needs to be adhered to in an immersive context, just as it does in a non-immersive context. This means that, ideally, the environment itself should be part of the piece. However, if the environment is essentially representational of a world, then how can implicit associations be avoided, which would distract the participant from the interaction of the audio and visual material? As discussed in Chapter 3, the use of completely abstract material when creating audiovisual work, is a strategy used by some to avoid

the complexity of semiotic, representational material. Further, regarding the idea of isolated structural incoherence, it was posited that if an element of the piece is structurally coherent in isolation, and remains so for the duration of the piece, it could act as a barrier to strong cross-modal binding.

However, the transition from theory to practice is never completely seamless. Theoretical concepts such as isolated structural incoherence exist in an ideal state. Artistic practice does not exist in this ideal state. Therefore it is often necessary for the composer to compromise. In light of this reality, the goal here was to create a pseudo-abstract landscape that attempts to guide the focus of the AV-participant, encouraging them to concentrate on the audiovisual sculpture rather than the environment. This was necessary, whilst at the same time providing enough sensory details within the environment to lay the foundations for a successful sense of presence. A balance between these concerns was attempted.

Environment Development

The surrounding material that made up the environment needed to be as unobtrusive as possible so as not to distract from the audiovisual sculpture. As discussed above and in Chapter 4, it was crucial to attempt to create a sense of presence, or more specifically, a strong sense of PI. Initially a white room was created in which to place the object. This was intended to mimic the contemporary gallery environment which would typically consist of exhibits displayed in white rooms. This seemed like an appropriate, although potentially obvious, jumping off point.

In order to construct this viewing room, a technique called cube mapping was used to attach textures to the inside of a cube. The AV-participant was then placed in the middle of the cube. See Fig. 8.1.

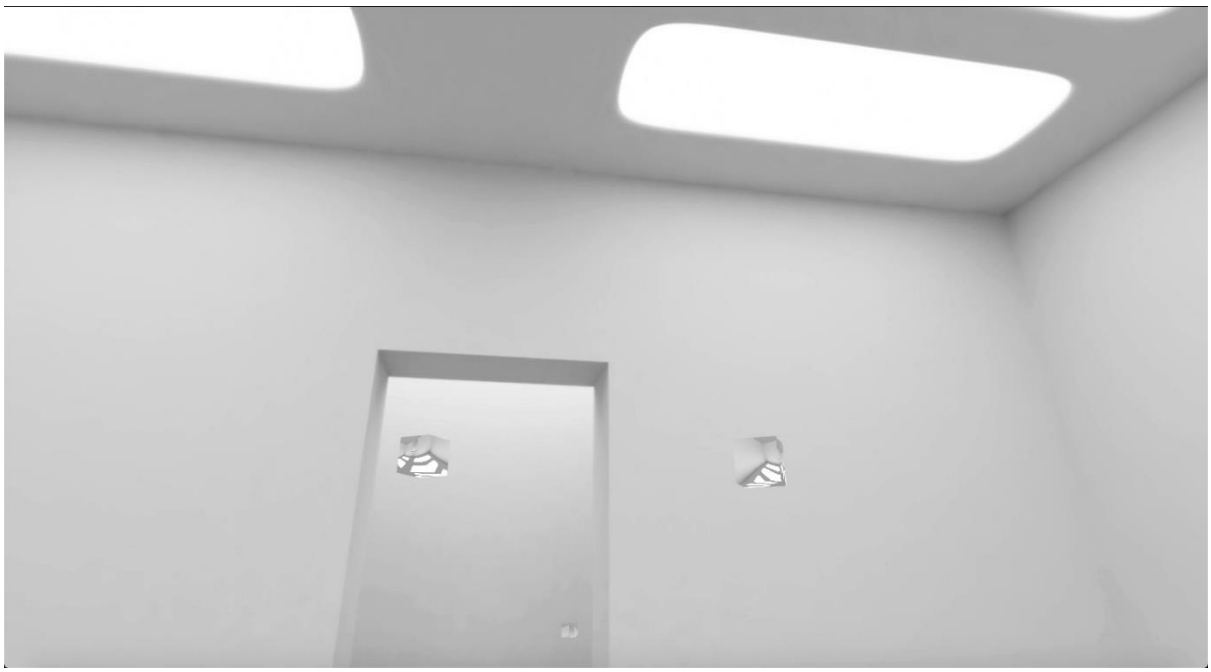

Figure 8.1 White room environment.

This approach presented a few problems. When creating a room in this way, the method used to map the cube textures meant that the texture coordinates were placed at infinity (see Appendix A.1 for the code used to achieve this). This means that when moving through the space, the perspective of the room did not react accordingly. This created a conflict between physical movement and the sensory information arriving at the eyes. It was felt that this would impede the sense of presence in the final piece. Another problem with this environment was that there was a doorway in one of the walls. This could imply that the AV-participant could go through the door. However since the extremities of the room were at infinity, this would have been impossible. This element of the environment could potentially frustrate the AV-participant and might distract them from the focus of the experience which was intended to be the specific audiovisual object in the space. See the example video *White Room Example Environment*[106] (*mediaFiles/obj_3/whiteRoomExample* in the accompanying media). As the camera moves through the space there is no sense that it is moving through the room. The only way movement can be sensed is when the reflective cubes are in the visual field and are moving relative to the camera. The door or walls of the room never move closer.

These factors precipitated a re-evaluation of the approach to the construction of the environment. Moving away from the idea of the white room, it was decided to place the audiovisual object on a simple plane. A desert plane was initially considered, featuring mountains in the distance (see Fig. 8.2). The mountains and sky were rendered using the same technique as the white room. That is, they

[106] https://youtu.be/aGhGHpGPdrk (accessed 02/09/2020).

were part of a cubemap texture situated at infinity to give the illusion of a fully-realised 360 degree environment. A separate textured quad was then rendered as the ground plane in order to provide a sense that the AV-participant was able to move within the environment. The shading and texture on the ground plane provided the correct changes in perspective to indicate movement, whereas the cubemap textures were stationary but gave the effect of a surrounding landscape. It was hoped that this would locate the AV-participant within the setting but also provide no implication that there was anything to explore except for the foreground audiovisual sculpture. In the video example *Desert and Mountain Environment Example*[107] (*mediaFiles/obj_3/desertAndMountainEx* in the accompanying media pack) notice the textured desert surface as the camera moves through the environment.

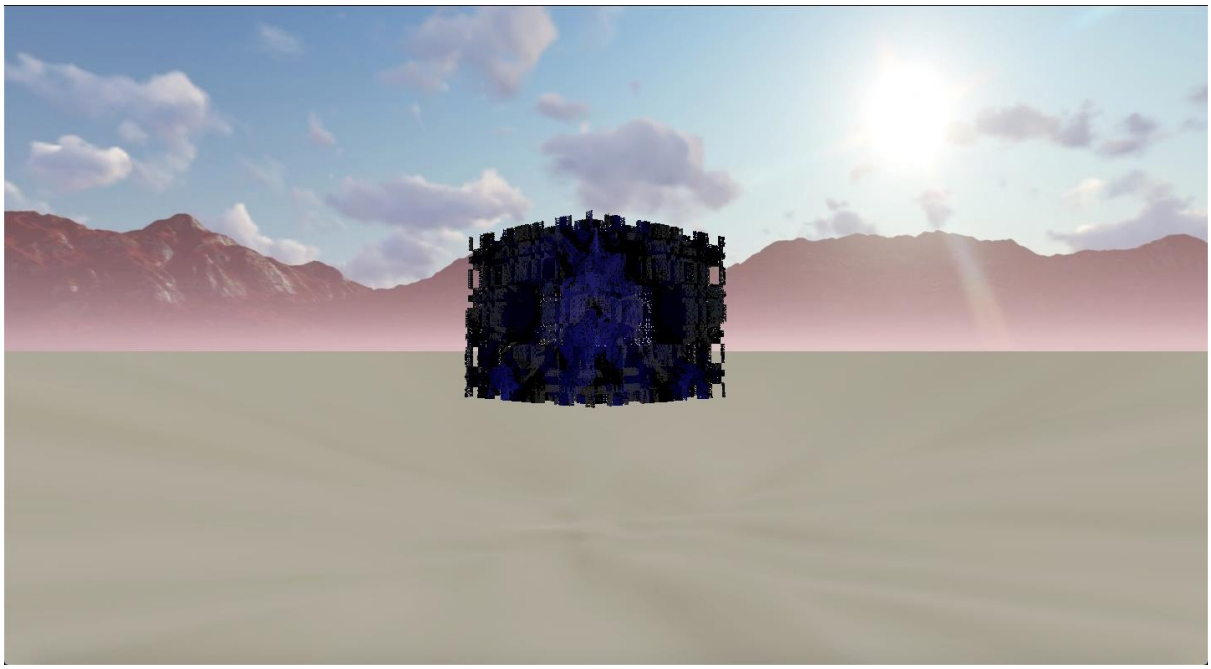

Figure 8.2 Desert and mountain environment.

This approach worked better than the previous white room. However, there was an aesthetic clash between the raymarched foreground material and the static surrounding skybox. The use of stock environmental textures seemed to weaken the aesthetic cohesion of the material. There was a concern that this, in turn, would negatively impact on the cohesion of the piece as a whole.

Following this reflection, it was decided to completely dispense with the textured environment and create a simple raymarched plane using a signed distance field and a shaping function (see Fig. 8.3). The video *Raymarched Environment and Mandelbulb Example*[108] (*mediaFiles/obj_3/raymarchedEnvEx* in the accompanying media pack) demonstrates movement

[107] https://youtu.be/07Lw3-nXQ7s (accessed 02/09/2020).
[108] https://youtu.be/zC8PqXtsWsg (accessed 02/09/2020).

throughout the scene. The plane was created in such a way that a series of ridges stretches into the distance in every direction. This was intended to give the AV-participant a sense of distance without having to use a textured mountain range. This approach also seemed more appropriate for an abstract piece in that the lack of photorealistic elements within the environment lent it a more surreal character. As all aspects of the virtual environment were now being rendered in the same fragment shader using signed distance fields, a simple fog effect was easily implemented that created a sense of distance and also some pleasing natural lighting effects. For a more detailed description of the code used to create the plane and lighting see Appendix A.1.ii.

Figure 8.3 Raymarched environment.

8.1.2 Foreground Material

The environmental visual material was developed in tandem with the foreground visual material. The foreground material marks a departure from the simple shapes of the *Ventriloquy* pieces. The motivation for this was to utilise the medium as an unreality simulator. The exploration of 3D fractal forms demonstrated exciting possibilities for appropriate source material.

Menger Sponge

The simple construction, and resulting level of visual complexity, motivated the initial experiments with the Menger sponge fractal (Quilez 2011). The fractal forums[109],which are a comprehensive

[109] https://fractalforums.org/ (accessed 16/05/2020)

source of information regarding rendering and experimenting with fractal forms, proved to be an invaluable resource during this research. The algorithm for constructing the Menger sponge is as follows:

1. Start with a solid cube.
2. Divide each face of the cube into nine squares.
3. Remove the middle cube from each face as well as the cube in the centre.
4. Repeat steps 2 and 3 as many times as desired to create the fractal.

Fig. 8.4 shows an early example of the visual results obtained with this approach. Here the fractal is raymarched with an appearance of blue tinted glass.

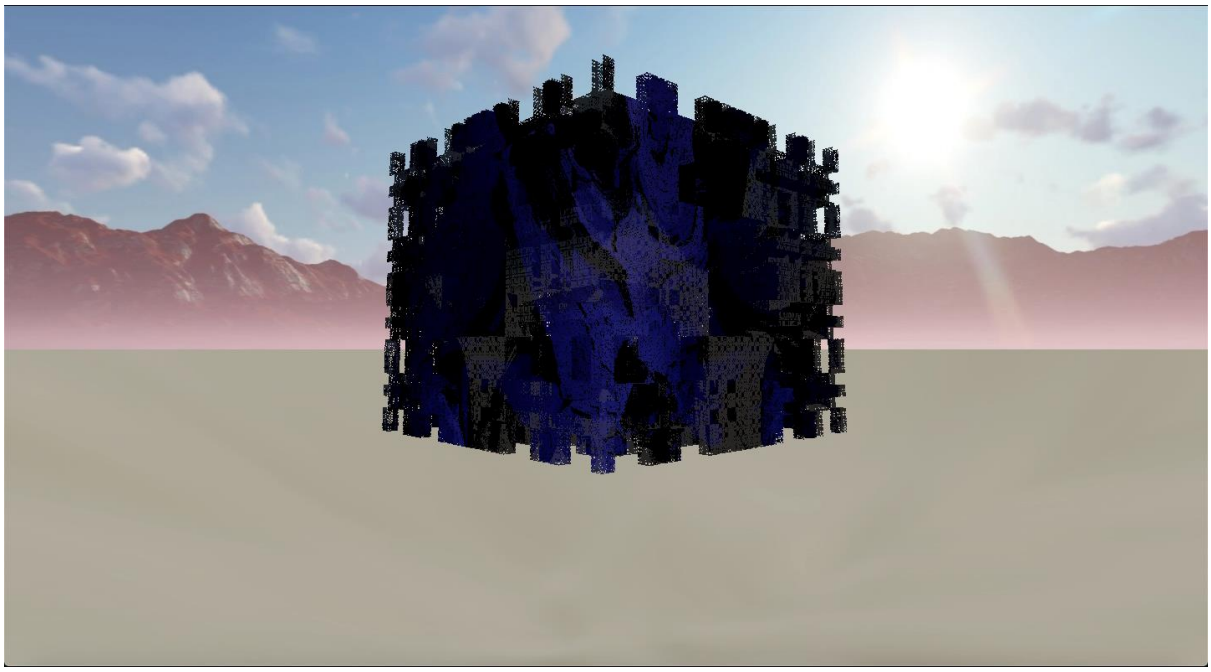

Figure 8.4 Glass menger cube.

During development, several aspects of the structure were explored. These included varying the size of the structure, level of detail and the material with which the structure was made. Although this approach provided some interesting results (see Fig. 8.5), it was still quite similar to the cubes used in the *Ventriloquy* pieces.. For this reason, it was decided to explore the Mandelbulb fractal.

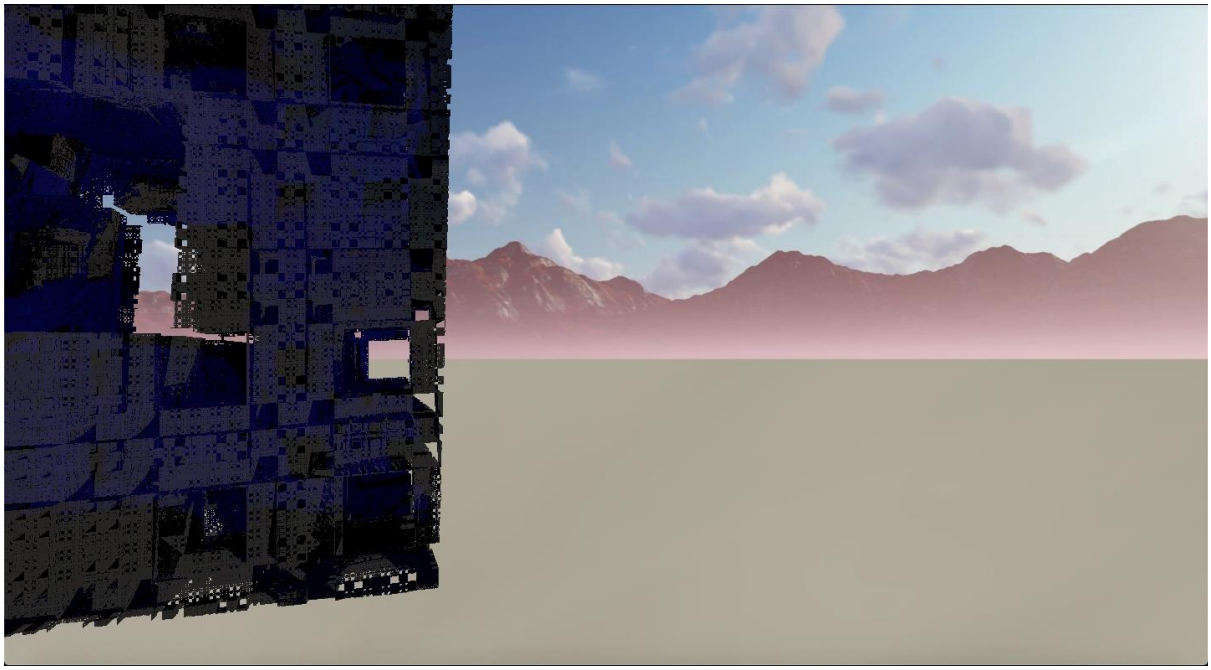

Figure 8.5 Glass menger detail.

Mandelbulb

The Mandelbulb is a 3D fractal that evolved from the 2D mandelbrot set. It was developed by Daniel White in collaboration with Paul Nylander. The evolution and derivation of the Mandelbulb formula is described in White (2009). As with the Menger sponge above, the fractal was rendered using a raymarching approach. This allowed me to run the application in real-time. The use of raymarching to render the object resulted in aesthetically pleasing fluid motion. See *mediaFiles/obj_3/glassMandelbulbEx.mov* in the accompanying media pack. The glass-like transparent material gave very pleasing results but was computationally intensive. This meant that it wouldn't run in real-time in VR. In the end an opaque, shiny material was rendered, which was efficient enough to run in real-time (see Figs. 8.6 and 8.7). The code for the signed distance field is adapted from *Glass Mandelbulb* (loicvdb 2019) and can be found at Appendix A.2.

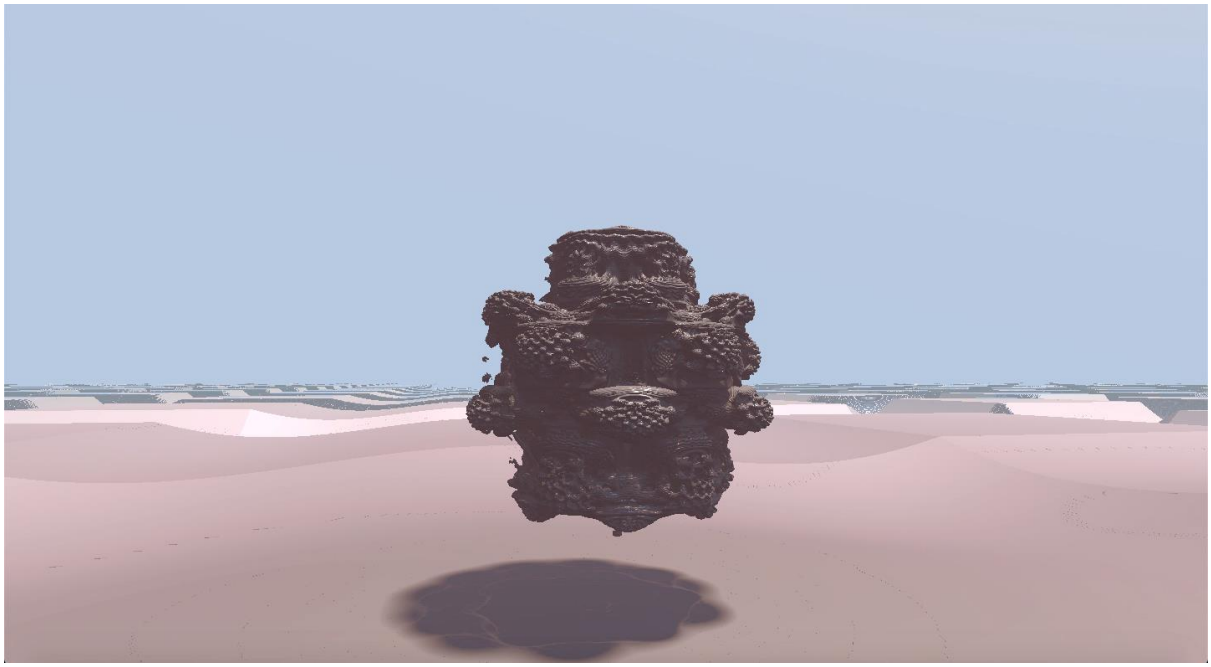

Figure 8.6 Opaque mandelbulb.

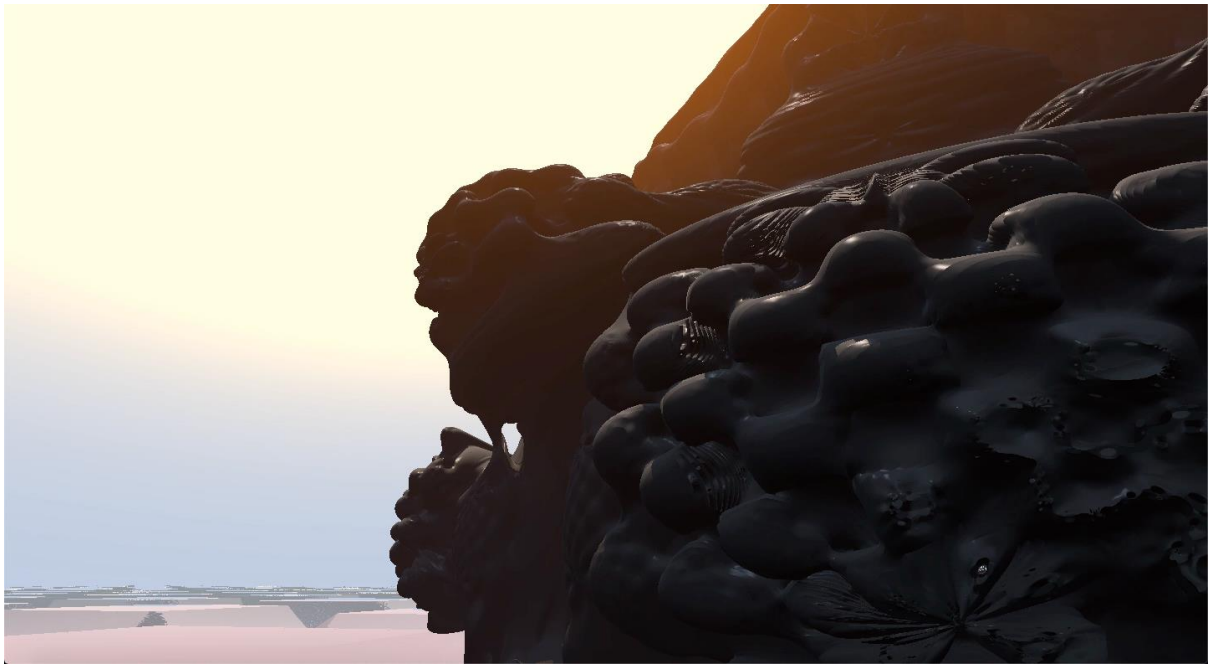

Figure 8.7 Mandelbulb surface.

8.2 Audio Elements

The sound is synthesised in a single Csound csd file. There are two separate sound processes in operation. The foreground sound that is mapped to the foreground visual object, and the environmental sound that is implicitly mapped to the surrounding environment.

8.2.1 EnvironmentalAudio

In order to create a virtual environment that fostered a strong sense of PI and Psi, sound would play an important role in the environmental setting. The environment is not only a visual artefact. It is tightly intertwined with sound also. An instrument in the Csound orchestra was created to provide some environmental audio textures. The instrument emits modulated pink noise in an effort to mimic the sound of wind across the plane. Pink noise is less harsh than white noise, containing an equal distribution of power per octave rather than per frequency band, so it was decided that this would be an appropriate method for sonifying the environment. The noise is modulated using FFT analysis and a control signal sent from *Studio::Update()*. This was done to introduce some variation in the texture. Care was taken for the texture to be as unobtrusive as possible. The instrument code, and a short description of it can be found at Appendix A.3.i.

8.2.2 Foreground Audio

The foreground sound design gradually evolved alongside the development of the visual object. During the experimentation with glass-like materials, it seemed logical, from a perceptual binding standpoint, to also generate glass-like sounds. Physical modelling synthesis techniques were explored, resulting in bowed glass audio textures. See *mediaFiles/obj_3/glassMandelBulbEx.mov* in the accompanying media file.

However, it soon became apparent that the glass material struggled to render smoothly in real-time. Consequently, a re-appraisal of the audio design approach was necessary to maintain aesthetic coherence. Building on the experience of developing *Ventriloquy I* and *II*, there was an awareness that the audio patch needed to display a wide expressive range. A granular synthesis approach was pursued, as there is large potential for timbral variation using granular techniques. Some characteristics of granular synthesis also aligned conceptually with the characteristics of fractals. Fractals are generated by relatively simple functions but are capable of wide variation and deep complexity. The granular process can be viewed in a similar light. Great timbral depth and complexity can be generated from a single signal. The granularity of the audio signal could also be thought of as self-similar, just like the structural details of a fractal. An instrument was created in Csound that used the *grain3* opcode. This opcode asynchronously granulates a synthesised sawtooth waveform. The instrument code and a description of it can be found in Appendix A.3.ii.

8.3 Mapping and Interaction

There are three mapping layers being implemented in this piece. The first mapping strategy places the foreground audio source within the scene using HRTF filters and distance calculations to situate the sound source at the centre of the audiovisual object. The data files containing HRTF measurements are based on the MIT database[110]. This is a functional mapping layer as opposed to compositional. The next layer is an audio reactive layer that analyses the audio signal and maps values to the shader to create visual movement on the surface of the object. This is a compositional layer that binds the audio to the object. Finally, a regression model is implemented using a neural network that maps data from the AV-participant's controller to audio and visual parameters.

8.3.1 Functional Mapping

The aim of this mapping layer is to situate an audio source within the scene in such a way that it behaves like a mono sound source in the real world. This process maps the foreground granular sound to the same position as the visual fractal object. The end result is that the sound appears to be emanating from the object and remains in place relative to the movement of the AV-participant. Using HRTF filters it is possible to situate a sound anywhere in the scene, anchoring it in place as the AV-participant moves around it. The code implementation of this mapping layer and accompanying description can be found in Appendix A.4.i.

8.3.2 Audio-Reactive Mapping

In addition to locating the sound source at the same place as the visual object, it was necessary to create a further mapping between the audio and visuals so that any time the foreground audio is heard, there is some surface movement on the visual object. This further strengthens the bond between the audio and visuals as the AV-participant can see visual movement synchronised to the audio signal. The AV-participant then sees that the object is vibrating. Lived experience of sound-producing objects tells us that they usually vibrate in some way to create sound. This is intended to act as a real element in the composition, as opposed to an unreal element.

An FFT is performed on the output of the granular instrument in Csound. Frequency amplitude data retrieved from the signal is then processed in *Studio::Update()* and routed to the fragment shader. The code for analysing the audio and routing the data values to the fragment shader is presented in Appendix A.4.ii. Two code snippets are included below to demonstrate how the values are used to affect the visual form.

[110] https://sound.media.mit.edu/resources/KEMAR.html (accessed 12/07/2020).

```
00 dr = pow(r, power - 1.0) * power * dr * (0.7 + lowFreqVal *
fftBinValScale) + 1.0;
```

Example 8.1 Mapped audio value used in mandelbulbSDF().

The code in Ex. 8.1 shows how the mapped value *lowFreqVal* is used in the *mandelbulbSDF()* function. It is placed in the calculation of the complex derivative. This value is then used in the distance estimation formula shown in Ex. 8.2.

```
00 return abs(0.5 * log(r) * r/dr);
```

Example 8.2 Mandelbulb distance estimation.

This formula is discussed informally in Christensen (2011b). Here he states that to truly understand the derivation of the formulas 'would require the attention of someone with a mathematical background'. This is an example of art and mathematics coming together and feeding into each other. However, the derivations of these formulas are beyond the scope of this thesis. By experimenting with the values that make up the formulas it is possible to get an intuitive sense of how aspects of the visual form are affected. Regarding *lowFreqVal*, it was found that mapping it to the derivative component had the effect of expanding and contracting the surface of the object. Direct use of the value displaced the surface too much, so it was necessary to scale it by some values. As shown above, *lowFreqVal* is multiplied by *fftBinValScale*. This is another uniform sent from *Studio::Update()*. This value is in the range 2.0 to 100.0. The product of this multiplication is then increased by 0.7 to make sure the surface doesn't completely disappear. These exact values were arrived at through a process of experimentation rather than rigorous analytical technique. This approach is suitable for artistic practice, as the focus of the work is on the rendered results rather than the technical formulas.

This audio-reactive mapping seemed to create a very tight bond between the audio and visual material. Sound is generally associated with physical movement so aligning these to features even with abstract virtual shapes hopefully works to create a sense of Psi in the scene.

8.3.3 Neural Network Mapping and Interaction

This piece utilises a similar IML mapping methodology to the previous *Ventriloquy* works. As *rapidLib* is integrated into the *ImmersAV* toolkit it was simple to use the neural network regression algorithm to quickly map data from the controllers to audio and visual parameters. This mapping layer creates indirect, metaphorical correspondences between the audio and the visuals. Whereas the audio-reactive and functional layers aim to instil a sense of Psi, this layer is the poetic, unreal element of the audiovisual interaction. The correspondences are looser here, than in the audio-reactive layer,

allowing room for interpretation. The use of position and rotation input parameters was intended to create a sense that the AV-participant was sculpting the audiovisual object. By moving and rotating the controllers they were moulding the sound and visual form. Then by moving to another location in the space they could explore different relationships with the same movements.

The novel use of IML techniques to control audiovisual parameters within VR could open up a wide range of possibilities for real-time immersive control of abstract audiovisual material. The AV-participant is empowered through real-time interaction, allowing them to enter the carefully crafted world of the composition itself and explore non-linear audiovisual relationships through a process of play and discovery. The accessible nature of the IML approach empowers the audiovisual composer with the potential to rapidly craft complex, non-linear and dynamic mappings between any input and output data that can be harnessed. Employing this approach in VR further expands the possibilities for control of both foreground and environmental audiovisual material. The specific attributes of the medium of VR now present an opportunity for audiovisual composers to create abstract experiences on a massive scale. The use of an IML control paradigm within this context supports the rapid creation of easily scalable mapping layers to match the scale of the medium.

The IML mapping layer implemented in *Obj_#3* is an extension of the previous screen-based implementations. The most immediate difference between these contexts is that here, a much larger input parameter space is used. Instead of the three input parameters utilised in the *Ventriloquy* pieces, there are fourteen input parameters. These include the three-dimensional positions of each of the controllers and the four-dimensional rotation quaternions of each of the controllers. In addition to this, the AV-participant is free to move throughout the space, around the fractal object. In *Ventriloquy I*, the input parameter range was restricted to the three-dimensional cone in front of the laptop camera. Here, the positional range is extended to room size. This was taken into account in the training approach utilised for the public demonstration.

The process of training and running the neural network is the same as that used in the *Ventriloquy* pieces. Firstly, the parameters are randomised by the AV-participant. These are defined in *Studio::Update()*. The audio parameters are sent to Csound where they affect the output of the *Granular Instrument*. The opcode *chnget* is used to retrieve the data from each channel. The parameters are then used as arguments for *grain3* as shown in Ex. 8.3.

```
00 kCps      chnget       "grainFreq"
01 kPhs      chnget       "grainPhase"
02 kFmd      chnget       "randFreq"
03 kPmd      chnget       "randPhase"
04 kGDur     chnget       "grainDur"
05 kDens     chnget       "grainDensity"
06 kFrPow    chnget       "grainFreqVariationDistrib"
07 kPrPow    chnget       "grainPhaseVariationDistrib"
08
09 kGDur = 0.01 + kGDur
10 kDens = 1 + kDens
11
12 aOut8     grain3       kCps, kPhs, kFmd, kGDur, kDens, iMaxOvr, kFn,
giWFn ,kFrPow, kPrPow
```

Example 8.3 Csound grain3 opcode.

At lines 09 and 10 above, some constant values are added to avoid audio discontinuities at startup. The visual parameters are sent as uniforms to the fragment shader. These are shown in Ex. 8.4.

```
00 uniform float randSize;
01 uniform float fftBinValScale;
02 uniform float phiScale;
03 uniform float thetaScale;
```

Example 8.4 Values sent to fragment shader.

The uniform *randSize* is used to scale the size of the mandelbulb by passing it to the signed distance function and multiplying the result by the same value again. See Ex. 8.5.

```
00 float scale = randSize;
01 mandelDist = mandelbulbSDF((pos + vec3(0.0, -1.7, 0.0)) / scale) *
scale;
```

Example 8.5 Scaling of the Mandelbulb.

By dividing the position vector and then multiplying the result the object is scaled proportionally. The uniform *fftBinValScale* is used to scale the *lowFreqVal* uniform that was described in the last section. It has the effect of accentuating the surface movement on the fractal. The last two uniforms *phiScale* and *thetaScale* are also applied to the generation of the mandelbulb. See Ex. 8.6.

```
00 theta = acos(z.y / r) * thetaScale;
01 phi = atan(z.z, z.x) * phiScale;
```

Example 8.6 Angular adjustment of the Mandelbulb.

By adjusting the angles of *theta* and *phi* it is possible to generate interesting shapes that do not necessarily look like a Mandelbulb. These shapes move away from the fractal-type detail and appear smoother. This added some aesthetic variation to the visual form as it morphed between the Mandelbulb and non-Mandelbulb shapes. See Fig. 8.8 for a non-Mandelbulb shape.

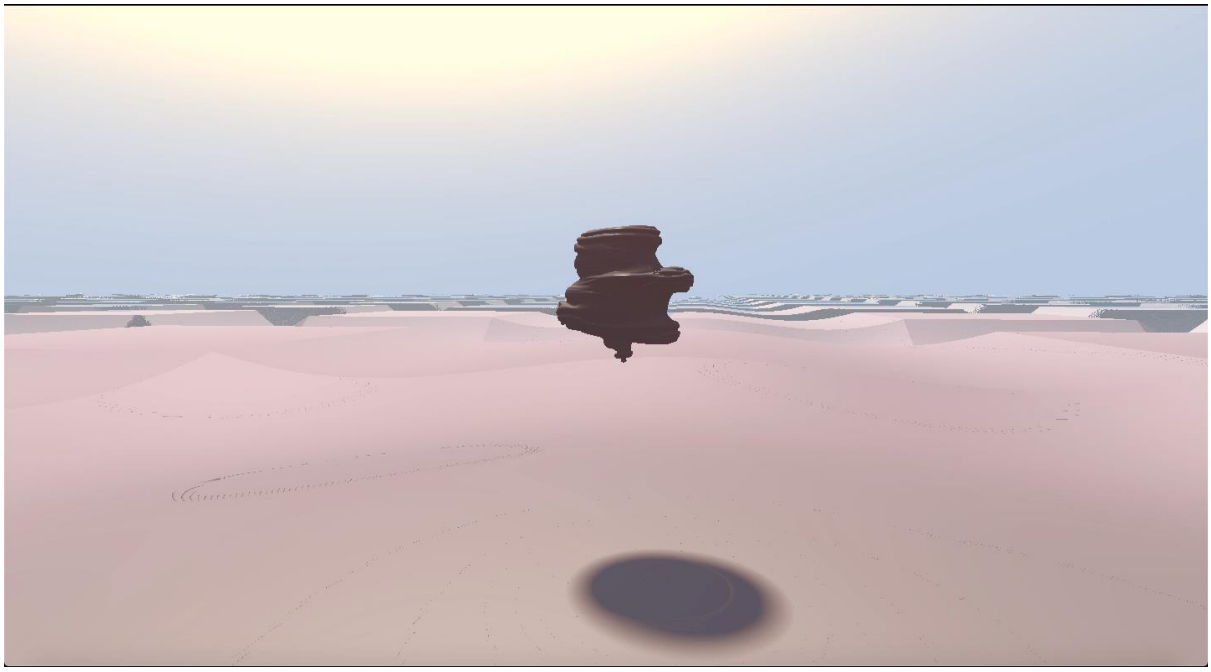

Figure 8.8 Non-Mandelbulb form.

Once the above parameters are randomised the AV-participant can decide if they like the audiovisual combination by recording a training example. As described in Chapter 5, these training examples are used as the training data for the neural network. Before recording examples, a placement plan, similar to the placement of training examples in the preparation for *Ventriloquy I*, was created. It was thought that it may be useful to spread the examples around the perimeter of the tracking space in a circle. See Fig. 64 below.

Figure 8.9 Placement of training examples for Obj_#3.

When the spatial placement of the training examples was finalised, the AV-participant then stood at the relevant position. This determined the six x, y and z values that are used as inputs to the neural network. When preparing for the public presentation, the controller was held upright just in front of the body as a baseline position. This had the effect of training the neural network, using the positions outlined in Fig. 8.9.

Once a sufficient amount of training examples are recorded, the AV-participant then trains the neural network. When the training has finished and the model is running, the AV-participant is able to control the form and the sound of the audiovisual sculpture simultaneously by moving and rotating the controllers. When playing with the system, an engaging feature consisted of turning off the model when an interesting form was found. By pressing the relevant button on the controller the AV-participant is able to pause the shape. When this happens, the audio still affects the surface movement ensuring there is a perceptual bond between the audio and visual material.

8.4 Public Presentation

Obj_#3 was demonstrated at a VR showcase in Goldsmiths, University of London.[111] This video can also be found in the accompanying media pack at *mediaFiles/obj_3/obj3_goldVRlabs_210220.mp4.* It was presented to a range of users in an informal setting. This event provided an opportunity to receive feedback on how the AV-participants perceived the experience.

[111] https://youtu.be/RdvezMCTt-I (accessed 17/09/2020).

8.4.1 Setup and Controller Configuration

The piece was presented using a HTC Vive Pro headset and two controllers. The system allowed for room size tracking so that the AV-participant could walk completely around the audiovisual sculpture. Due to the nature of the presentation context, the AV-participant did not have long to spend in the experience. For this reason a decision was made to pre-train the model, so that the AV-participant could begin playing with the system immediately. When each AV-participant entered into the virtual environment they were instructed to load and run the model. They were then free to interact with the audiovisual sculpture. The controls to start and stop the model were mapped to the grip button on the right controller. This button is hidden from view so most of the participants didn't know where it was. In future presentations it would be better to map that functionality to a more obvious place. Other than that, no one had any issues with the controllers.

8.4.2 Feedback

A system of giving feedback was implemented by Diana Lengua on the evening of the event. She utilised a whiteboard and post-it notes to give the AV-participants the chance to offer their thoughts on their experiences. See Fig. 8.10. The responses were separated into three categories relating to the AV-participant's subjective sensation of:

- Body
- Image
- Space

Figure 8.10 Participant feedback.

There were three different VR pieces being presented. The notes were divided into three different colours, one for each piece. The green notes referred to *Obj_#3*. Under the body category the responses were:

1. Meditative, relaxing, immersive.
2. I didn't feel I had a body. I was focused on the alien creature.
3. Tai-chi confrontation with another.
4. Controllers meant you had a good sense of your arms.
5. Body turns into sound.
6. Interaction.

These comments reveal some interesting insights from the AV-participants. Regarding the sensation of a virtual body, it is not surprising that one person stated they didn't feel they had a body, as there was no implementation of a virtual body. However, the mere presence of the controller models gave another person a sense of their arms existing in the virtual world. Four of the comments (1, 2, 3 and 5) seem to suggest that the AV-participant experienced a sense of presence and were engaged with the virtual environment. Comment number 5 suggests that the audio was a prominent part of the experience. This is encouraging as it suggests that at least one AV-participant was sensitive to the interaction between the audio and the environment. The next category was *image*, with the following comments:

1. Alien Planet. The shadow and light here make it feel like I am a camera.

2. On a windswept plain.
3. Impressive. Hypnotic. Fluid. Control. Mysterious.
4. I saw a face and then all faces then were not related to the shape.
5. Alien piece of coral. Why only one? I want more.
6. Spinning.
7. Giger-like a/v object and experience.
8. Lovecraft monstrous desolation with throbbing hovering screeches. Immersive interactive sound. Frightening.
9. Texture. Interactivity.
10. Liquid Sound. Total Immersion.

Looking at the above comments it's apparent that there are six that relate to visuals (1, 2, 4, 5, 7, and 8). Comment number 4 was interesting in that there is no representational imagery in the audiovisual object whatsoever. Comments 1, 5, 7 and 8 refer to an impression of alien or sci-fi material. This suggests that although an attempt was made to create an abstract environment free from representation, people may place their own representation on the material anyway. Sound is mentioned in three comments (7, 8 and 10). Comments 8 and 10 seem to equate sound with immersion. In comment 8 the sound is described as frightening. This is interesting in contrast to comment number 1 under the *body* heading which used the words relaxing and meditative. It seems these users had a completely opposing experience. Other characteristics mentioned by several AV-participants are related to movement (3 and 6) and interaction (3, 8 and 9). The final category under which AV-participants were asked about their experience was *space*:

1. Mysterious. Religious. Mystical. Sci-fi.
2. Object (outside-in).
3. __ tool!
4. Seaside.
5. Alien, highly in___ engaging.

The first comment again highlights an association with the world of sci-fi. This is the second time the word mysterious is used (see comment 3 under *image*). Interestingly this person seems to suggest a religious aspect to the environment. Unfortunately comments 3 and 5 were incomplete. However, comment 5 mentions the word alien again. Comments 2 and 4 are visually oriented statements. Similar to the alien comments under the *image* heading above, comment 4 here illustrates the

tendency for our perception to resolve ambiguities by using ‘a knowledge base of previously acquired information’ (Ernst and Bülthoff 2004: 162).

Some of the feedback received was by way of discussion throughout the session. Several of the AV-participants suggested that they would like some more fine-tuned control of the sculpture. They were happy with the large gestural interaction but when they found an interesting form they reported that they would like to be able to engage in fine detailed interaction. Another AV-participant was interested in zooming in on the surface of the fractal, similar to how it is possible to zoom in to two-dimensional Mandelbrot fractals. As the AV-participant moves closer to the surface they could use the fine control to perform a fractal zoom to unlock the infinite possibilities inherent in the structure.

8.5 Analysis

The piece primarily consists of material that can be separated into four overlapping categories; environmental, foreground, real and unreal material. The environmental material consists of all the audio and visual elements that make up the surrounding environment within which the AV-participant is situated. The foreground material is made up of all the audio and visual elements that are intended to capture the AV-participant’s attention, namely, the fractal object and the granular audio textures. The real, or representational, material consists of elements that are intended to somewhat resemble elements from the lived-experience of the AV-participant. The unreal, or abstract, material consists of elements that could not exist in the real world. The elements that make up these four broad categories serve different compositional functions. These will now be discussed.

A significant goal of this piece was to instil a sense of presence, as is understood by Slater, in the AV-participant. Slater’s understanding of presence was discussed in Chapter 4. Some people will identify this concept as immersion. The compositional aim of the arrangement of environmental elements was to foster this sense of presence. The ground plane acts as a virtual representation of the floor. The physical sensation of standing on a floor in the real world is mirrored in the virtual environment through the virtual plane. The ridges on the virtual plane aim to create an illusion of distance. The implementation of pink noise is intended to mimic, to some degree, the impression of wind blowing across the plane. These elements are intentionally representative of the real world, and as such, aim to provide a counter-balance to the more abstract elements of the piece. The surreal presence of the giant sun in the sky is intended to instil a sense of the unreal in the AV-participant.

Similarly, the hue of the plane exists in a more surreal, or unreal state. The foreground material, the fractal object and granular sound, also contributes to the unreal aspect of the piece.

The balance between real and unreal is important, and was carefully considered to encourage the AV-participant to perceive the foreground material in as present-a-manner as possible. Once an attempt to foster a sense of presence was made, through implementation of the real, environmental elements, the unreal elements were intended to encourage a sense of wonder by providing the AV-participant with an experience that would be impossible in their lived experience of the world. If the balance between real and unreal elements was off, the AV-participant may not have experienced the sensation of presence, or may have experienced presence but no sense of engagement or wonder. The perceptual effects of this balance may be similar to how the audio-spectator might perceive certain audiovisual mapping strategies. Overuse of transparent one-to-one mappings may become uninteresting, whereas mappings that are too ambiguous may fail to create any significant perceptual binding. Similarly, if the virtual environment is too close to everyday reality it could become uninteresting, acting like a one-to-one mapping of the real world to the virtual world. However, if it is too abstract, it may act like an opaque mapping, not providing enough of a connection to lived-experience to allow the AV-participant to experience a sense of presence.

Across the three feedback categories discussed in section 8.5, the concept of immersion was mentioned eight times, which indicates that these AV-participants experienced a certain level of presence. The material seems to have also provided a sense of an alien world with nine instances of terms relating to sci-fi, aliens and a mysterious or even religious sensation. These responses indicate that a sense of the unreal was also experienced by several of the AV-participants. Once the environmental material is balanced along the real/unreal axis, the foreground material might then be presented with the goal of exploring other axes of audiovisual balance such as those discussed in Chapter 3.

The foreground material represents the core focus of interaction within the piece. Here, the AV-participant interacts with material by rotating and moving the controller through the space. In this way, they are exploring audiovisual relationships between the fractal object and the granular audio texture. There are two audiovisual mapping layers within the foreground material. An audio-reactive layer and a neural network layer. Whilst overall, the foreground material exists in the unreal category, the individual layers can also be analysed through the lens of the real/unreal duality.

The structural form of the object is continually changing and amorphous. The audio texture is loosely pitched and is emitted in irregular bursts of sound. These aspects of the material are intentionally structurally ambiguous. It was posited in Chapter 3 that by generating structurally ambiguous material that may be perceived as incoherent in isolation, may help to achieve a certain level of unification of the audio and visual elements.

The audio-reactive layer is intended to anchor the audiovisual relationship between the fractal object and granular audio texture in the real world. Lived experience often provides examples of objects that visually vibrate when they emit sound. A speaker cone moves in and out to create differences in air pressure which correspond to sound waves. A string visually vibrates on a guitar causing the sound board to also vibrate and create sound. As described in section 8.3.2, audio frequency amplitudes are mapped to the signed distance function that describes the volume of the visual shape. The result of this mapping is a larger surface displacement for sounds that have more power in the lower frequencies. When this happens, the relative temporal motion is tightly synchronised and would appear perceptually balanced. However, the overall agency for this motion resides solely in the audio material. Further, the mapping is quite transparent and direct.

To counterbalance this behaviour, the neural network acts as an opaque, implicit mapping layer that exists within the unreal category. Here, the rotation and position of each of the controllers are mapped to parameters that affect the pitch and timbre of the audio texture, and the size and form of the visual shape. As the pitch and timbre change, so does the form and size of the visual object. As the AV-participant rotates the controllers, the visual form also rotates. This is almost a direct, transparent mapping, although the nature of the neural network means that it is still non-linear. A result of this is that the rotation and size of the visual object is implicitly relative to the pitch and timbre of the audio texture. Pitch-based movement is reflected in the rotating and morphing form of the visual object. This relative temporal motion is generally tightly synchronised but quite opaque, acting in a non-linear way as the AV-participant moves around the object.

Compared to the studies in Chapter 5 and the *Ventriloquy* pieces, the nature of the audiovisual relationships explored using the neural network mapping layer is different. In the previous pieces, the neural network mapped positional coordinates to audio and visual material that were each generated with noise-based algorithms responsible for the movement of the individual elements. Both the audio and visual patches contained their own kinetic agency that resulted in motion that approached noise. Here, the movement of the visual object is tied to the audio-reactive layer, and also the rotation and position of the controllers. It does not independently move. The audio-reactive layer provides the

perception of very tightly synchronised relative temporal motion between the media. The neural network layer counter-balances this by allowing for independent visual motion that is implicitly relative to pitch and textural motion. This implicit mapping is non-linear, just like the mapping between the position and rotation of the controllers, and the audiovisual parameters. This produces a more ambiguous sense of relative temporal motion than is apparent with the audio-reactive mapping and attempts to offset the imbalance in kinetic agency between the audio and visual material.

The choice of initial audiovisual objects, acting as training examples, was not conceptualised here in the same way as in the *Ventriloquy* pieces. Previously, the main aim of carefully choosing the initial objects was to create specific cadential points within the parametric space that could be used as areas of release as opposed to the intermediate forms that presented areas of tension. Perhaps the decision to move away from this conceptualisation is a result of the medium. The *Ventriloquy* pieces were temporal compositions with a beginning and an end. A strong temporal structure was needed to provide a performative arc. During the development of *Obj_#3* this was less of a concern. The AV-participant is engaged in a free exploration of the parameter space.

The original intention was to allow the AV-participant to choose their own examples and train their own models. During the course of the presentation it became clear that this was not practical due to time constraints. For this reason, in preparation for the public presentation, it was decided to choose initial audiovisual objects that attempted to maximise the relative expressive range of the material. The intention was that the AV-participant would then have the full range of the material available to explore. This was attempted by using initial training examples that were spread across the range of the material. As the AV-participant moves through the space there is a wide range of visual forms coupled with audio textures that create new correspondences depending on their position. On reflection, it may be appropriate to present the material pre-trained in an attempt to maximise its expressive potential. It takes time to work with the audio and visual material to achieve this. It may also be helpful to provide guidance for the AV-participant by visually marking positions in the environment that represent locations where the audiovisual parameters provide strong cross-modal bindings, similar to the cadential locations on the performance interfaces of the studies and the *Ventriloquy* pieces.

It was posited in Chapter 3 that the aesthetic richness of the individual audio and visual elements contribute to the perceived relative expressive range of the material. The visual object here is associated with a single audio texture. However, only four comments in the feedback mention the sonic aspect of the environment specifically. It is not clear why so few people mentioned the audio

elements. It is possible that the relative expressive range is unbalanced at the level of the individual elements. The granular audio texture may not be quite as rich in aesthetic detail as the visual material. During informal discussion, one AV-participant noted that the granulation was static and that the same grains were repeated throughout the experience. This perception of a static texture may contrast with the wider range of visual textures provided by the fractal object. Further, the level of detail of the visual object oscillated between smooth and detailed fractal surfaces. There was an attempt made to map the grain density with the level of detail of the fractal object. This was embedded in the neural network mapping layer and represented by the *grainDensity* and *fftBinValScale* parameters. However, on reflection, the mapping was not perceptually transparent enough. Perhaps such a direct mapping needs to be explicit and linear rather than implicit. Further, perhaps some more randomisation in the grain generation would have provided a more diverse audio texture to balance the visual textures.

8.6 Conclusion

Obj_#3 is the first full realisation of the core research goals. The use of a neural network to interactively shape and mould the audiovisual sculpture provides a novel form of interaction within the virtual environment. Without the use of the regression model, it would have been many times more difficult to implement the implicit, non-linear mapping that was achieved here. Further, the capability to re-map the parameters at the whim of the AV-participant would be impossible without the IML techniques that the piece was built around. This capability provides the potential for the piece to be experienced in a multitude of ways, and perhaps would encourage the AV-participant to revisit and learn how to work with the material. The evolution of the Ventriloquy pieces demonstrated the need to spend time exploring the expressive possibilities of the generative patches in order to find the areas that allow for dramatic performance. When demonstrating a VR piece such as this, the time constraints are usually too tight to allow for this familiarisation process. Perhaps in future this could be counteracted by building elements into the world that guide the AV-participant through the material.

Great care was taken here to develop an environment that fostered a sense of presence and focused the AV-participant's attention on the foreground material. Place and plausibility-illusion influenced the development of the material. As did the balance between real and unreal elements. These concepts could be considered specific characteristics of the medium of VR. The practice of audiovisual composition, with a focus on the perception of audiovisual balance, can expand into this emerging medium, harnessing these characteristics to achieve the expressive and aesthetic potential afforded by full sensory immersion.

Ag Fás Ar Ais Arís, presented in Chapter 9, will build on the concepts explored here. The line between foreground and environmental material will be blurred, further exploring the immersive possibilities of the medium. The IML control paradigm will also be applied in a novel way to exploit the scalability of the technology.

Chapter 9 Dissolving the Object: *Ag Fás Ar Ais Arís*

The primary goal of *Ag Fás Ar Ais Arís* is to blur the lines between foreground and environmental elements. Through doing this it is hoped to maximise the expressive possibilities of the entire virtual environment, where both the foreground, and elements of the surrounding environment, are capable of being controlled by the AV-participant. The use of an interactive machine learning (IML) control paradigm is central to achieving this goal.

The title is in Irish and can be translated as *Growing Back Again.* This describes the cyclical evolution of the central structure. The piece is an evolution of the approach to immersive audiovisual composition that was explored in *Obj_#3*. Whilst that piece was centred on a specific object, an audiovisual sculpture that the AV-participant interacted with, *Ag Fás Ar Ais Arís* is concerned with including the environment in the composition. It encourages exploration whilst being surrounded by continuously evolving audio and visual material. The idea of the audiovisual object is still present, however the firm boundaries between the environment and the object have now become blurred.

The approach to creating this piece was also influenced by Iannis Xenakis' *Polytopes* compositions such as *Polytope de Cluny* (1974). These were a series of large scale multimedia works created between 1967 and 1987 (Harley 1998). It was posited that VR would be capable of recreating a sense of grand scale without relying on the massive resources that went into creating Xenakis' works. The idea that VR could exploit morphing and shifting architecture on a large scale was also influential, thereby expanding on Xenakis' ideas and developing his spatial concept of art and architecture. It has been argued that Xenakis' work with multimedia and space can be seen as belonging to the tradition of Wagner's *Gesamtkunstwerk,* and that contemporary VR can be seen as a vehicle for continuing that legacy.

> Through the masterly use of the latest technological tools in the Polytopes, architecture becomes an art of time and music an art of space. In this way, these spatialized light and sound scenographies take part in the tradition that links Wagner's conception of the total art work (the Gesamtkunstwerk) with contemporary notions of cyberspace, in the sense that they both deal with the creation of an immersive and artificial environment. (Sterken 2001: 263)

It must be noted however that in his *Polytopes*, 'Xenakis purposefully dissociates the musical and visual discourse' (ibid. 2001: 267) as his aim was 'to play with the diversity of the senses and not to create correspondences in their expression' (ibid. 2001: 271). This is where the work differs as it is exploring concepts specific to audiovisual composition, which, by its very nature, is concerned with correspondences arising from the tight integration of audio and visual material.

The material within *Ag Fás Ar Ais Arís* is conceptualised as occupying three states:

- Foreground
- Intermediate
- Environmental

The foreground material is made up of the primary audiovisual object which includes the fractal structure and associated audio texture. The intermediate material is made up of elements, both audio and visual, that exist between the environment and the foreground. The ambiguous nature of this material will be discussed throughout the chapter. Finally, the environmental material is made up of the surrounding visual structures and audio texture.

This piece drills into the relationship between the object and environment. It asks how an immersive audiovisual composition can move beyond a rigidly defined duality into a more dynamic dialogue between the elements. The source code for this piece can be found at its GitHub repository[112] and in the accompanying media pack under *sourceCode/agFasArAisAris*.

9.1 Audio Implementation

The code used to generate the audio elements of the piece can be found in the file *agFasArAisAris.csd.* This file can be found in the media pack at *sourceCode/agFasArAisAris/data/agFasArAisAris.csd*. It can also be found at its public Github repository[113].

In order to create ambiguity between the foreground and environmental elements, the intermediate audio texture at times exists as an environmental element and at other times extends into the foreground material. This behaviour was achieved through a combination of several audio synthesis techniques that are intended to create a sufficiently wide range of aesthetically rich textures that both inhabit the wider soundscape for the virtual environment and also help to connect with the foreground material. In addition to this, another audio element was developed that remains solely within the surrounding environment and is not interactive. The aim of this is to provide a foundational texture

[112] https://github.com/bDunph/agFasArAisAris/tree/development (accessed 25/03/2022).
[113] https://github.com/bDunph/agFasArAisAris/blob/development/data/agFasArAisAris.csd (accessed 12/03/2022).

that would remain perceptually in the background when the intermediate texture extends into the foreground material.

The aesthetic aim of the foreground material is to achieve a more balanced perception of relative-expressive-range and relative-temporal-motion, than was achieved in *Obj_#3*. This was attempted through experimentation with granular synthesis techniques and the triggering of instruments in the *csd* file.

Another important element of the audio development in this piece was the placement of audio sources within the virtual environment. It was felt that some dynamic placement of audio sources, combined with the more ambiguous, structurally incoherent, foreground visual structure, would help in achieving the main goal of exploring the relationship between foreground and environmental elements. The implementation of the above material will be discussed throughout this section.

9.1.1 Environmental Audio

This audio element was designed to inhabit the soundscape of the surrounding environment. To achieve this, an autonomous audio texture was developed that is not directly controlled by the AV-participant. The functional nature of this element means that it should not command the attention of the AV-participant over the other intermediate and foreground elements. However, it should still be audible during passages where the intermediate and foreground material is more subdued. This suggested an audio texture that could both blend-in to the intermediate and foreground texture and also subtly cut through them. In order to achieve this, a noise-based texture was implemented that also exhibited pointillistic characteristics, enabling it to be heard through the other audio elements. To hear the texture play the file *chpt9_audio1.wav* in the media pack. This can be found at *mediaFiles/agFasArAisAris.*

The initial experimentation towards developing this texture can be heard in the audio file *24cellRow.wav*. This can be found in the media pack at *mediaFiles/agFasArAisAris.* This file contains a soundwave synthesised from the symmetry group of the theoretical *24-cell* polytope. The 24-cell is a shape that can only exist in four physical dimensions. It is a regular polytope, which means it is fully symmetrical. To say that a shape is symmetrical is to say that ‘there is a congruent transformation which leaves it unchanged as a whole, merely permuting its component elements’ (Coxeter 1948: 44). An example of a congruent transformation would be if you rotate a cube around its centre by 90 degrees. The shape will appear the same but the vertices, sides and edges that make up the cube will

be in different positions. This is one of a certain number of symmetrical transformations that can be applied to the cube. The transformations themselves make up the symmetry group.

> Such a congruent transformation is called a *symmetry operation*. Clearly, all the symmetry operations of a figure together form a group (provided we include the identity). This is called the *symmetry group* of the figure. (Ibid.)

The concept of 4D physical space was explored during the early development of *Obj_#3* with a focus on 4D polytopes. During the exploration of these structures, an attempt to find a way to audiovisualise the symmetry group was undertaken. The mathematical, group-theory, representation of these shapes was posited as a source from which to generate audio. It was reasoned that these symmetrical structures were a visual representation of the group description. Therefore there might be a way to represent the same description as an audio signal. A method was found that allowed for the sonification of the symmetry groups of the regular polytopes using an implementation of the *Todd-Coxeter* algorithm from Brown (2011). This software runs in the terminal and can write the Coxeter matrix for a given symmetry group to a text file. The Todd-Coxeter method is described in mathematical detail in Todd and Coxeter (1936). It is also described step by step in Brown (2011).

This Todd-Coxeter algorithm is recursive in nature, which can be interpreted as a type of abstract periodic oscillation. According to Heintz, McCurdy and Neukom (2015), if a process 'exhibits certain features such as periodic oscillation with a frequency range of 20 to 20,000Hz, it will produce sound'. Following this line of reasoning, a small application was written to convert the matrix values to digital audio values. Csound was then used to feed those values into an *f-table* which was then rendered out to a *wav* file. This application, and Csound instrument, along with the *txt* files and *wav* files for several polytopes can be viewed at the *PolytopeSound*[114] GitHub repository. This source code can also be found in the accompanying media pack at *sourceCode/polytopeSound*. Experimentation was carried out with reading the *f-table* at different speeds and it was found that reading it at a speed of 1Hz produced an aesthetically pleasing gritty texture.

The sound file *24cellRow.wav* is used within *Ag Fás Ar Ais Arís* as the raw audio material for the environmental audio texture. To further refine the sound and break the texture down into a more pointillistic form, the *partikkel* opcode is used to granulate it. This opcode was originally developed by Øyvind Brandstegg, Thom Johansen and Torgeir Strand Henriksen. It is an 'all-in-one implementation' (Brandtsegg, Saue and Johansen 2011: 39) of the granular synthesis techniques outlined in Curtis Roads' *Microsound* (2004). This opcode is implemented in the *ClickPopStatic*

[114] https://github.com/bDunph/PolytopeSound (accessed 14/09/2020).

instrument in the *agFasArAisAris.csd* file. The instrument is based on the second *partikkel* example[115] by Joachim Heintz and Øyvind Brandtsegg. Notes are triggered randomly using the *ClickPopStaticTrigger* instrument. This creates layered instances of the triggered instrument and randomises parameters, helping to create a dynamic texture that changes over time. For a more detailed description of the code used to build the instrument see Appendix B.1.i.

9.1.2 Intermediate Audio

This audio texture, at times, exists as an environmental element and at other times a foreground element. The intention here is to create a connection between the background and foreground material in order to try to break down the strict duality that existed in *Obj_#3*. In order to fulfil its objective, the texture needed to be capable of a sufficiently wide, expressive range so that it could both recede into the background, and command attention in the foreground, when needed. In order to contrast the completely noise-based environmental audio, the initial idea in developing this texture was to explore sounds with drone characteristics. In the author's experience, drone textures can have a similar aesthetic effect to noise-based textures in that they tend to allow the perception of the listener to wander, focusing on different aspects of the soundscape.

The raw audio material for this element is generated using a physical modelling approach. The *ModalSynth* instrument generates a continuous drone using the *wgbow*[116] and *mode*[117] opcodes. The *wgbow* opcode was developed by John Ffitch and is based on a physical model developed by Perry Cook. It is a waveguide emulation of a bowed string. This means the mathematical model takes into account the fact that the string is fixed at both ends and accounts for the physical implications of this (Heintz, McCurdy and Neukom 2015).

W*gbow* is used as a source of excitation for a bank of *mode* opcodes. The *mode* opcode was originally developed by François Blanc and translated to C code by Stephen Yi. It is a type of filter that models a mass-spring-damper system. This is a system that emulates the oscillations of a weight attached to a spring. After being set into motion the competing forces acting on the weight and the spring cause it to oscillate between states and eventually come to a stop (*ibid.)*. The file *chpt9_audio2.wav* gives an example of the sound produced by this instrument. This can be found in the media pack at

[115] http://www.csounds.com/manual/html/examples/partikkel-2.csd (accessed 09/07/2020).
[116] https://csound.com/docs/manual/wgbow.html (accessed 14/03/2022).
[117] https://csound.com/docs/manual/mode.html (accessed 14/03/2022).

mediaFiles/agFasArAisAris. For a more detailed description of the code used to build the instrument see Appendix B.1.ii.

The drone texture is then used as source material for a granulation process centred around the *sndwarp*[118] opcode. This process is implemented in the *ModalSampler* and *ModalSamplerTrigger* instruments. The *sndwarp* opcode reads sample values from a function table and allows for dynamic control of a pointer to those samples. The audio can then be dynamically time-stretched and pitch-shifted independently. When the application is launched, the *ModalSynth* plays for a number of seconds. This audio is written to a function table which is then read by the *ModalSampler* instrument. The *ModalSamplerTrigger* instrument triggers score file events for the *ModalSampler*. This instrument, in practice, relates to triggering of grains and ,consequently, is responsible for controlling the characteristics of each grain and the timbre of the overall texture. The following parameters are output from the neural network and allow for control of the texture by the AV-participant:

- Grain frequency
- Grain size
- Grain amplitude
- Maximum number of *ModalSampler* instances
- Sndwarp window size
- Filter cutoff
- Filter resonance

Grain frequency, size and amplitude refer to the characteristics of the individual grains that are triggered by *ModalSamplerTrigger*. Variation of the maximum number of *ModalSampler* instances affects the density of the audio texture. Window size is a parameter of *sndwarp*, and refers to the size of the window, in samples, that is used to scale the signal in time. The following section contains some audio examples of the perceptual effect of varying this parameter. Filter cutoff and resonance each refer to a *moogvcf2*[119] filter opcode that is placed on the output of *sndwarp*. Controlling these parameters, the AV-participant is able to explore a range of textures from sparse individual grains (*chpt9_audio3.wav*) to dense grain clouds (*chpt9_audio4.wav*). The instrument is also capable of producing overlapping stretched grains that can range from thin (*chpt9_audio5.wav*) to densely-

[118] https://csound.com/manual/sndwarp.html (accessed 14/03/2022).
[119] https://csound.com/docs/manual/moogvcf2.html (accessed 15/03/2022).

layered (*chpt9_audio6.wav*) soundscapes. It is hoped that this range of textures would help this intermediate audio element to transition from environmental to foreground material when needed.

9.1.3 Foreground Audio

The foreground audio element is intended to provide a balanced counterpart to the foreground visual element. In a similar way to *Obj_#3*, it was decided to continue the exploration of granular processes as a conceptual pairing with visual fractal forms. However, in an effort to further develop the noise-based approaches of the earlier *Ventriloquy* pieces, a noisier soundscape was desired than was developed for *Obj_#3*. It was hoped that this noisy soundscape would also be perceived as structurally incoherent, in tandem with the visual fractal structure. The aim here is to allow perceptual space for the foreground audio and visual structures to bind perceptually across the senses.

In order to achieve this, a recording of a heavy rainfall was made. It was reasoned that the textural characteristic of rainfall approaches noise and is naturally granular. This recording can be heard by listening to the file *Rain_1.wav* in the media pack at *mediaFiles/agFasArAisAris.* The source file is stored in a function table in Csound and read back by the *GranulatedRain* instrument. This instrument is based on the *sndwarp* opcode and is triggered by the *GranulatedRainTrigger* instrument. The output of the *GranulatedRain* instrument is also sent to a reverb instrument, *GranulatedRainReverb*. The wet and dry signals are then mixed again before output.

This audio element is directly controlled by the AV-participant. The parameters used to explore this element are:

- Grain frequency
- Grain size
- Maximum number of *GranulatedRain* instances
- Reverb feedback
- Reverb cutoff
- *Sndwarp* resample value
- *Sndwarp* window size

As with *ModalSampler* and *ModalSamplerTrigger* described above, the grain frequency and size parameters refer to the grain characteristics as they are triggered by *GranulatedRainTrigger*. The maximum number of *GranulatedRain* instances affects the density of the texture. Reverb feedback

and cutoff refer to characteristics of the reverb that is added to the output signal. Resample value refers to the *sndwarp* opcode. This parameter dictates how much the signal will be pitched up or down. As above, window size refers to the *sndwarp* opcode. The perceptual effect of varying this parameter results in a type of muddy distortion when the value is very low, contrasted with more clarity in the signal when the value is higher. The example file *chpt9_audio7.wav* demonstrates this with a window size of 80. Following this, *chpt9_audio8.wav* demonstrates a window size of 800. Finally, *chpt9_audio9* demonstrates a window size of 4800. These files can be found in the media pack at *mediaFiles/agFasArAisAris.*

Utilising these parameters, the AV-participant is able to explore a range of audio textures from deep pitched grains with wide filter sweeps (*chpt9_audio10.wav*) to rapidly fired grains in a single layer (*chpt9_audio11.wav*). The instrument is also capable of producing dense clouds of grains. These can range from deep and muddy (*chpt9_audio12.wav*) to shrill and frantic (*chpt9_audio13.wav*). These textures will be perceptually mapped to the foreground visual element, through an interactive machine learning (IML) process, with the aim of creating states of audiovisual balance. In this way, it is hoped to develop strong cross-modal bonds in the perception of the AV-participant.

9.1.4 Sound-Source Placement

The sound-source placement strategy deviates from the one pursued in *Obj_#3*. There, the foreground audio was simply placed in the centre of the audiovisual object and the environmental audio was sent directly to the stereo output. This was appropriate in the context of the piece due to the fact that there was a clear separation of object and environment. In *Ag Fás Ar Ais Arís,* the delineation between foreground and environmental material is not as absolute. . This presented an opportunity to experiment with the placement of the different audio elements. Each of the sound sources are placed in the environment using the *SoundLocaliser* instrument. A more detailed description of the code used to implement this can be found in Appendix B.1.iii.

Pierre Schaeffer's concept of '*kinematic relief',* the dissemination of sound into a space using 'mobile sound sources' (Harley 1998: 56) provided some initial inspiration for implementing spatially dynamic sound sources within the virtual environment. Both the environmental and foreground audio elements are disseminated within the virtual environment in this way. The environmental sound source travels in a clockwise direction around the outer boundary of the scene. This results in a natural reduction and increase in volume as the sound-source moves away from and towards the AV-participant respectively. It is mapped to a large sphere that creates a sense of scale within the

environment. The visual element and audio-reactive mapping will be discussed below. For an example of the sonic effects of the implementation see *chpt9_vid1.mp4*. This file can be found in the media pack at *mediaFiles/agFasArAisAris.*

The foreground audio element is situated at three locations within the environment. Firstly, it is situated at the origin. This is where the centre of the fractal object is also placed. This sound source is static. The nature of the fractal object means that it is not a uniform structure and there are often different parts of the structure spread around the space. Due to this visual de-centralisation, it was decided to place three sound sources orbiting the main area where the fractal expands into. These sound sources orbit the space anti-clockwise to provide some contrasting motion to the environmental sound source orbiting further out. It was decided to visually mark the sound sources with small spheres. There is no audio-reactive mapping here though as it was felt that this may detract from the desired effect of binding the foreground audio with the fractal structure. See *chpt9_vid2.mp4* for an example of the orbiting sound sources. As the video progresses it is possible to perceive a raise in volume of the audio texture and panning effect as the spheres pass by the camera. In this example the audio texture and the fractal structure are perceived as separate entities. However, when training the neural network responsible for mapping the controller position to the foreground material, a similar procedure was followed to the preparation for the *Ventriloquy* pieces. That is, an effort was made to find audiovisual couples that demonstrated balanced relative-temporal-motion. During periods of balanced relative-temporal-motion, the audio texture binds with the fractal and it appears that the audio and visual elements are unified. When this happens, the spheres tend to lose their perceptual binding with the foreground audio and become purely visual elements within the scene. This will be discussed further later in the chapter.

Finally, it was decided to route the intermediate audio source directly to the stereo out channel rather than locating it in a specific place in the environment. This choice was made due to the explicit mapping strategy employed with this element, which directly maps aspects of the sound texture to lighting and colour elements in the environment. The pervasive nature of these mappings meant that it seemed appropriate to implement the audio texture itself in a pervasive way. Care was taken to balance the sound sources correctly so that when the AV-participant is closer to the foreground material, the audio is not overwhelmed by the intermediate element.

9.2 Visual Implementation

All of the visual elements are generated within the fragment shader file *agFasArAisAris.frag*. The file can be found in the media pack at *sourceCode/agFasArAisAris/data*. It can also be found in its public Github[120] repository.

The visual scene is made up of several structural components; two planes, several spheres and a Sierpinski pyramid fractal. Each structure is raymarched according to the process described in Chapter 7. Patterns on the structures are generated using a line-based orbit trap. Colours and shading are generated using smoothing operations and global illumination techniques. The environmental and intermediate elements evolved through experimentation with basic spherical and planar forms. The goal of the experimentation was to generate a landscape that would provide a sense of presence whilst at the same time appear surreal. The balance between the real and unreal is important here. The use of the Sierpinski fractal was motivated by a desire to move away from the well-defined object of *Obj_#3*. A technique for folding the fractal was used to dissolve the structure and create a wide range of visual forms helping to create a more ambiguous foreground element that allowed room for interaction with the intermediate audio and visual elements. The implementation of these visual elements will be discussed below.

[120] https://github.com/bDunph/agFasArAisAris/blob/development/data/agFasArAisAris.frag (accessed 15/03/2022).

9.2.1 Environmental Structures

The environmental structures comprise the lower plane, the repeated large hollow hemispheres and the large orbiting sphere. The lower plane provides a terrain on which the AV-participant can walk. This is a functional element of the environment intended to reinforce the AV-participant's sense of place illusion (PI). It is generated using the same signed distance function (SDF) as the plane in *Obj_#3*. The repeated hemisphere structures are initially generated using Inigo Quilez's sphere SDF[121]. The SDF is altered here to allow for a hollow sphere. The surface is also displaced to create the hemisphere shape. This gives the impression of a type of domed-stage backdrop (see Fig. 9.1). For a discussion on the code used to render the sphere see Appendix B.2.i.

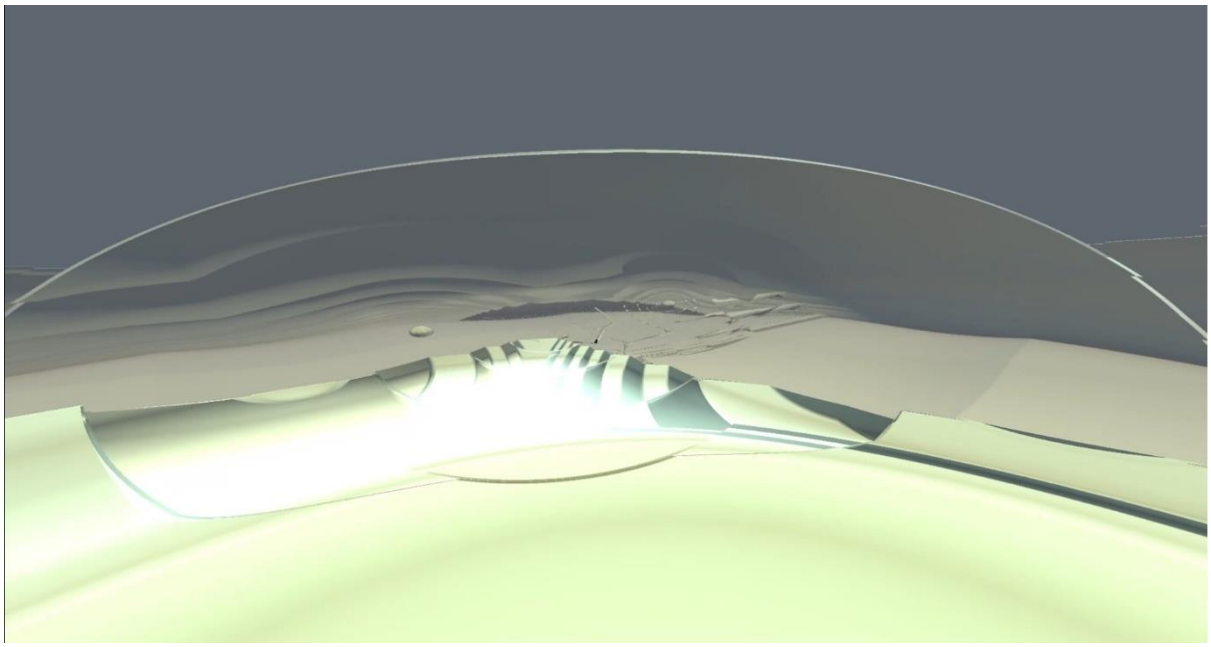

Figure 9.1 Domed-stage backdrop.

Following the initial call to the sphere SDF, it is then called repeatedly at a distance of the radius, plus a distance factor. This creates the effect seen in Fig. 9.2 where fragments of hollow hemispheres stretch to the horizon. The large orbiting sphere (see *chpt9_vid1.mp4)* is also created using the sphere SDF. For this element however, the centre of the sphere is translated to a position of (26, -5, 26) before being rotated around the *y* axis. This results in the sphere orbiting around the scene at the specified distance. The motivation to generate this element was to create a sense of scale. The object is intended as being perceived as a very large orbiting object that is occasionally visible in the background. It was hoped that this would suggest to the AV-participant that they are standing in a vast space. It was also hoped that this would add an element of unreality to the environment.

[121] https://iquilezles.org/articles/distfunctions/ (accessed 28/03/2022).

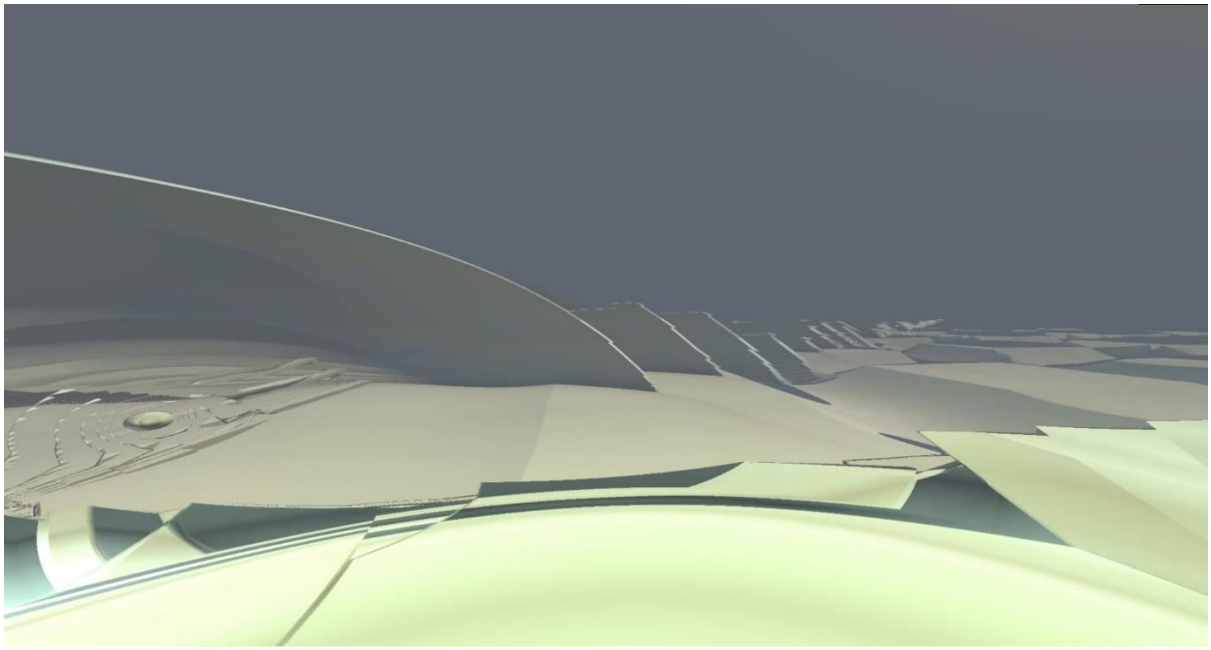

Figure 9.2 Sphere fragments stretching into the distance.

9.2.2 Intermediate Structures

The upper plane and three small rotating spheres can be conceptualised as intermediate material. In a similar fashion to the intermediate audio elements, these structures traverse the space between the surrounding environment and the foreground material. They provide both practical and expressive functions. The upper plane is intended to create variation in the terrain on a practical level, but aspects of it are also controlled by the AV-participant through the neural network mapping. This mapping will be discussed later in the chapter. The rotating spheres provide visual representations of the orbiting foreground sound-sources on a functional level, but they also reflect the changing colours and shape of the surrounding environment. In this way they become a dynamic part of the unfolding audiovisual space.

The implementation of the upper-plane arose out of experimentation with folding spatial coordinates. A similar SDF to the *Obj_#3* plane was initially implemented. This was extended to include an algorithm that tests the position of the ray, then reflects it through a 2D plane such that the point is folded across the 3D axes. This technique is discussed further with regard to the fractal structure below. For more detailed discussion on the code see Appendix B.2.ii. This led to some interesting visual artefacts which enhanced the visual palette of the composition. The video file *chpt9_vid3.mp4* displays some of these artefacts. As the camera moves around the scene, fractal-like patterns seem to float in the air. The camera also pans across the ridges that lead to the upper

plane. These effects are due to the folding algorithm and demonstrate the expressive potential of raymarching techniques. These elements are also controlled subtly by the AV-participant. A fractional Brownian motion (fBm) algorithm implemented by Morgan McGuire[122] is applied to the folded points resulting in some wave-like movement. The parameters used here are:

- fBm amplitude
- fBm speed

The fBm amplitude controls the depth of the noise displacement whilst the fBm speed controls the rate at which the noise values change. The effect of varying these parameters can be seen in *chpt9_vid4.mp4*. The ridges leading up the upper plane and the floating artefacts towards the centre of the scene vary in thickness and motion.

The original motivation for the rotating spheres was discussed above. They are generated using the same SDF as the large orbiting sphere. They are initially positioned at positions on the circumference of a circle before being rotated around the *y* axis. The nature of the raymarching approach to rendering means that the surrounding scene is reflected in the spheres. This, in turn, means that they reflect the dynamically changing environment and in that way become part of the intermediate material.

9.2.3 Foreground Structure

The foreground fractal structure was developed with an aim to generate a more structurally ambiguous form than was produced in *Obj_#3*. The intention here was to help to blur the lines between foreground and background. Also, it was an effort to explore the deeper question of audiovisual balance through the concept of isolated-structural-incoherence as discussed in Chapter 3.

The structure presented here is a kaleidoscopic-iterated-function-system (KIFS) fractal derived from the Sierpinski gasket. This is a fractal named after Waclaw Sierpinski (1882 - 1969). It is constructed by repeatedly dividing an equilateral triangle into four smaller equilateral triangles, then removing the central triangle (Riddle 2020). See the video file *chpt9_vid5.mp4* for an example of the fractal close to its initial state. An iterated function system (IFS) is a general mathematical method for constructing many different types of fractals using repeated affine transformations (Bradley 2010).

[122] https://www.shadertoy.com/view/4dS3Wd (accessed 17/03/2022).

The use of the term, *kaleidoscopic*, distinguishes KIFS fractals from traditional IFS fractals due to the folding method discussed above. This folding technique was introduced on the *Fractal Forums* website[123] by the user *knighty*.

In order to generate different structural forms, a rotation matrix is applied within the KIFS SDF. This rotation, combined with the above folding technique, and variation of several parameters generates many different forms. The AV-participant can explore the various forms by varying the following parameters:

- Fractal angle
- Number of iterations
- Scale value
- Offset value

The fractal angle parameter is the size of the angle that the point is rotated by. The number of iterations refers to the number of times the folding algorithm is performed within the SDF. The scale value affects the size of the fractal and the offset value affects the spacing between the triangles in the fractal algorithm. For examples of the range of forms capable of being generated see *chpt9_vid6.mp4*. The movement of the fractal is generated using the same fBm algorithm that was used as part of the upper-plane movement discussed above. This motion was introduced to lend the fractal some independent kinetic agency so that it could be balanced with the relative-temporal-motion of the audio. The AV-participant can also control the following parameters related to the fBm algorithm:

- fBm amplitude
- fBm speed

In a similar fashion to the upper-plane element, these parameters affect the amount of the displacement and the speed of the noise. The perceptual effect of this can be seen in the way the fractal moves. This spatial movement can be seen to range between fast motion with large displacement and slow motion with small displacement. For a discussion on the code used to generate this structure see Appendix B.2.iii.

[123] http://www.fractalforums.com/sierpinski-gasket/kaleidoscopic-(escape-time-ifs)/ (accessed 28/03/2022).

9.2.4 Colouring and Shading

The colour scheme evolved gradually through a process of experimentation with a series of smoothing calculations, colour mixtures and the use of *orbit trapping* (Quilez 1999). Orbit trapping is a technique where geometrical forms can be interwoven into the fractal structures. The geometric form used here is a simple line which creates the different coloured veins running through the fractal and surrounding environment. Any geometric forms can be used in this way and can result in the rendering of varied patterns. The smoothing calculations are adapted from a formula developed by M.H. Christensen (2011). The derivation of this formula can be found in Vepstas (1997).

Its general effect is to create smooth continuous blending of colours rather than the banded colours commonly seen in fractal renders. For a discussion of the code used to calculate the material colours and surface normals see Appendix B.2.iv. The base material colour was inspired by an emerald palette that lends a slight art-deco character to the structural elements. This is intended to instil a vague sense of retro familiarity to the AV-participant. This aspect of the environment could potentially lean towards the real end of the real/unreal scale. Some colour parameters are output from the neural network during performance which is intended to subtly change the hue of the scene. This will be discussed in the mapping section below.

After the material colour and normals are calculated, a global lighting rig based on Quilez (2013) is generated. There are three light sources modelled on the sun, the sky and an indirect source. For a discussion on the code used to generate the global illumination see Appendix B.2.v. Included in the global illumination, there are some ambient occlusion calculations being performed. Ambient occlusion is an effect that makes surfaces darker the more they are surrounded by other surfaces. The function used to calculate ambient occlusion here was taken from a thread on the *pouet.net* forums that collates useful algorithms used in raymarching shaders. This formula was posted by the user *las.*[124]

After the global illumination and ambient occlusion are calculated, fog is applied to the scene. This fog algorithm is adapted from Quilez (2010) and is essentially the same function used in *Obj_#3*. Finally, gamma correction is applied to the final *colour* value before being output from the shader.

[124] http://www.pouet.net/topic.php?which=7931&page=1&x=3&y=14 (accessed 15/07/2020).

9.3 Mapping Strategy

The primary mapping strategy used in this piece utilises neural networks to map controller positions to audio and visual parameters. There is a separate regression model implemented for each controller. This allows the right hand to control the foreground material and the left hand to control the intermediate material. This is a novel approach to controlling audiovisual material in a virtual environment. It could be seen as an extension to the parallel configuration of regression models that was discussed in Chapter 5 and implemented in *Study No.3*. However, instead of controlling both regression models with the same input, there are two separate sets of inputs here. In addition to the neural networks, there are several audio-reactive mappings implemented in this piece. These are intended to create tight synchronised motion across the audio and visual elements to balance the more opaque neural network mapping systems. Another function of the audio-reactive mappings are to reinforce the AV-participant's sense of plausibility illusion (Psi) and to act as *real* sound producing objects to balance to other *unreal* elements of the composition.

9.3.1 Foreground Mapping

The foreground material, consisting of the fractal structure and the *GranulatedRain* instruments, is controlled by the right hand of the AV-participant. The spatial coordinates of the right hand controller are used as input into a neural network. These coordinates are calculated in camera space. This means that the headset is considered the origin. Headset-relative coordinates are used instead of world coordinates to retain the same localised control area regardless of where the AV-participant is in the scene. This is intended to make the navigation of the parameter space intuitive for the AV-participant. The initial training couples will always be in the same relative spatial location during performance, thus allowing the AV-participant to retain a reliable sense of the performance space. The output parameters are as follows:

Audio:

- Grain frequency
- Grain size
- *Sndwarp* window size
- *Sndwarp* resample value
- Maximum number of *GranulatedRain* instances
- Reverb feedback
- Reverb cutoff frequency

Visual:

- fBm amplitude
- fBm speed
- Fractal angle
- Number of fractal iterations
- Fractal scale value
- Fractal offset value

Although the above parameters are not mapped directly to one-another, the simultaneous control of them through the neural network creates complex, non-linear relationships between the audio and visual material. The grain frequency and size is intended to roughly map to the fBm amplitude, fBm speed and the fractal angle. It is hoped that these associations could help to create close cross-modal bindings due to well-balanced relative-temporal-motion. The *sndwarp* window size is intended to relate to the number of fractal iterations. Lower iteration values result in less-detailed surfaces and lower window sizes result in rough-sounding, low-resolution audio textures. The maximum number of *GranulatedRain* instances, *sndwarp* resample value, reverb feedback and reverb cutoff frequency parameters are intended to create non-linear, opaque associations with the fractal scale and offset values. The audio parameters here affect the perceived density, pitch and timbre of the sounds. The visual parameters affect the size and thickness of the fractal structure. It is hoped that this may lead to cross-modal correspondences relating roughly to pitch and size.

The neural network was trained using a similar procedure to *Study No.2* and the *Ventriloquy* pieces. Initial audiovisual objects were chosen through a process of randomising the parameters and observing the perceptual results. Five objects were chosen and were associated with physical locations in space. The physical locations in space correspond to the same locations used during the training process for *Ventriloquy I*. These locations are illustrated in Fig. 39. The cartesian coordinates at these physical locations were used as input data for the training examples, whilst the audiovisual parameters representing the audiovisual objects are used as output data. The neural network was then trained using these five training examples. The training examples can be viewed in the media pack at *mediaFiles/agFasArAisAris/foreground_av*.mp4*. The files are numbered from 1 to 5. For example, the first audiovisual object video file is named *foreground_av1.mp4*. A demonstration of the trained regression model can be viewed in *foreground_demo.mp4*.

9.3.2 Intermediate Mapping

The intermediate material, consisting of the upper plane, indirect illumination, material colour and the *ModalSampler* instruments, is controlled by the left hand of the AV-participant. The input values to the neural network are the spatial coordinates of the controller relative to the headset. The output parameters of the neural network are as follows:

Audio:

- Grain frequency
- Grain size
- Grain amplitude
- *Sndwarp* window size
- Maximum number of ModalSampler instances
- *Moogvcf2* cutoff frequency
- *Moogvcf2* resonance

Visual:

- fBm amplitude
- fBm speed
- Red value
- Green value
- Blue value

In a similar way to the foreground material above, the grain frequency, size and amplitude audio parameters are intended to associate perceptually with the fBm amplitude and speed visual parameters. However, this regression model utilises different fBm parameter ranges than the foreground regression model to try to better match the motion of the visual material to the audio texture. Here, the audio grains tend to move slower and are less pointillistic than the foreground audio. To attempt to approximate this behaviour in the visual material, the fBm amplitude range is larger and the speed range is lower. This is intended to create large slower sweeping motion in the visual material to match the slower sweeps of the intermediate audio texture. The maximum number of ModalSampler instances, *sndwarp* window size, *moogvcf2* cutoff frequency and resonance parameters are intended to relate to the red, green and blue values in the visual material. These colour values are mapped to the calculations of the material colour of the visual structures and also

the indirect global illumination light source. This intention here is to loosely associate the timbre of the audio with the colour of the surrounding environment.

The neural network was trained in the same way as the foreground neural network above. The same spatial locations were used as training input data. The audiovisual training examples can be viewed at *mediaFiles/agFasArAisAris/intermediate_av*.mp4.* The files are numbered from 1 to 5. For example, the first audiovisual object video file is named *intermediate_av1.mp4.* A demonstration of the trained regression model can be viewed in *intermediate_demo.mp4.*

9.3.3 Audio-Reactive Mapping

Several transparent, one-to-one, audio-reactive mappings were implemented to create tightly-synchronised elements that would help to create strong cross-modal perceptual bindings. These elements include:

- The spectral amplitude of the *ClickPopStatic* texture is mapped to a function that displaces the surface of the large orbiting sphere. See the video file *chpt9_vid8.mp4* for an example of this mapping.
- The root-mean-squared (RMS) power of the *ModalSampler* is mapped to the brightness of the material colour and the indirect global illumination of the scene. This results in the brightness of the scene increasing with the RMS value. See the video file *chpt9_vid9.mp4* for an example of this mapping.
- The spectral amplitude of the *GranulatedRain* texture is mapped to the distance estimation calculation for the KIFS fractal. This results in the fractal expanding when the amplitude value increases. See the video file *chpt9_vid10.mp4* for an example of this mapping.

9.4 Performance

For a performance of the above material see *agFasArAisAris_performance.mp4*. The structure of the performance is as follows:

- Section A (0:00 - 3:27): Exploration of intermediate material.
- Section B (3:28 - 8:37): Exploration of foreground material.
- Section C (8:38 - 13:59): Improvisation with both regression models.

Section A is an exposition and exploration of the intermediate material. The video opens with the scene in darkness. The *ClickPopStatic* texture is heard in isolation. At 0:07 the *ModalSampler* texture enters. As the sound enters the scene lights up, demonstrating a transparent connection between these elements. The audio texture is pitched and changes randomly. The lighting in the scene reacts to the volume of each note. The relative-temporal-motion between the lighting and audio is balanced here, helping to perceptually bind the material. However, the kinetic energy generating this motion originates in the audio texture and is transferred to the luminance of the scene. This imbalance in agency will be addressed as the section unfolds. At 0:34 the intermediate regression model is activated and the left-hand controller sweeps into the centre of the performance area. During training, this spatial area was associated with *intermediate_av5*. This training example was chosen to create a subdued area within the parameter space. Here, the high-frequency material is filtered out, the audio texture is not very dense and could be characterised as a low-frequency rumble. The floating upper-plane fragments could be described as chunky, reflecting the low-frequency character of the audio. The lighting is also quite static. The aim here is for the intermediate material to retreat into the background, as opposed to occupying the foreground. This effect can be seen when the foreground material enters at 3:28. However, at this point in time, it serves as an appropriate starting point for exploring the intermediate material.

From 0:47 to 1:00 the controller reaches towards the left side of the performance area. This area is associated with *intermediate_av1*. The audio swells, with high frequencies entering the audible range. The colour of the floating fragments and upper-plane folds adopt a yellow hue in tandem with this audio swell. Here, the temporal motion of the material colour is balanced with the spectral motion of audio texture. This could be thought of as an example of well-balanced relative-temporal-motion. These elements are not directly mapped. The temporal motion within each perceptual modality is controlled simultaneously through the neural network. In addition to providing well-balanced relative-temporal-motion, the motion in the visual elements is intended to try to counter the imbalance in kinetic agency introduced by the audio-reactive mapping between the RMS of the *ModalSampler* and the luminance of the scene. The visual motion here is not a direct result of audio analysis. Instead, the temporal motion is independently instigated through the neural network.

From 1:00 to 1:28 the controller moves from the left-hand side of the performance area to the top. This area is associated with *intermediate_av3*. At 1:26 there is a long gap between notes in the audio texture. This corresponds with a dramatic reduction in luminance. At 2:06 the controller reaches the right-hand side of the performance area, which is associated with *intermediate_av2*. Here, the audio texture changes from long to short notes. As the controller moves to this part of the performance area,

the upper-plane floating fragments also become thinner. At 2:23 the rhythmic speed of the audio texture increases further. This seems to skew the perceived audiovisual balance as the rapid notes dominate the perceptual space. Here, the audio seems to be more structurally coherent due to the enhanced tonality and rhythmic characteristics. An attempt to counter this imbalance will be discussed below. Following this, at 2:36, the controller continues to the bottom part of the performance area, which corresponds to *intermediate_av4*. Here, the notes in the audio texture are sustained for a longer period. This is matched in the visual material as the plane fragments expand to become thicker. Finally, the controller moves back up towards the left-hand side of the performance area before returning to the centre. At 3:01 the controller pulls back from the centre. Here the filter parameter changes dramatically which corresponds to spatial movement in the floating upper-plane fragments. This marks the end of section A.

At 3:28 the foreground regression model is activated. The fractal object appears in a structurally coherent Sierpinski pyramid form. The vein textures running through the object are quite distinct. The colour of these textural elements are controlled by the intermediate neural network. This is an example of how the intermediate material is intended to blur the boundary between the foreground and environment. In addition to this, the veins themselves run out from the fractal onto the ground plane, creating a further connection between the foreground and the environment.

The AV-participant moves through the fractal to the opposite side of the virtual space. At 4:08 the right-hand controller sweeps across the performance area and reaches the left side. This spatial area is associated with *foreground_av1*. Here, the spectral amplitude of the audio texture is directly mapped to the surface calculation of the fractal resulting in a tightly synchronised movement. Between 5:10 and 5:28 the controller moves from the left side of the performance area to the bottom. This area is associated with *foreground_av2*. The visual form morphs from organic-looking strands to a large pyramid structure. As the scale of the visual object expands the audio texture becomes deeper. This creates a strong cross-modal binding. From 6:19 to 6:40 the controller slowly moves from the bottom of the performance area to the right-hand side. This part of the performance space corresponds to *foreground_av4*. As the object becomes more dense the audio also moves towards a dense noise texture. From 7:17 to 7:30 the controller gradually moves from the right-hand side of the performance area to the top. This area is associated with *foreground_av5*. As the controller travels between these areas the visual structure becomes more structurally complete and solid. This could potentially cause an audiovisual imbalance as the audio here is quite noise-like and structurally incoherent. The visual figure morphs into a pyramid shape with continuously moving chunks. Here, the audio texture also becomes chunky, with deep guttural sounding grains, resulting in balanced

relative-temporal-motion. From 8:15 to 8:24 the controller finally moves from the top of the performance space to the centre. The visual object condenses into a more compact form and the audio drops in pitch. The pulsating nature of the audio then speeds up. This area of the performance space is associated with *foreground_av3*. This audiovisual object is intended to align compositionally with *intermediate_av5* as a more subdued section of the performance space. Throughout this section, the intermediate material remains somewhat static, retreating into a background state, supporting the foreground material by filling out the environmental audio and providing a static colour scheme.

Section C begins at 8:38. Throughout this section the AV-participant is controlling the intermediate and foreground material simultaneously. Here, the intermediate material moves between foreground and background blurring the lines between the different areas of the virtual environment. This shift from background to foreground can be perceived from 8:40 as the left controller gradually moves from the centre of the performance area out to the left-hand side. As this happens, the intermediate audio material opens up with the entry of high frequencies, this happens in tandem with a change in colour of the veins running along the ground plane and through the fractal object. From 9:17 to 9:26 the left-hand controller moves from the bottom-left up to the top-left side of the performance area. This corresponds to a timbral change in the intermediate audio material and in the colours of the veins. Here, the intermediate material is mixing with the foreground material, corresponding to the noisy, pulsating character of the foreground audio and the fragmented structural form of the fractal.

From 9:30 to 9:38 the right-hand controller moves from the left to the bottom of the performance area. This corresponds to an expansion of the fractal form into a large, fragmented pyramid structure. The foreground audio here also seems to expand from a dense pulsating character to a lighter, more expansive noise texture. Whilst this is happening in the foreground material, the intermediate audio is characterised by a high pitched drone that slowly dissipates into a noisy state. The colour also changes from an orange tinge to yellow. By 9:40 the whole character of the scene has changed to a more expansive, non-pitched, fragmented state. The relative expressive range, between the two regression models, across this section is quite well balanced and results in a passage where each of the elements are working in a unified way.

At 10:33 the intermediate audio changes to a distinct texture characterised by randomly pitched notes of short duration. The rhythmic and tonal structural characteristics skew the audiovisual balance of the scene, creating a moment of tension. There is an attempt to balance the structural coherency of the material here as the foreground visual material morphs through several forms before coming to rest in a more solid state at 11:41. The nature of the intermediate audio throughout this passage lends

a frantic nature to the visual deformations. At 12:37 the two controllers join together at the top of the performance area and move in tandem down towards the centre. This signifies the final section. The fractal is condensed into a tightly packed area. As the frequencies of the audio drop into the lower registers, the colour scheme becomes darker and the floating fragments become larger. At 13:01 the left-hand controller moves out to the left before jumping back into the centre. It then moves up before jumping back again. It finally moves down before jumping back a final time. The AV-participant is utilising the jump resolve technique that evolved through the studies in Chapter 5. The AV-participant then moves away from the centre of the scene across the upper plane. The intermediate and foreground material fades until the performance ends with the environmental material in isolation.

9.5 Conclusion

Ag Fás Ar Ais Arís is the culmination of the research carried out in this thesis. The central aim of the piece was to explore the blurring of boundaries between the surrounding environment and foreground material. In this way, it was hoped to maximise the expressive potential of the interactive machine learning control paradigm within the emerging medium of VR. Whilst *Ventriloquy I & II,* and *Obj_#3,* were focused on specific audiovisual objects in isolation, this piece embraced the surrounding environment as an active part of the composition. As such, the primary outcomes from this piece are:

- The implementation of an intermediate layer of material to act between the foreground and the surrounding environment.
- The use of separate neural networks for each controller to allow the AV-participant to simultaneously control the foreground and intermediate material.

These outcomes are built on top of the outcomes that have emerged through the practice presented in the rest of the thesis. The initial development of the core material for this piece was grounded in the concepts of relative-expressive-range, relative-temporal-motion and isolated-structural-incoherence. Care was also taken to present a virtual environment within which a sense of presence was fostered through convincing PI and Psi. This aspect of the composition was dependent upon a balance between *real* and *unreal* elements. The technical implementation of the piece was dependent on the *ImmersAV* toolkit. The focused nature, and intimate knowledge, of the software, allowed for the freedom to experiment with the material in a way that felt natural to the author. The compositional approach of choosing initial audiovisual training examples and arranging them in the performance space was continued from the studies in Chapter 5 and *Ventriloquy I*.

The culmination of this piece aligns with the culmination of the research presented in this thesis. It does not, however, represent the end of the practice that has been developed over the course of the project. Chapter 10 will discuss the overall outcomes of the research and suggest some ways in which it can be continued.

Chapter 10 Conclusion

The core research question posed at the beginning of the thesis asked how IML can be used to control audiovisual compositions in the emerging medium of VR. Throughout the thesis, the constituent strands of inquiry that relate to the core research question were explored, before being combined into a unified statement. These strands include:

- The practice of audiovisual composition.
- The use of IML as a control paradigm for audiovisual compositions.
- The use of VR as a medium, through which, audiovisual compositions can be presented.

This chapter will look at the research outcomes and demonstrate how they combine to provide new knowledge around the research question. Following this discussion, future directions for the research will be suggested.

10.1 Categorisation of Practices

The first strand of inquiry is related to the practice of audiovisual composition. This is the context within which the exploration of IML and VR technologies took place. To provide the foundation for this contextualisation, Chapter 2 presented a survey of work within the field of Audiovisual Art. The aim of this survey was to provide a map within which the rest of the work could be situated. Some issues around terminological clarity were identified alongside some contradictory conceptualisations of aesthetic styles.

In contemporary art and music practice in general, the blurring of boundaries long ago became commonplace and it would be impossible to draw hard borders around any practice. However, it was argued that it remains useful to be able to identify the aims of artistic practices and the principles that ground them.

An argument was made for the separation of aesthetic style from presentation context. Some practices, such as *live cinema,* are, in part, defined by their presentation context. However, practices such as *visual music* can be found in several contexts, both live and fixed-format. The combination of style and context highlights different flavours of practice. With visual music, the aim of many composers is to translate musical form and movement into the visual realm. This can be separated from the presentation context and, therefore, accounts for the fact that there are live visual music performances as well as fixed-format visual music screenings. Live visual music is a different flavour of audiovisual

expression to fixed-format visual music due to the presentation context. However, in their essence, they are still visual music pieces. When talking about *generative audiovisual art*, this also applies. This separation of content and context helps to focus on aesthetic concerns, which can then be discussed within the context of the chosen presentation framework. The presentation of generative audiovisual art can be seen in many different places such as in live performance, immersive contexts, online spaces and also public installations. Just like visual music above, each of these presentation contexts contributes to a different flavour of generative audiovisual art.

This separation of aesthetic style from presentation context is an original contribution to the field of audiovisual art. The work done here helps to contextualise and clarify the terminology used in the subsequent theoretical discussion surrounding the practice of audiovisual composition, which, in turn, provides the foundation for the compositional practice within which the use of IML and VR technologies were explored.

10.2 Audiovisual Balance

The concept of equality-of-material in audiovisual composition was identified as an avenue of theoretical exploration that would help to provide compositional goals for the pieces presented later in the thesis.

The concept of *audiovisual balance* emerged as a way of conceptualising the idea of equality in audiovisual compositions. Emphasising the position that audiovisual art is an art of interaction between sensory modalities, it was argued that those interactions present themselves most clearly when the material is in a state of balance. Some forces that affect the perceived balance of the material were proposed. These forces were identified as *isolated-structural-incoherence, relative-temporal-motion* and *relative-expressive-range*.

Extending the practice of audiovisual composition into VR, fostering a sense of presence was identified as an important, medium-specific characteristic that may have an effect on the perception of audiovisual balance. The implications of this mean that, when working in VR, the audiovisual composer is tasked with arranging material according to three senses instead of two; the senses of sight, hearing and presence. It was posited that the sense of presence may be fostered through careful development of material that provides realistic representations of PI and Psi. However, it was also noted that, in the context of audiovisual compositions, the creation of implausible, abstract material, is also important for maintaining the AV-participant's interest. This suggested that a balance might

be struck between *real* and *unreal* material. This axis of balance was incorporated into the forces that affect audiovisual balance in the composition.

These new terms for audiovisual analysis contribute to the practice of audiovisual composition and act as the theoretical foundation upon which the compositions in the portfolio were built. They contribute to the core research aims by providing the analytical basis for assessment of the compositions in the portfolio.

10.3 ImmersAV Toolkit

The *ImmersAV* toolkit directly contributes to the main research goals by providing the means by which IML techniques were used to control audiovisual compositions in VR. The toolkit presents a focused workflow for immersive audiovisual composition utilising an IML control paradigm. The toolkit was designed from the ground up with a focus on creating a streamlined environment within which audiovisual composition, IML and VR could be combined and explored. A significant feature of the toolkit is the ability to map data from any part of the toolkit to another. This allows for audio-reactive mappings, visual-to-audio mapping of data, including the mapping of pixel data directly from the GPU to CPU-based audio processes, and simultaneous mapping of data from third party processes to audio and visual material. This functionality was motivated by the desire to give the audiovisual composer complete control over audiovisual balance within the composition. The ability to harness the power of modern VR systems in a way that is completely open source is also an important aspect of the toolkit. In this way it offers an alternative path to experimenting in VR than is currently offered by commercial game engines.

10.4 Artistic Output

The audiovisual compositions presented in the portfolio demonstrate the development of compositional methodologies that directly explore the ways that IML can be used to control audiovisual material.

In Chapter 5, a multilayer-perceptron, feedforward neural network was used to implement a regression algorithm that maps input control data to output parameter data. This technique is a novel method of simultaneously controlling audio and visual parameters in real time. Four studies were then created that explored some ways in which this approach could be used. From these studies, a range of observations were made that influenced the subsequent work. The outcomes that proved most important for the later compositions were as follows:

- Implementing neural networks in parallel to allow multiple audiovisual objects to be controlled simultaneously.
- The use of the jump resolve performance technique to explore tension and release through repetition.
- The use of randomisation to find audiovisual parameters that provided strong cross-modal bindings.
- The use of balanced relative-temporal-motion to bind the audio textures to the movement of the visuals.
- The method of associating initial audiovisual couples with certain spatial areas of the performance interface.
- The use of improvisation to explore the parametric space.
- The use of structurally complete visual elements or tonal audio events to purposefully skew the audiovisual balance, creating tension, before returning to cadential areas to relieve that tension.

The primary compositional methodology used to develop material emerged through these studies. This was expanded during the development of *Ventriloquy I*. Here, the input control method was expanded from a 2D paradigm to 3D in Chapter 6. *Ventriloquy I* utilised a three dimensional performance interface for the first time. The work done here in placing the initial training examples in the 3D space proved useful for the later VR pieces. The placement of the training examples in the 3D performance space allowed for intuitive navigation of the parametric interface during performance. Another compositional outcome from this piece was the discovery that placing similarly chaotic audiovisual objects within similar spatial regions of the performance space allowed for smooth entry and exit of visual objects during performance. The 3D spatial control paradigm was substituted for a tactile controller for *Ventriloquy II*. Although this provided some desirable tactile response during performance, the intuitive spatial layout of the initial training examples was lost. This reinforced the effectiveness of the spatial paradigm developed in *Ventriloquy I*.

In order to implement the above control methods within VR, the *ImmersAV* toolkit was developed. The first piece developed using the toolkit was discussed in Chapter 8. *Obj_#3* was the initial realisation of the core research aims. This piece carefully presented material with a focus on fostering a strong sense of presence. The use of real and unreal material provided a way to create a strong sense of PI and Psi in addition to creating interest through the abstract material. The IML control paradigm was implemented in such a way that positional and orientation data from both controllers were

utilised. The compositional methodology was altered so that the initial training examples were arranged in world space around the Mandelbulb. Also, the IML mapping layer was augmented by audio-reactive mappings to enhance PI. The mapping of RMS power values to the surface of the Mandelbulb provided tightly synchronised movement to bind the audio and visual material perceptually whilst the IML mapping layer provided a looser, more expressive association between the media.

The final piece, *Ag Fás Ar Ais Arís*, was built upon the knowledge gained throughout the research. It was built with the *ImmersAV* toolkit and aimed to maximise the expressive potential of the virtual environment. Here, the relationship between environmental and foreground material was examined. The use of an intermediate layer of material, that traversed the gap between environmental and foreground material, was instrumental in blurring the boundaries between the two elements. The control paradigm borrowed heavily from the earlier *Ventriloquy* work, utilising camera space coordinates as the input data for the training examples. This ensured that the performance area was the same wherever the AV-participant was standing in the virtual environment. Finally, the use of parallel neural networks was influenced by the earlier studies. However, in this implementation, separate input data was used for each neural network. This allowed the AV-participant to simultaneously control the foreground and intermediate material.

10.5 Future Work

The research presented in this thesis suggests several avenues for future exploration. The *ImmersAV* toolkit would benefit from further development. It may be possible to abstract the functionality of the *Studio()* class to an external text file that would function similarly to the *csd* file and fragment shader. That would mean the *ImmersAV* toolkit could exist as a pre-built binary with compositions being loaded at run-time. This would simplify the process of creating new pieces and would provide a compact and portable environment. The toolkit is currently built on OpenVR. It would be beneficial to include other VR SDKs, which would increase the range of systems it could be used with. It may also be beneficial to replace OpenGL with Vulkan, and possibly Metal for macOS specifically. These graphics APIs are potentially faster and more efficient than OpenGL.

The IML control paradigm also suggests many avenues for exploration going forward. *Ag Fás Ar Ais Arís* utilised two neural networks with the camera space coordinates of each controller as the input data. It would be possible to add at least two more neural networks with the orientation of each controller acting as input data to the neural networks. These extra mapping layers could be used to

control more elements of the environment or the foreground material. Another mapping layer could be implemented utilising the asynchronous read-back capability of the ImmersAV toolkit. A classification algorithm could be employed to activate different audio and visual elements depending on what the AV-participant is looking at in the virtual environment. The cyclical mapping technique described in Chapter 7 could be developed to introduce some random oscillatory behaviour between the audio and visuals. This technique could also be explored further to see what other audiovisual feedback effects could be developed.

Virtual environments are well-placed to simulate non-euclidean spaces. The exploration of such spaces could provide rich and dynamic audiovisual environments within which the AV-participant could explore. With the continual development of VR hardware, it is only a matter of time until the tactile sense of touch is commonplace in VR environments. This would further extend the possibilities for immersive experiences. The manipulation of tactile sensations as part of the composition would be an interesting avenue of exploration.

10.6 Conclusion

The outcomes described in this chapter contribute to the main research question by presenting some ways in which it could be answered. As discussed above, there are several other ways it could be answered going forward.

The development of the pieces in the portfolio shone a light on the ways in which these emerging technologies can be expressively employed in an art of multisensory perception. The act of creating these works also shone a reciprocal light on the nature of the author's perception, providing some welcome insight along the way. According to Arandas, Grierson and Carvalhais (2020: 2) the 'abstract, discrete systems we build to make art, when modelled like us, can help us understand a little bit more about the world through the art we are trying to make with it'. Ultimately, I hope that the ideas, software and artworks presented in this thesis will contribute positively to the field and provide a jumping-off point for further exploration.

Bibliography

Abbado, A. (1988) ‘Perceptual Correspondences of Abstract Animation and Synthetic Sound’, *Leonardo*, Supplemental Issue Vol 1. Electronic Art, 1, pp. 3–5.

Abbado, A. (2017) *Visual Music Masters: Abstract Explorations: History and Contemporary Research*. Milan: Skira Editore.

Agrawal, S. et al. (2020) ‘Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences’, *Journal of the Audio Engineering Society*, 68(6), pp. 404-417. doi: https://doi.org/10.17743/jaes.2020.0039.

Alexander, A. (2009) *Audiovisual Live Performance, See This Sound*. Linz: Ludwig Boltzmann Institute; Lentos Art Museum Linz; Academy of Visual Arts Leipzig. Available at: http://www.see-this-sound.at/compendium/maintext/54/6.html (Accessed: 27 September 2020).

Alves, B. (2005) ‘Digital Harmony of Sound and Light’, *Computer Music Journal*, 29(4), pp. 45–54. doi: 10.1162/014892605775179982.

Amershi, S. et al. (2014) ‘Power to the People: The Role of Humans in Interactive Machine Learning’, *AI Magazine*, 35(4), pp. 105–120. doi: 10.1609/aimag.v35i4.2513.

Arandas, L., Grierson, M. and Carvalhais, M. (2020) ‘Sonic and Visual Relationships : The Mind, Contemporary Composition and Complex Signals’, in *SIIDS 2020 - Sound, Image and Interaction Design Symposium*. Funchal: University of Madeira, pp. 1–3. Available at: https://siids.arditi.pt/wp-content/uploads/2020/08/SIIDS_2020_paper_7.pdf.

Arfib, D. et al. (2002) ‘Strategies of Mapping Between Gesture Data and Synthesis Model Parameters Using Perceptual Spaces’, *Organised Sound*. Cambridge University Press, 7(02), pp. 127–144. doi: 10.1017/S1355771802002054.

Atherton, J. and Wang, G. (2018) ‘Chunity: Integrated Audiovisual Programming in Unity’, in *New Interfaces for Musical Expression*. Blacksburg. Available at: http://www.nime.org/proceedings/2018/nime2018_paper0024.pdf (accessed 15 April 2022).

Basanta, A. (2017) ‘Shades of Synchresis: A Proposed Framework for the Classification of Audiovisual Relations in Sound-and-Light Media Installations’, *eContact!*, 19(2). Available at: https://econtact.ca/19_2/basanta_synchresis.html.

Battey B. (2015) ‘Towards A Fluid Audiovisual Counterpoint’, *Sonic Ideas*, 4(17), pp. 26-32.

Battey, B. (2016) ‘Creative Computing and the Generative Artist’, *International Journal of Creative Computing*, 1(2/3/4), pp. 154–173. doi: 10.1504/ijcrc.2016.076065.

Battey, B. (2020) ‘Technique and Audiovisual Counterpoint in the Estuaries Series’, in Knight-Hill, A. (ed.) *Sound and Image: Aesthetics and Practices*. ebook. Abingdon, New York: Routledge, pp. 263–280.

Bauer, M. (2018) ‘Self-Imposed Fetters: The Productivity of Formal and Thematic Restrictions’, *Connotations*, 27, pp. 1–18. Available at: http://www.connotations.de/debate/fetters.

Bergstrom, I. et al. (2017). ‘The Plausibility of a String Quartet Performance in Virtual Reality’, *IEEE Transactions on Visualization and Computer Graphics*, *23*(4), 1332–1339. https://doi.org/10.1109/TVCG.2017.2657138

Bergstrom, I. and Lotto, R. B. (2016) ‘Soma: Live Musical Performance in Which Congruent Visual, Auditory and Proprioceptive Stimuli Fuse to Form a Combined Aesthetic Narrative’, *Leonardo*, 49(5), pp. 397–404. doi: 10.1162/LEON_a_00918.

Bernardo, F. et al. (2017) ‘Interactive Machine Learning for End-User Innovation’, in *Proceedings of the AAAI Symposium Series: Designing the User Experience of Machine Learning Systems*. Association for the Advancement of Artificial Intelligence. Available at: http://research.gold.ac.uk/19767/1/BernardoZbyszynskiFiebrinkGrierson_UXML_2017.pdf (accessed 14 April 2022).

Betts, T. (2013) *Generative AV With Pure Data And Unity*. Available at: http://www.nullpointer.co.uk/content/generative-av/ (accessed 27 September 2020).

Birtwistle, A. B. (2006) *Cinesonica : Sounding the Audiovisuality of Film and Video*. PhD thesis. Goldsmiths, University of London. Available at: http://research.gold.ac.uk/id/eprint/28634 (accessed 14 April 2022).

Boucher, M. (2020) 'Capturing Movement: A Videomusical Approach Sourced in the Natural Environment', in Knight-Hill, A. (ed.) *Sound and Image: Aesthetics and Practices*. ebook. Abingdon, New York: Routledge, pp. 226–239.

Boucher, M. and Piché, J. (2020) 'Sound/Image Relations in Videomusic: A Typological Proposition', in Knight-Hill, A. (ed.) *Sound and Image: Aesthetics and Practices*. ebook. Abingdon, New York: Routledge, pp. 13–29.

Bradley, L. (2010) *Iterated Function Systems, Chaos & Fractals*. Available at: https://www.stsci.edu/~lbradley/seminar/ifs.html (accessed 29 September 2020).

Brandtsegg, Ø., Saue, S. and Johansen, T. (2011) 'Particle Synthesis – A Unified Model for Granular Synthesis', in Neumann, F. and Lazzarini, V. (eds) *Linux Audio Conference 2011*. Maynooth, Ireland: NUI Maynooth, pp. 39–46. Available at: http://lac.linuxaudio.org/2011/papers/39.pdf (accessed 15 April 2022).

Brett, T. (2016) 'Virtual Drumming: A History of Electronic Percussion', in Hartenberger, R. (ed.) *The Cambridge Companion to Percussion*. Illustrated. Cambridge: Cambridge University Press, pp. 82–94.

Brown, K. (2011) *The Todd-Coxeter Procedure,* lecture notes, Mathematics 6310, Cornell University, delivered April 2011. Available at: https://pi.math.cornell.edu/~kbrown/6310/toddcox.pdf (accessed 14 April 2022).

Brunel, L., Carvalho, P. F. and Goldstone, R. L. (2015) 'It Does Belong Together: Cross-Modal Correspondences Influence Cross-Modal Integration During Perceptual Learning', *Frontiers in Psychology*, 6(358), pp. 1–10. doi: 10.3389/fpsyg.2015.00358.

Buckley, Z. and Carlson, K. (2019) 'Towards a Framework for Composition Design for Music-Led Virtual Reality Experiences', *26th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019 - Proceedings*, pp. 1497–1499. doi: 10.1109/VR.2019.8797768.

Burdea, G. C. and Coiffet, P. (2003) *Virtual Reality Technology*. 2nd edn. New Jersey: John Wiley & Sons.

Callear, S. (2012) *Audiovisual Particles: Parameter Mapping as a Framework for Audiovisual Composition*. PhD Thesis. Bath Spa University.

Candy, L. (2007) 'Constraints and Creativity in the Digital Arts', *Leonardo*, 40(4), pp. 366–367. doi: 10.1162/leon.2007.40.4.366.

Carvalho, A. et al. (2015) *The AudiovisualBreakthrough*. Edited by A. Carvalho and C. Lund. Berlin: Fluctuating Images.

Cassidy, J. (2012) *They Upped Their Game After The Oranges.* Available at http://www.janecassidy.net/they-upped-their-game-after-the-oranges.html (accessed 14 April 2022).

Chion, M. (1994) *Audio-Vision: Sound on Screen*. Edited by C. Gorbman and W. Murch. New York: Columbia University Press.

Christensen, M. H. (2011) *Distance Estimated 3D Fractals (II): Lighting and Coloring*, Syntopia. Available at: http://blog.hvidtfeldts.net/index.php/2011/08/distance-estimated-3d-fractals-ii-lighting-and-coloring/ (accessed 29 September 2020).

Christensen, M. H. (2011a) *Distance Estimated 3D Fractals (III): Folding Space*, Syntopia. Available at: http://blog.hvidtfeldts.net/index.php/2011/08/distance-estimated-3d-fractals-iii-folding-space/ (accessed 29 September 2020).

Christensen, M. H. (2011b) Distance Estimated 3D Fractals (V): The Mandelbulb & Different DE Approximations, Syntopia. Available at: http://blog.hvidtfeldts.net/index.php/2011/09/distance-estimated-3d-fractals-v-the-mandelbulb-different-de-approximations/ (accessed 29 September 2020).

Colucci, D. et al. (1999) ‘The Vision Dome: Fully Immersive VR, in a Simple, Scaleable, and Collaborative Environment’, *SID Symposium Digest of Technical Papers*, 30(1), p. 620. doi: 10.1889/1.1834099.

Cook, N. (1998) *Analysing Musical Multimedia*. Oxford: Clarendon Press.

Correia, N. (2015) *Gen.AV 2*. Available at: http://www.gen-av.org/ (accessed 26 May 2017).

Correia, N. (2013) *Interactive Audiovisual Objects*. PhD Thesis. Helsinki: Aalto University School of Arts, Design and Architecture.

Correia, N. and Tanaka, A. (2014) ‘User-Centered Design of a Tool for Interactive Computer--Generated Audiovisuals’, in Sa, A., Carvalhais, M., and McLean, A. (eds) *ICLI 2014 - INTER-FACE: International Conference on Live Interfaces*. Porto: Porto University, pp. 86–99. ISBN 978-989-746-060-9. Available at: https://gala.gre.ac.uk/id/eprint/20971/ (accessed 15 April 2022).

Coulon, R. et al. (2020a) ‘Non-Euclidean Virtual Reality III: Nil’, *arXiv*. Available at: http://arxiv.org/abs/2002.00369 (accessed 15 April 2022).

Coulon, R. et al. (2020b) ‘Non-Euclidean Virtual Reality IV: Sol’, *arXiv*. Available at: https://arxiv.org/abs/2002.00369 (accessed 15 April 2022).

Coulter, J. (2010) ‘Electroacoustic Music with Moving Images: The Art of Media Pairing’, *Organised Sound*. Cambridge University Press, 15(1), pp. 26–34. doi: 10.1017/S1355771809990239.

Coulter, J. (2007) ‘The Language of Electroacoustic Music with Moving Images’, EMS : Electroacoustic Music Studies Network.

Coulter, J. (2005) ‘Multimedia Composition As Research’, in EMS: Electroacoustic Music Studies Network. Montreal, pp. 1–11.

Coxeter, H. S. M. (1948) *Regular Polytopes*. London: Methuen and Co. Ltd.

Cruz-Neira, C., Sandin, D. J. and DeFanti, T. A. (1993) 'Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE', *Proceedings of ACM SIGGRAPH 93 Conference on Computer Graphics*, pp. 135–142.

Csound (2020) *Introduction, Canonical Csound Manual*. Available at: https://csound.com/docs/manual/Introduction.html (accessed 29 September 2020).

Cytowic, R. (1995) 'Synesthesia: Phenomenology And Neuropsychology A Review of Current Knowledge', *Psyche: An Interdisciplinary Journal of Research on Consciousness*, 2(10), pp. 1–22.

Czernuszenko, M. et al. (1997) 'The ImmersaDesk and Infinity Wall Projection-Based Virtual Reality Displays', *Computer Graphics (ACM)*, 31(2), pp. 46–49. doi: 10.1145/271283.271303.

D-Fuse (2006) *VJ: Audio-Visual Art and VJ Culture*. Edited by M. Faulkner and D-Fuse. London: Laurence King Publishing.

Dannenberg, R. B. (2005) 'Interactive Visual Music: A Personal Perspective.', *Computer Music Journal*, 29(4), pp. 25–35. doi: 10.1016/S0041-1345(97)01103-2.

Dobrian, C. and Koppelman, D. (2006) 'The "E" in NIME: Musical Expression with New Computer Interfaces', *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression*. Paris: Zenodo, pp. 277–282. doi: 10.5281/zenodo.1176893.

Doornbusch, P. (2002) 'Composers' Views on Mapping in Algorithmic Composition', *Organised Sound*. Cambridge University Press, 7(02), pp. 145–156. doi: 10.1017/S1355771802002066.

Duff, A. (2014) *Vectrex: Minijack-input Mod*. Available at: http://users.sussex.ac.uk/~ad207/adweb/assets/vectrexminijackinputmod2014.pdf (accessed 14 April 2022).

Eisenstein, S. (1957) The Film Sense. Edited by J. Leyda. New York: Meridian Books.

Eisenstein, S. (2010). *Sergei Eisenstein Selected Works, Volume II, Towards a Theory of Montage*. Edited by Glenny, M. and Taylor, R. Translated from Russian by Glenny, M. London: I. B. Tauris.

Ernst, M. O. and Bülthoff, H. H. (2004) ‘Merging the Senses into a Robust Percept’, *Trends in Cognitive Sciences*, 8(4), pp. 162–169. doi: 10.1016/j.tics.2004.02.002.

Evans, B. (2005) ‘Foundations of a Visual Music’, *Computer Music Journal*, 29(4), pp. 11–24. doi: 10.1162/014892605775179955.

Evans, K. K. and Treisman, A. (2010) ‘Natural Cross-Modal Mappings Between Visual and Auditory Features.’, *Journal of Vision*, 10(1), pp. 1–12. doi: 10.1167/10.1.6.

Fiebrink, R. and Sonami, L. (2020) ‘Reflections on Eight Years of Instrument Creation with Machine Learning’, *International Conference on New Interfaces for Musical Expression*, pp. 237-242.

Fiebrink, R., Trueman, D. and Cook, P. R. (2009) ‘A Meta-Instrument for Interactive, On-The-Fly Machine Learning’, *New Interfaces for Musical Expression (NIME)*, Pittsburgh.

Fiebrink, R. and Trueman, D. (2012) ‘End-User Machine Learning in Music Composition and Performance’, *CHI 2012 Workshop on End-User Interactions with Intelligent and Autonomous Systems*, pp. 14–17.

Freed, A. et al. (2009) ‘Musical Applications and Design Techniques for the Gametrak Tethered Spatial Position Controller’, *Proceedings of the 6th Sound and Music Computing Conference, SMC 2009*. Porto: Sound and Music Computing Network, pp. 189–194.

Fry, R. (1920) *Vision and Design*. London: Chatto & Windus.

Fry, R. (1996) *A Roger Fry Reader*. Illustrated. Edited by C. Reed. University of Chicago Press.

Gainza, M., Lawlor, B. and Coyle, E. (2004) ‘Convention Paper’, in 117th Audio Engineering Society Convention. San Francisco: Audio Engineering Society, pp. 1–7.

Garro, D. (2012) ‘From Sonic Art to Visual Music: Divergences, Convergences, Intersections’, *Organised Sound*, 17(2), pp. 103–113. doi: 10.1017/S1355771812000027.

Gillies, M. (2016) ‘What is Movement Interaction in Virtual Reality For?’, *MOCO ’16: Proceedings of the 3rd International Symposium on Movement and Computing*. Thessaloniki: Association for Computing Machinery, pp. 1–4. doi: 10.1145/2948910.2948951.

Grierson, M. (2005) *Audiovisual Composition*. PhD Thesis. University of Kent.

Grierson, M. and Kiefer, C. (2011) ‘Maximilian: An Easy to Use, Cross Platform C++ Toolkit for Interactive Audio and Synthesis Applications’, in *Proceedings of The International Computer Music Conference* 2011. Huddersfield: University of Huddersfield, pp. 276–279.

Guldemond, J., Bloemheuvel, M. and Keefer, C. (2012) ‘Oskar Fischinger: An Introduction’, in Keefer, C. and Guldemond, J. (eds) *Oskar Fischinger 1900 - 1967: Experiments in Cinematic Abstraction*. Amsterdam, Los Angeles: EYE Filmmuseum, Center for Visual Music, pp. 10–30.

Hadley, M. (2020) Basics of Generating Meshes from an Image, openFrameworks: ofBook. Available at: https://openframeworks.cc/ofBook/chapters/generativemesh.html (Accessed: 29 September 2020).

Harley, M. A. (1998) ‘Music of Sound and Light: Xenakis’s Polytopes’, *Leonardo*, 31(1), pp. 55–65. doi: 10.2307/1576549.

Hart, V. et al. (2017a) ‘Non-Euclidean Virtual Reality I: Explorations of H^3’, *Bridges 2017 Conference Proceedings*, pp. 33–40. Available at: http://arxiv.org/abs/1702.04004 (accessed 15 April 2022).

Hart, V. et al. (2017b) ‘Non-Euclidean Virtual Reality II: Explorations of H^2 x E’, *Bridges 2017 Conference Proceedings*, pp. 41–48. Available at: http://arxiv.org/abs/1702.04862 (accessed 15 April 2022).

Heintz, J., McCurdy, I. and Neukom, M. (2015) ‘Physical Modelling’, *Csound FLOSS Manual*. Available at: http://floss.booktype.pro/csound/g-physical-modelling/ (accessed 16 June 2020).

Higgins, D. and Higgins, H. (2001) ‘Intermedia’, *Leonardo*, 34(1), pp. 49–54. doi: 10.1162/002409401300052514.

Holzer, D. (2017) ‘Vector Synthesis: An Investigation into Sound-Modulated Light’, *eContact!*, 19(2). Available at: https://econtact.ca/19_2/holzer_vectorsynthesis.html (accessed 14 April 2022).

Hunt, A. D., Wanderley, M. M. and Kirk, R. (2000) ‘Towards a Model for Instrumental Mapping in Expert Musical Interaction’, in *Proceedings of the International Computer Music Conference (ICMC 2000)*, ICMA: Berlin, Germany, pp. 209–12.

Hunt, A. and Kirk, R. (2000) ‘Mapping Strategies for Musical Performance’, in Wanderley, M. M. and Battier, M. (eds) *Trends in Gestural Control of Music*. Paris: Ircam - Centre Pompidou, pp. 231–258.

Hyde, J. (2012) ‘Musique Concrète: Thinking in Visual Music Practice: Audiovisual Silence and Noise, Reduced Listening and Visual Suspension’, *Organised Sound*, 17(2), pp. 170-178.

Ikeda, R. (2008) *datamatics*. Available at: http://www.ryojiikeda.com/project/datamatics/ (accessed 29 September 2020).

Ikeshiro, R. (2013) *Studio Composition: Live Audiovisualisation Using Emergent Generative Systems*. PhD Thesis. Goldsmiths, University of London.

Jaeger, H. (2014) *Conceptors: An Easy Introduction*. Paper. Jacobs University Bremen. Available at: http://arxiv.org/abs/1406.2671 (accessed 14 April 2022).

Katan, S. (2012) *Sight , Sound , the Chicken , and the Egg: Audio-Visual Co-dependency in Music*. PhD Thesis. Brunel University.

Katan, S. (2016) *Conditional Love*. Available at https://simonkatan.co.uk/projects/conditionallove.html (accessed 14 April 2022).

Kopeček, I. and Ošlejšek, R. (2008) ‘Hybrid Approach to Sonification of Color Images’, in 2008 Third International Conference on Convergence and Hybrid Information Technology. Busan: IEEE, pp. 722–727. doi: 10.1109/ICCIT.2008.152.

Latham, W. et al. (2021) ‘Exhibiting Mutator VR: Procedural Art Evolves to Virtual Reality’, *Leonardo*, 54(3), pp. 274-281. doi: https://doi.org/10.1162/leon_a_01857. Available at: http://research.gold.ac.uk/27620/ (accessed 15 April 2022).

Le Grice, M. (1982) *Abstract Film and Beyond*. Massachusetts: The MIT Press.

Lee, M., Freed, A. and Wessel, D. (1991) ‘Real-Time Neural Network Processing of Gestural and Acoustic Signals’, in *Proceedings of the International Computer Music Conference (ICMC)*. Montreal: Michigan Publishing, pp. 277–280.

Levin, O. (2018) ‘The cinematic time of the city symphony films: time management, experiential duration and bodily pulsation’, *Studies in Documentary Film*. Taylor & Francis, 12(3), pp. 225–238. doi: 10.1080/17503280.2018.1504370.

Lindroth, S. (2020) ‘Modal Frequency Ratios’, *The Canonical Csound Reference Manual*. Available at: http://www.csounds.com/manual/html/MiscModalFreq.html (accessed 29 September 2020).

Livingstone, S. R., Palmer, C. and Schubert, E. (2012) ‘Emotional Response to Musical Repetition’, *Emotion*, 12(3), pp. 552–567. doi: 10.1037/a0023747.

Lombard, M. and Ditton, T. (1997) ‘At the heart of it all: The concept of presence’, *Journal of Computer-Mediated Communication*, 3(2). doi: 10.1111/j.1083-6101.1997.tb00072.x.

Lund, C. and Lund, H. (eds) (2009) *Audio.Visual: On Visual Music and Related Media*. Illustrated. Stuttgart: Arnoldsche Art Publishers.

Ma, M.-Y. S. (2020) There is no soundtrack: Rethinking art, media, and the audio-visual contract. Manchester: Manchester University Press.

MacGillivray, C. et al. (2013) ‘The Diasynchronoscope : Bringing a new dimension to animation’, in *Confia (International Conference of Illustration and Animation)*, pp. 367–379. doi: 978-989-97567-6-2.

Magnusson, T., Eldridge, A. and Kiefer, C. (2020) ‘Instrumental Investigations at the Emute Lab’, *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 509–

513. Available at: https://www.nime.org/proceedings/2020/nime2020_paper97.pdf (accessed 14 April 2022).

Makela, M. (2008) *The Practice of Live Cinema*. Available at: http://i40474.wixsite.com/miamakela/publications (accessed 5 February 2018).

Martin, R. (2014) *Christian Marclay: Video Quartet*. Available at: https://www.tate.org.uk/art/artworks/marclay-video-quartet-t11818 (accessed 14 April 2022).

McCormack, J. and Dorin, A. (2001) 'Art, Emergence, and the Computational Sublime', in *Proceedings of the Second International Conference on Generative Systems in the Electronic Arts*. Victoria: Monash University Publishing, pp. 67–81. doi: 10.1.1.16.6640.

McCreery, M. P. et al. (2013) 'A Sense of Self: The Role of Presence in Virtual Environments', *Computers in Human Behavior*, 29(4), pp. 1635–1640. doi: 10.1016/j.chb.2013.02.002.

McDonnell, M. (2010) 'A Composition of the "Things Themselves": Visual Music in Practice', *eContact!*, 15(4). Available at: https://econtact.ca/15_4/mcdonnell_visualcomposition.html (accessed 14 April 2022).

McDonnell, M. (2014) 'Visual Music', *eContact!*, 15(4). Available at: http://www.econtact.ca/15_4/mcdonnell_visualmusic.html (accessed 14 February 2017).

McMahan, A. (2003) 'Immersion, Engagement, and Presence: A Method for Analyzing 3-D Video Games', in Wolf, M.J.P. and Perron, B. (eds.) *The Video Game Theory Reader*. New York; London: Routledge, pp. 89-108.

Minsky, M. (1980) 'Teleprescence', *Omni Magazine*, June.

Miranda, E. R. (2013) *Readings in Music and Artificial Intelligence*. ebook. London: Routledge. doi: 10.4324/9780203059746.

Mitchell, T. M. (1997) *Machine Learning*. New York: McGraw-Hill.

Mollaghan, A. (2015) *The Visual Music Film*. Basingstoke, United Kingdom: Palgrave MacMillan.

Momeni, A. and Henry, C. (2006) 'Dynamic Independent Mapping Layers for Concurrent Control of Audio and Video Synthesis', *Computer Music Journal*, 30(1), pp. 49–66. doi: 10.1162/comj.2006.30.1.49.

Moody, N., Fells, N. and Bailey, N. (2007) 'Ashitaka: An Audiovisual Instrument', in *Proceedings of the 2007 Conference on New Interfaces for Musical Expression* (NIME07), pp. 148-153. doi: 10.1145/1279740.1279767.

Moritz, W. (1998) 'Restoring the aesthetics of early abstract films', in Pilling, J. (ed.) A Reader in Animation Studies. Bloomington: Indiana University Press, pp. 221–227. Available at: https://muse.jhu.edu/book/40033.

Moritz, W. (1986) *Towards an Aesthetics of Visual Music*, Asifa Canada. Available at: http://www.centerforvisualmusic.org/TAVM.htm (accessed 14 May 2017).

Murch, W. (2000) 'Stretching Sound to Help the Mind See', *New York Times*, 1 October. Available at: http://www.filmsound.org/murch/stretching.htm (accessed 14 April 2022).

Murray, J. H. (1997) *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. New York: Free Press.

Nilsson, N. C., Nordahl, R. and Serafin, S. (2016) 'Immersion Revisited: A Review of Existing Definitions of Immersion and their Relation to Different Theories of Presence', *Human Technology*, 12(2), pp. 108–134. doi: 10.17011/ht/urn.201611174652.

Nunn, D. (2018) 'Software for vector synthesis and performance', in *Vector Hack Festival*. Zagreb, Ljubljana. Available at: http://arro.anglia.ac.uk/id/eprint/703981 (accessed 14 April 2022).

Ox, J. and Britton, D. (2000) 'The 21st Century Virtual Reality Color Organ', *IEEE MultiMedia*, 7(3), pp. 2–5.

Ox, J. and Keefer, C. (2008) *On Curating Recent Digital Abstract Visual Music*, Center for Visual Music. Available at: http://www.centerforvisualmusic.org/Ox_Keefer_VM.htm (accessed 23 June 2020).

Perl, T., Venditti, B. and Kaufmann, H. (2013) 'PS Move API: A Cross-Platform 6DoF Tracking Framework', *Proceedings of the Workshop on Off-The-Shelf Virtual Reality*. Orlando: IEEE Virtual Reality, p. 8. Available at: http://publik.tuwien.ac.at/files/PubDat_218820.pdf (accessed 15 April 2022).

Piché, J. (2004) *Interview with Paul Steenhuisen*. Available at: http://www.jeanpiche.com/Textes/Interview.htm (accessed 29 September 2020).

Puckette, M. and Zicarelli, D. (1990) *MAX - An Interactive Graphic Programming Environment*, Opcode Systems, Menlo Park, CA.

Punto y Raya (2020) Punto y Raya. Available at: https://www.puntoyrayafestival.com/ (Accessed: 7 September 2020).

Putnam, L., Latham, W. and Todd, S. (2017) 'Flow Fields and Agents for Immersive Interaction in Mutator VR: Vortex', *Presence: Teleoperators and Virtual Environments*, 26(2), pp. 138–156. doi: https://doi.org/10.1162/PRES_a_00290.

Quilez, I. (1999) *Geometric Orbit Traps*. Available at: https://www.iquilezles.org/www/articles/ftrapsgeometric/ftrapsgeometric.htm (accessed 29 September 2020).

Quilez, I. (2010) *Better Fog*. Available at: https://www.iquilezles.org/www/articles/fog/fog.htm (accessed 29 September 2020).

Quilez, I. (2011) *Menger Fractal*. Available at: https://iquilezles.org/www/articles/menger/menger.htm (accessed 29 September 2020).

Quilez, I. (2013) *Outdoor Lighting*. Available at: https://www.iquilezles.org/www/articles/outdoorslighting/outdoorslighting.htm (accessed 29 September 2020).

Quilez, I. (2015) *Normals for an SDF*. Available at: https://iquilezles.org/www/articles/normalsSDF/normalsSDF.htm (accessed 29 September 2020).

Quilez, I. (2020) Inigo Quilez. Available at: https://www.iquilezles.org/ (Accessed: 29 September 2020).

Quilez, I. (2020a) *Distance Functions*. Available at: https://www.iquilezles.org/www/articles/distfunctions/distfunctions.htm (accessed 29 September 2020).

Ramachandran, V. S. and Hubbard, E. M. (2001) 'Synaesthesia - A Window Into Perception, Thought and Language', *Journal of Consciousness Studies*, 8(12), pp. 3–34.

Rees, A. L. (2011) *A History of Experimental Film and Video*. 2nd edn. London: British Film Institute.

Richardson, J., Gorbman, C. and Vernallis, C. (eds) (2013) *The Oxford Handbook of New Audiovisual Aesthetics*. Illustrated. Oxford: Oxford University Press.

Riddle, L. (2020) *Wacklaw Sierpinski (1882 - 1969), Classic Iterated Function Systems*. Available at: https://larryriddle.agnesscott.org/ifs/siertri/sierbio.htm (accessed 29 September 2020).

Roads, C. (2004) *Microsound*. Cambridge, Massachusetts: MIT Press.

Rodgers, R. B. (1952) 'Cineplastics : The Fine Art of Motion Painting', The Quarterly of Film Radio and Television, 6(4), pp. 375–387. Available at: http://www.jstor.org/stable/1209948.

Rogers, H. (2013) *Sounding the Gallery: Video and the Rise of Art-Music*. Oxford: Oxford University Press.

Rogers, H. (2014). The Musical Script: Norman McLaren, Animated Sound, and Audiovisuality. *Animation Journal*, *22*, 68–84.

Rogers, H. (2019) 'The Audiovisual Eerie: Transmediating Thresholds in the Work of David Lynch', in Vernallis C., Rogers, H. and Perrott L. (eds.) *Transmedia Directors: Artistry, Industry and the New Audiovisual Aesthetics.* New York: Bloomsbury, pp. 241-270.

Rovan, J. B. et al. (1997) ‘Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance’, in Camurri, A. (ed.) *Kansei-The Technology of Emotion Workshop. Proceedings of the AIMI International Workshop*. DIST, pp. 68–73. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.4788&rep=rep1&type=pdf.

Russell, S. and Norvig, P. (2010) *Artificial Intelligence: A Modern Approach.* 3rd edn. Illustrated. Upper Saddle River: Prentice Hall. doi:10.1017/S0269888900007724.

Sá, A. (2016) *A Perceptual Approach to Audio-Visual Instrument Design, Composition and Performance*. PhD Thesis. Goldsmiths, University of London.

Sá, A., Caramieux, B. and Tanaka, A. (2014) ‘The Fungible Audio-Visual Mapping and its Experience’, *CITAR - Journal of Science and Technology of the Arts*, 6(1), pp. 85–96. doi: 10.7559/citarj.v6i1.131.

Sá, A., Caramieux, B. and Tanaka, A. (2014a) ‘A Study About Confounding Causation In Audio-Visual Mapping’, *Proceedings xCoAx 2014 - Computation, Communication, Aesthetics and X*, (2014), pp. 274-288. ISBN 978-989-746-036-4.

Schirm, J., Tullius, G. and Habgood, J. (2019) ‘Towards an Objective Measure of Presence: Examining Startle Reflexes in a Commercial Virtual Reality Game’, *CHI PLAY 2019 - Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play*, pp. 671–678. doi: 10.1145/3341215.3356263.

Seeing Sound (2016) *2016 Screenings*. Available at: http://www.seeingsound.co.uk/2016-screenings/ (accessed 7 September 2020).

Şentürk, S. et al. (2012) ‘Crossole: A Gestural Interface for Composition, Improvisation and Performance using Kinect’, *Proceedings of International Conference on New Interfaces for Musical Expression*. Ann Arbor: University of Michigan. Available at: https://www.nime.org/proceedings/2012/nime2012_185.pdf (accessed 15 April 2022).

Servotte, J.C. et al. (2020) ‘Virtual Reality Experience: Immersion, Sense of Presence, and Cybersickness’, *Clinical Simulation in Nursing*, 38, pp. 35–43. doi: 10.1016/j.ecns.2019.09.006.

Shams, L., Kamitani, Y. and Shimojo, S. (2002) 'Visual Illusion Induced by Sound', *Cognitive Brain Research*, 14(1), pp. 147–152. doi: 10.1016/S0926-6410(02)00069-1.

Siemens (2019) *Sound Pressure, Sound Power, and Sound Intensity: What's the difference?* Available at: https://community.sw.siemens.com/s/article/sound-pressure-sound-power-and-sound-intensity-what-s-the-difference (accessed 29 September 2020).

Sinnett, S., Spence, C. and Soto-Faraco, S. (2007) 'Visual Dominance and Attention: The Colavita Effect Revisited', *Perception and Psychophysics*, 69(5), pp. 673–686. doi: 10.3758/BF03193770.

Slater, M. (1999) 'Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire', *Presence: Virtual and Augmented Reality*, 8(5), pp. 560–565. doi: 10.1162/105474699566477.

Slater, M. (2003) 'A Note on Presence Terminology', *Presence - Connect*, 3(3), pp. 1–5.

Slater, M. (2009) 'Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), pp. 3549–3557. doi: 10.1098/rstb.2009.0138.

Slater, M. (2018) 'Immersion and the Illusion of Presence in Virtual Reality', *British Journal of Psychology*, 109(3), pp. 431–433. doi: 10.1111/bjop.12305.

Slater, M. et al. (1996) 'Immersion, Presence and Performance in Virtual Environments: An Experiment with Tri-Dimensional Chess', *VRST '96: Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. Hong Kong: Association for Computing Machinery, pp. 163–172. doi: 10.1145/3304181.3304216.

Slater, M. and Sanchez-Vives, M. V. (2016) 'Enhancing Our Lives with Immersive Virtual Reality', *Frontiers in Robotics and AI*, 3(74), pp. 1–47. doi: 10.3389/frobt.2016.00074.

Slater, M. and Steed, A. (2000) 'A Virtual Presence Counter', *Presence: Teleoperators and Virtual Environments*, 9(5), pp. 413–434. doi: 10.1162/105474600566925.

Slater, M., Usoh, M. and Steed, A. (1995) ‘Taking Steps: The Influence of a Walking Technique on Presence in Virtual Reality’, *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2(3), pp. 201–219. doi: 10.1145/210079.210084.

Sound/Image (2018). Available at: http://www.gre.ac.uk/ach/events/soundimage (Accessed: 7 September 2020).Seeing Sound (2020). Available at: http://www.seeingsound.co.uk/ (Accessed: 7 September 2020).

Spence, C. (2011) ‘Crossmodal Correspondences: A Tutorial Review’, *Attention, Perception, & Psychophysics*, 73(4), pp. 971–995. doi: 10.3758/s13414-010-0073-7.

Sterken, S. (2001) ‘Towards a Space-Time Art: Iannis Xenakis’s Polytopes’, *Perspectives of New Music*, 39(2), pp. 262–273. Available at: http://www.jstor.org/stable/833570 (accessed 15 April 2022).

Tanaka, A. and Ortiz, M. (2017) ‘Gestural Musical Performance with Physiological Sensors, Focusing on the Electromyogram’, in Lesaffre, M., Maes, P.J., and Leman, M. (eds.) *The Routledge Companion to Embodied Music Interaction*. Oxford: Routledge, pp. 45–58.

The Light Surgeons (2007) *True Fictions*. Available at: http://www.lightsurgeons.com/art/true_fictions/ (accessed 29 September 2020).

Thomas, R. (2017) Malcolm LeGrice and Keith Rowe: After Leonardo, *The Wire*. Available at: https://www.thewire.co.uk/in-writing/essays/p=10428 (accessed 9 June 2017).

Todd, J. A. and Coxeter, H. S. M. (1936) ‘A Practical Method for Enumerating Cosets of a Finite Abstract Group’, *Proceedings of the Edinburgh Mathematical Society*, 5(1), pp. 26–34. doi: 10.1017/S0013091500008221.

Todorovic, D. (2008) ‘Gestalt Principles’, *Sholarpedia*, 3(12), p. 5345. Available at: http://www.scholarpedia.org/article/Gestalt_principles (accessed 15 April 2022).

Valcke, J. (2008) *Static Films and Moving Pictures: Montage in Avant-Garde Photography and Film*. PhD Thesis. University of Edinburgh. Available at: http://hdl.handle.net/1842/4061 (accessed 14 April 2022).

Varela, F. J., Thompson, E. and Rosch, E. (2016) 'Enaction: Embodied Cognition', in *The Embodied Mind: Cognitive Science and Human Experience*. Rev. ed. Cambridge, Massachusetts; London, England: MIT Press, pp. 147–183. Available at: https://muse.jhu.edu/book/49607 (accessed 15 April 2022).

Vepstas, L. (1997) *Renormalizing the Mandelbrot Escape*. Available at: http://linas.org/art-gallery/escape/escape.html (accessed 15 September 2020).

Weech, S., Kenny, S. and Barnett-Cowan, M. (2019) 'Presence and Cybersickness in Virtual Reality are Negatively Related: A Review', *Frontiers in Psychology*, 10(FEB), pp. 1–19. doi: 10.3389/fpsyg.2019.00158.

Weinel, J. (2019) 'Cyberdream VR: Visualizing Rave Music and Vaporwave in Virtual Reality', *ACM International Conference Proceeding Series*, pp. 277–281. doi: 10.1145/3356590.3356637.

Weisling, A. et al. (2018) 'Surveying the Compositional and Performance Practices of Audiovisual Practitioners', *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 344–345.

White, D. (2009) *The Unravelling of the Real 3D Mandelbulb*. Available at: https://www.skytopia.com/project/fractal/mandelbulb.html (accessed 18 June 2020).

Whitelaw, M. (2008) 'Synesthesia and Cross-Modality in Contemporary Audiovisuals', *The Senses and Society*, 3(3), pp. 259–276. doi: 10.2752/174589308x331314.

Whitney, J. (1980) *Digital Harmony: On the Complementarity of Music and Visual Art*. Peterborough New Hampshire: Byte Books / McGraw-Hill.

Whitney, M. (1997) 'The Whitney Archive: A Fulfillment of a Dream', *Animation World Magazine*, August.

Witmer, B. G., Jerome, C. J. and Singer, M. J. (2005) 'The Factor Structure of the Presence Questionnaire', *Presence: Virtual and Augmented Reality*, 14(3), pp. 298–312. doi: 10.4324/9780203071175-15.

Witmer, B. G. and Singer, M. J. (1998) ‘Measuring Presence in Virtual Environments: A Presence Questionnaire’, *Presence: Teleoperators and Virtual Environments*, 7(3), pp. 225–240. doi: 10.1162/105474698565686.

Youngblood, G. (1970) *Expanded Cinema*. New York: P. Dutton & Co. Inc.

Zbyszyński, M. *et al.* (2021) ‘Gesture-Timbre Space: Multidimensional Feature Mapping Using Machine Learning and Concatenative Synthesis’, in Kronland-Martinet, R., Ystad, S., and Aramaki, M. (eds) *Perception, Representations, Image, Sound, Music*. Springer International Publishing, pp. 600–622. doi: 10.1007/978-3-030-70210-6_39.

List of Artwork

Bach, J. S. (BWV565) Tocatta and Fugue BWV565, Deutsche Grammophon. doi: 427668-2.

Belson, J. (1970) World, Center for Visual Music. Available at: https://vimeo.com/89192891 (Accessed: 29 September 2020).

Björk (2011) Biophilia, Mobile App. Available at: https://apps.apple.com/gb/app/björk-biophilia/id434122935 (Accessed: 29 September 2020).

Bute, M. E. (1938) Synchromy No.4, Whitney Museum of American Art. Available at: https://www.youtube.com/watch?v=YRmu-GcClls (Accessed: 29 September 2020).

Bute, M. E. (1948) Color Rhapsodie, Center for Visual Music. Available at: https://vimeo.com/208709266 (Accessed: 29 September 2020).

Buyukberber, C. and Uyanik, Y. (2016) Morphogenesis. Available at: https://canbuyukberber.com/morphogenesis-vr (Accessed: 29 September 2020).

Cardew, C. (1967) Treatise, Edition Peters. Available at: https://www.editionpeters.com/product/treatise/ep7560 (Accessed: 29 September 2020).

Cassidy, J. (2012) They Upped Their Game After The Oranges. Available at: http://www.janecassidy.net/installations.html (Accessed: 29 September 2020).

Conrad, T. (1966) The Flicker, LUX. Available at: https://lux.org.uk/work/the-flicker (Accessed: 29 September 2020).

Correia, N. (2008) AVOL. Available at: http://www.nunocorreia.com/portfolio/avol (Accessed: 29 September 2020).

Correia, N. (2010) AV Clash. Available at: http://www.nunocorreia.com/portfolio/av-clash (Accessed: 29 September 2020).

Eats, M. (2015) This City. Available at: https://vimeo.com/channels/staffpicks/140019134 (Accessed: 29 September 2020).

Eggeling, V. (1924) Symphonie Diagonale, Museo Nacional Centro De Arte, Reina Sofia. Available at: https://www.youtube.com/watch?v=w7aeub6Wanc&vl=en (Accessed: 29 September 2020).

Fischinger, O. (1932) Ornament Sounds, Center for Visual Music. Available at: https://vimeo.com/ondemand/26951 (Accessed: 29 September 2020).

Glass, P. and Wilson, R. (1979) Einstein on the Beach, Nonesuch 79323-2. Available at: https://philipglass.com/recordings/einstein_on_the_beach_none/ (Accessed: 29 September 2020).

Greenaway, P. (1978) Vertical Features Remake, The Early Films of Peter Greenaway Volume 2, DVD. BFIVD565. Available at: https://www2.bfi.org.uk/blu-rays-dvds/early-films-peter-greenaway-volume-2 (Accessed: 29 September 2020).

Greenaway, P. (1980) The Falls, The Early Films of Peter Greenaway Volume 2, DVD. BFIVD565. Available at: https://www2.bfi.org.uk/blu-rays-dvds/early-films-peter-greenaway-volume-2 (Accessed: 29 September 2020).

Greenaway, P. (2007) The Tulse Luper Suitcases, Output includes three feature films, a TV series, 92 DVDs, CD-ROMs, books and live VJ performances. Available at: http://www.tulseluperjourney.com/index.jsp (Accessed: 29 September 2020).

Grierson, M. (2005) ‘Light Speak’, in Audiovisual Composition. PhD. Thesis. London: University of Kent.

Hein, W. and Hein, B. (1968) Rohfilm, LUX. Available at: https://lux.org.uk/work/rohfilm (Accessed: 29 September 2020).

Ikeda, R. (2008) datamatics[prototype-ver2.0]. Available at: http://www.ryojiikeda.com/project/datamatics/ (Accessed: 29 September 2020).

Ikeda, R. (2009) data.tron[8K enhanced version]. Available at: http://www.ryojiikeda.com/project/datamatics/#datatron_8k_enhanced_version (Accessed: 29 September 2020).

Ikeshiro, R. (2011) Construction in Zhuangzhi. Available at: https://ryoikeshiro.wordpress.com/2011/07/01/construction-in-zhuangzi/ (Accessed: 29 September 2020).

Ikeshiro, R. (2013) Construction in Kneading, Optical Research DVD, Hardcore Jewellery JEWEL-002. Available at: https://ryoikeshiro.wordpress.com/2013/05/01/construction-in-kneading/ (Accessed: 29 September 2020).

Kandinsky, W. (1914) Improvisation 35, Kunstmuseum, Basel. Available at: https://www.wassilykandinsky.net/work-513.php (Accessed: 29 September 2020).

Katan, S. (2012) Cube With Magic Ribbons. Available at: https://simonkatan.co.uk/projects/cubewithmagicribbons.html (Accessed: 29 September 2020).

Katan, S. (2016) Conditional Love. Available at: https://simonkatan.co.uk/projects/conditionallove.html (Accessed: 29 September 2020).

Kiefer, C. (2018) *10K video.* Performance for neural networks, sound and video. Available at: https://vimeo.com/268980331 (accessed 17 December 2021).

Kurokawa, R. (2016) unfold.alt. Available at: http://www.ryoichikurokawa.com/project/ualt.html (Accessed: 29 September 2020).

Land, O. (1966) Film in Which There Appears Edge Lettering, Sprocket Holes, Dirt Particles Etc., LUX. Available at: https://lux.org.uk/work/film-in-which-there-appear-sprocket-holes-edge-lettering-dirt-particles-etc (Accessed: 29 September 2020).

Latham, W., Putnam, L. and Devlin, S. (2016) Mutator VR. Available at: http://mutatorvr.co.uk/ (Accessed: 29 September 2020).

Le Grice, M. (1967) Little Dog For Roger. Available at: https://www.malcolmlegrice.com/1960s (Accessed: 29 September 2020).

Le Grice, M. (1973) After Leonardo, LUX. Available at: https://lux.org.uk/work/after-leonardo (Accessed: 29 September 2020).

Léger, F. (1924) Ballet Mécanique, Museum of Modern Art, New York. Available at: https://www.moma.org/learn/moma_learning/fernand-leger-ballet-mecanique-1924/ (Accessed: 29 September 2020).

Loicvdb (2019) *Glass Mandelbulb*, ShaderToy. Available at: https://www.shadertoy.com/view/tdtGRj (accessed 29 September 2020).

Light Surgeons, The (2007) True Fictions, The Light Surgeons. Available at: http://www.lightsurgeons.com/art/true_fictions/ (Accessed: 29 September 2020).

Lutoslawski, W. (1961) Venetian Games. Available at: http://www.lutoslawski.org.pl/en/composition,52.html (Accessed: 29 September 2020).

McDonnell, M. and McDonnell, B. (2016) Duel Tones. Available at: https://www.mauramcdonnell.com/duel-tones (Accessed: 29 September 2020).

McKenna, B. (2017) *Continuously Variable Colour Field.* Installation with custom electronics, CRT monitors and multichannel audio. Available at: https://vimeo.com/250821574 (accessed 03 February 2022).

McLaren, N. (1971) Synchromy, National Film Board of Canada. Available at: https://vimeo.com/29399459 (Accessed: 29 September 2020).

Miyama, C. (2011) Quicksilver. Available at: https://www.youtube.com/watch?v=nkKjHdfiK14&feature=youtu.be (Accessed: 29 September 2020).

Piché, J. (2004) Sieves. Available at: http://www.jeanpiche.com/sieves.htm (Accessed: 29 September 2020).

Prudence, P. (2015) Cyclotone III. Available at: https://www.transphormetic.com/Cyclotone-III (Accessed: 29 September 2020).

Radiohead (2014) Polyfauna. Available at: https://universaleverything.com/projects/polyfauna (Accessed: 29 September 2020).

Rhodes, L. (1972) Dresden Dynamo, LUX. Available at: https://lux.org.uk/work/dresden-dynamo (Accessed: 29 September 2020).

Rhodes, L. (1977) Light Music, LUX. Available at: https://lux.org.uk/work/light-music (Accessed: 29 September 2020).

Richter, H. (1921) Rhythmus 21, Museum of Modern Art, New York. Available at: https://www.youtube.com/watch?v=239pHUy0FGc (Accessed: 29 September 2020).

Rimmer, D. (1970) Surfacing on the Thames, LUX. Available at: https://lux.org.uk/work/surfacing-on-the-thames1 (Accessed: 29 September 2020).

Ruttmann, W. (1921) Lichtspiel Opus 1. Available at: https://vimeo.com/262027844 (Accessed: 29 September 2020).

Sorensen, V. (2015) Mayur. Available at: http://vibeke.info/mayur/ (Accessed: 29 September 2020).

Tanaka, A. and Kogelsberger, U. (2020) Atau Tanaka & Uta Kogelsberger - ARRAY Music Festival - Saturday 6th June 2020. Available at: https://www.youtube.com/watch?v=0WE-omAUxpw&feature=youtu.be (Accessed: 29 September 2020).

Trope (2013) One, Two, Three... Available at: http://www.trope-design.com/exhibited-artworks/one-two-three/ (Accessed: 29 September 2020).

Trope (2014) Dandelion. Available at: http://www.trope-design.com/exhibited-artworks/dandelion/ (Accessed: 29 September 2020).

Trope (2014) Codex. Available at: http://www.trope-design.com/exhibited-artworks/codex/ (Accessed: 29 September 2020).

Vasulka, S. (1978) Violin Power, The Smithsonian American Art Museum. Available at: https://americanart.si.edu/artwork/violin-power-77216 (Accessed: 29 September 2020).

Vasulka, W. (1976) No. 25, Daniel Langlois Foundation. Available at: https://www.fondation-langlois.org/html/e/media.php?NumObjet=11250 (Accessed: 29 September 2020).

Weinel, J. (2019) Cyberdream VR. Available at: http://www.jonweinel.com/projects.html (Accessed: 29 September 2020).

Whitney, John and Whitney, James (1944) Five Film Exercises, Academy of Motion Picture Arts and Sciences. Available at: https://www.oscars.org/film-archive/collections/whitney-collection (Accessed: 29 September 2020).

Whitney, J. (1972) Matrix III, Pyramid 09191. Available at: https://www.youtube.com/watch?v=ZrKgyY5aDvA (Accessed: 29 September 2020).

Whitney, J. (1975) Arabesque. Available at: https://www.youtube.com/watch?v=w7h0ppnUQhE&feature=youtu.be (Accessed: 29 September 2020).

Whitney, J. (1991) Moon Drum. Available at: https://archive.org/details/JohnWhitneyMoonDrum1991 (Accessed: 29 September 2020).

Xenakis, I. (1956) Pithoprakta, Timpani Recordings (2008), Arturo Tamayo, Luxembourg Philharmonic Orchestra, Pierre Carré (Graphical Score). Available at: https://www.youtube.com/watch?v=nvH2KYYJg-o (Accessed: 29 September 2020).

Xenakis, I. (1974) Polytope de Cluny, Xenakis: Electronic Works, Vol. 2, NaxosofAmerica(On behalf of Node Records). Available at: https://www.youtube.com/watch?v=nVx0PvK9TnQ (Accessed: 29 September 2020).

Appendix A: *Obj_#3* Code

This appendix contains some of the code used in the development of *Obj_#3*. A brief description of the code functionality is given with each code snippet.

A.1 Environment

i) Ex. A.1 is part of the code used to render the white room described in section 8.1.1. The method used to map the cube textures meant that the texture coordinates were placed at infinity as shown in the code example. This is achieved in the vertex shader by setting the *z* component to *w* so that its value is always 1.0.

```
00 vec4 projectedPos = projMat * viewMat * modelMat * vec4(position, 1.0);
01 gl_Position = projectedPos.xyww;
```

Example A.1. Z coordinate placed at infinity.

ii) The plane is created using a simple signed distance function (see Ex. A.2). This function is adapted from Quilez (2020a).

```
00 float planeSDF(vec3 pos, vec4 normal)
01 {
02     return dot(pos, normal.xyz) + normal.w;
03 }
```

Example A.2. Plane SDF adapted from Quilez (2020b).

In the code block above the dot product is calculated between the ray position and the normal to the plane. The result of the dot product will be 0 if the two vectors are orthogonal. This means that if *pos* is; on the plane the result will be 0, above the plane the result will be positive, below the plane the result will be negative. The *w* component of *normal* represents the height of the plane from the origin, which is 0 in this case. In the *scenedSDF* function, the ray position is subject to several transformations as shown in Ex. A.3.

```
00 vec3 newPos = pos;
01 float function1x = 0.09 * sin(newPos.x * 0.4) * newPos.x;
02 float function1z = 0.09 * sin(newPos.z * 0.4) * newPos.z;
03
04 float function2x = clamp(function1x, 0.0, 3.0);
05 float function2z = clamp(function1z, 0.0, 3.0);
06
07 newPos.y += function2x;
08 newPos.y += function2z;
09
10 planeDist = planeSDF(newPos, PLANE_NORMAL);
```

Example A.3. Transformation of point.

At line 00 the vector *newPos* represents the ray position. The *x* and *z* components are used to generate repeating sine patterns at lines 01 and 02. These are then clamped between 0.0 and 3.0 which results in the flat topped ridges seen in the plane. These values are then added to the *y* component of the ray position before it is passed to *planeSDF()*. These transformations were derived through trial and error. Experiments with mathematical functions were conducted using the *Graphical Function Explorer*[125]. These functions were then implemented in code and further adjusted to arrive at the desired result. This process was inspiring in that it revealed a playful and profoundly artistic approach to visualising pure mathematical functions. The colour of the plane is achieved through basic Phong illumination. The large sun is achieved through a combination of environment lighting adapted from Quilez (2013) and the implementation of a *fog()* function also adapted from Quilez (2010).

A.2 Visual Rendering

The Mandelbulb signed distance function is adapted from *Glass Mandelbulb* (loicvdb 2019). The function is shown in Ex. A.4. The main Mandelbulb formula is shown at line 25. Here the *y* and *z* terms are swapped to rotate the object. The rendered object is affected in real-time using the uniforms *thetaScale*, *phiScale* and *fftBinValScale*. These uniforms contain data values mapped from the audio material.

[125] https://www.mathopenref.com/graphfunctions.html (accessed 17/09/2020).

```
00 float mandelbulbSDF(vec3 pos)
01 {
02     float Power = 8.0;
03     float r = length(pos);
04     if(r > 1.5) return r-1.2;
05     vec3 z = pos;
06     float dr = 1.0, theta, phi;
07 
08     for (int ® = 0; ® < 5; i++)
09     {
10           r = length(z);
11           if (r > 1.5) break;
12 
13           // convert to polar coordinates
14           theta = acos(z.y / r) * thetaScale;
15           phi = atan(z.z, z.x) * phiScale;
16 
17           // length of the running complex derivative
18           dr =  pow(r, Power-1.0) * Power * dr * (0.7 + lowFreqVal *
fftBinValScale) + 1.0;
19 
20           // scale and rotate
21           theta *= Power;
22           phi *= Power;
23 
24           // mandelbulb formula with y and z terms swapped to rotate the
object
25           z = pow(r, Power) * vec3(sin(theta) * cos(phi), cos(theta),
sin(phi) * sin(theta)) + pos;
26     }
27     return abs(0.5 * log® * r / dr);
28 }
```

Example A.4. Mandelbulb SDF adapted from loicvdb(2019).

A.3 Audio Rendering

i) The instrument that generates the environmental noise is shown in Ex. A.5.

```
00 ;*****************************************************************
01    instr 1 ; Real-time Spectral Instrument - Environmental Noise
02 ;*****************************************************************
03
04 ; get control value from application
05 kSineControlVal  chnget       "sineControlVal"
06
07 ; pink noise generator
08 ares      fractalnoise ampdbfs(-24), 1
09
10 ifftsize = 2048
11 ioverlap = ifftsize / 4
12 iwinsize = ifftsize * 2
13 iwinshape = 0
14
15 fsig      pvsanal      ares, ifftsize, ioverlap, iwinsize, iwinshape
16
17 ; get info from pvsanal and print
18 ioverlap, inbins, iwindowsize, iformat     pvsinfo      fsig
19 print   ioverlap, inbins, iwindowsize, iformat
20
21 ifn = 1
22 kdepth = 0.99 + (0.01 * kSineControlVal)
23
24 fmask   pvsmaska       fsig,   ifn,    kdepth
25
26 aOut0   pvsynth fmask
27 outs    aOut0 * 0.05,   aOut0 * 0.05
28
29 endin
```

Example A.5. Real-time spectral instrument.

The core of the instrument is the *fractalnoise* opcode at line 08 that generates a pink noise signal. All Csound opcode documentation can be found online in *The Canonical Csound Reference Manual.*[126] At line 15, the *pvsanal* opcode applies an FFT to the output of *fractalnoise*. The frequency domain signal is then passed to *pvsmaska* at line 24 which dynamically filters the signal based on values from a function table created using the *GEN08* subroutine. This subroutine creates smooth curved functions. The dynamic element of the opcode is controlled by the *kdepth* variable. This variable is modulated by *kSineControlVal* at line 22 which is sent from *Studio::Update()* in the main C++ application. The variable *kSineControlVal* is a float calculated using a sine function that is given the

[126] http://www.csounds.com/manual/html/ (accessed 29/09/2020).

current time value from the OpenGL context. The value output by *pvsmaska* is always between 0.99 and 1.0. This means that the signal is modulated almost entirely by the values from the function table with a subtle amount of variation. The environmental noise is intended to be unobtrusive and it was felt that too much modulation would have distracted from the foreground elements. The signal is then passed from *pvsmaska* to *pvsynth* at line 26 where an inverse FFT is applied. The re-synthesised signal is then passed to the main stereo output.

ii) A granular instrument was implemented within the CSound orchestra. This instrument makes up the core of the foreground audio in this piece (see Ex. A.6).

```
;*********************************************************************
   instr 3 ; Granular Instrument
;*********************************************************************

kCps            chnget  "grainFreq"
kPhs      chnget  "grainPhase"
kFmd      chnget  "randFreq"
kPmd      chnget  "randPhase"
kGDur     chnget  "grainDur"
kDens     chnget  "grainDensity"
kFrPow    chnget  "grainFreqVariationDistrib"
kPrPow    chnget  "grainPhaseVariationDistrib"

; initialisation to avoid perf error 0.0
kGDur = 0.01 + kGDur
kDens = 1 + kDens

iMaxOvr = 2000
kFn = 3

aOut3    grain3  kCps, kPhs, kFmd, kPmd, kGDur, kDens, iMaxOvr, kFn,
giWFn, kFrPow, kPrPow

kAmp    linseg 0.0,     p3 * 0.1,       0.95,   p3 * 0.1,       0.8,
p3 * 0.6,       0.8,    p3 * 0.1,       0.0

kfe  expseg p4, p3*0.3, p5, p3*0.1, p6, p3*0.2, p7, p3*0.3, p8, p3*0.1,
p9

; vary resonance
kres linseg 0.1, p3 * 0.2, 0.3, p3 * 0.4, 0.25, p3 * 0.2, 0.5, p3 * 0.2,
0.35

afil moogladder aOut3, kfe, kres

gaGranularOut = afil * kAmp

endin
```

Example A.6. Granular instrument.

The instrument is centred around the *grain3* opcode at line 20. The following parameters are manipulated at control rate:

- Grain frequency in Hz (*kCps*)
- Grain phase (*kPhs*)
- Random variation in grain frequency (*kFmd*)
- Random variation in start phase (*kPmd*)
- Grain duration in seconds (*kGDur*)
- Number of grains per second (*kDens*)
- The distribution of grain frequency variation (*kFrPow*)
- The distribution of random phase variation (*kPrPow*)

These parameter values are all sent from the *Studio::Update()* function in the C++ application. The generation of these values are discussed in section 8.4.3. The variable *kFn* at line 18 holds the number of a function table that contains the grain waveform. Here it is a sawtooth wave. Although it is possible to switch waveforms at the control rate, a single table is implemented here. The implementation of a range of tables was attempted but switching between them resulted in audible discontinuities.

The signal is then sent to the *moogladder* filter at line 29 whose frequency is dynamically modulated using an exponential function, *expseg* at line 24, and whose resonance is modulated using a linear function, *linseg* at line 27. The values used by *expseg* vary each time the instrument is triggered. The duration of *linseg* is dictated by the duration value sent by the triggering instrument (see Example A.7 below). This filtered signal is then multiplied by an *adsr* envelope (*kAmp*) before being sent to a global output channel (*gaGranularOut*). The output is then sent to the sound localisation instrument that situates the sound source within the virtual environment.

The granular instrument is triggered every 1 to 5 seconds with a random duration in the range of 0.1 to 10 seconds. This code uses the *schedkwhen* opcode at line 19 to trigger the granular instrument (see Ex. A.7).

```
00 ;*********************************************************************
01    instr 2 ; note scheduler
02 ;*********************************************************************
03
04 kGaussVal gauss 6.0
05
06 seed 0
07 kRand random 0.1, 10.0
08
09 seed 0
10 kRand2 random 1, 5
11
12 kTrigger metro kRand2
13 kMinTim = 0
14 kMaxNum = 1
15 kInsNum = 3
16 kWhen = 0
17 kDur = kRand
18
19 schedkwhen kTrigger, kMinTim, kMaxNum, kInsNum, kWhen, kDur,
1000+kGaussVal, 1400+kGaussVal, 1200+kGaussVal, 800+kGaussVal,
700+kGaussVal, 1000+kGaussVal
20
21 aOut oscil 0,   100
22
23 outs aOut, aOut
24
25 endin
```

Example A.7. Note scheduler.

The parameter *kTrigger*, at line 12, is the output of the *metro* opcode. This outputs a value of 1 at a time interval dictated by *kRand2*. This value is calculated at line 10 using *random*. The parameter *kDur* at line 17 is assigned a random value between 0.1 and 10.0 using the *random* opcode at line 07. This determines the length of the note. The parameters listed after *kDur* relate to what would be *p-fields* in the score file for the instrument being triggered. These are used to dictate the shape of the *linseg* and *expseg* envelopes of the granular instrument. These parameters are made up of random values with a gaussian distribution between the range -6.0 and 6.0. This is achieved using the *gauss* opcode at line 04. Using this method, randomisation was introduced into the tonal characteristics of the granular instrument.

A.4. Mapping Layers

i) In the *Studio::Update()* function a single sound source is defined. Ex. A.8 contains the code used to position the sound source in the scene.

```
00 // situated sound source
01 glm::vec4 objPosition = glm::vec4(0.0f, 0.0f, 0.0f, 1.0f);
02 glm::mat4 objModelMatrix = glm::mat4(1.0f);
03
04 glm::vec4 soundPosCameraSpace = viewMat * objModelMatrix * objPosition;
05 glm::vec4 viewerPosCameraSpace = glm::vec4(0.0f, 0.0f, 0.0f, 1.0f);
06
07 // distance value camera space
08 float distCamSpace = sqrt(pow(soundPosCameraSpace.x, 2) +
pow(soundPosCameraSpace.y, 2) + pow(soundPosCameraSpace.z, 2));
09
10 // azimuth in camera space
11 float valX = soundPosCameraSpace.x - viewerPosCameraSpace.x;
12 float valZ = soundPosCameraSpace.z - viewerPosCameraSpace.z;
13 float azimuth = atan2(valX, valZ);
14 azimuth *= (180.0f / PI);
15
16 // elevation in camera space
17 float oppSide = soundPosCameraSpace.y - viewerPosCamerSpace.y;
18 float sinVal = oppSide / distCamSpace;
19 float elevation = asin(sinVal);
20 elevation *= (180.0f / PI);
21
22 // assign values to Csound pointers
23 *hrtfVals[0] = (MYFLT)azimuth;
24 *hrtfVals[1] = (MYFLT)elevation;
25 *hrtfVals[2] = (MYFLT)distCamSpace;
```

Example A.8. Obj_#3 sound placement.

The location of the object is defined with the vector *objPosition* at line 01 in Ex. A.8. It is positioned at the origin. The object's model matrix is assigned to *objModelMatrix*. There are no translations, scaling or rotating operations that are relevant to the sound source so the model matrix is simply an identity matrix. These are multiplied by the view matrix at line 04 to give the sound position in camera space. The viewer position in camera space is then given as the origin. This is not explicitly needed but it is useful for readability. After these vectors have been calculated, the distance value in camera space is found. This is given by the formula at line 08. This distance is then assigned to the third Csound float at line 25. The azimuth angle is then calculated. This is an angle on the horizontal plane. Therefore, only the *x* and *z* components of the relevant vectors are needed. These values are used as arguments for *atan2()* at line 13. This function calculates the arc tangent of *valX / valZ* and returns the value in radians. At line 14 this value is converted to degrees before being sent to Csound at line

23. Finally the elevation angle is calculated. This is an angle on the vertical plane between the line of sight of the camera and the height of the object. At line 17 the difference between the sound source *y* component and the camera *y* component is found. The sine of the angle is found by dividing the float *oppSide* by the distance between the camera and the sound source. The elevation in radians is calculated using *asin()* at line 19. This is then converted to degrees before being sent to Csound at line 24.

```
00 kPortTime linseg       0.0, 0.001, 0.05
01
02 kAzimuthVal      chnget       "azimuth"
03 kElevationVal    chnget       "elevation"
04 kDistanceVal     chnget       "distance"
05
06 kDist            portk        kDistanceVal, kPortTime
07
08 aSig = gaGranularOut * 0.5
09
10 aLeftSig, aRightSig    hrtfmove2    aSig, kAzimuthVal, kElevationVal,
"hrtf-48000-left.dat", "hrtf-48000-right.dat", 4, 9.0, 48000
11
12 aLeftSig = aLeftSig / (kDist + 0.00001)
13 aRightSig = aRightSig / (kDist + 0.00001)
14
15 aL = aLeftSig * 0.8
16 aR = aRightSig * 0.8
17
18 outs      aL, aR
```

Example A.9. 3D source location instrument.

The *3D Source Location Instrument* in Csound (see Ex. A.9) receives the azimuth, elevation and distance values from *Studio::Update()* at lines 02 to 04 above. At line 08 the global audio output from the *Granular Instrument* is assigned to *aSig*. This signal is then used as the input to *hrtfmove2* at line 10. This opcode also takes *kAzimuthVal* and *kElevationVal* as input data. It uses the data files *hrtf-48000-left.dat* and *hrtf-48000-right.dat* as sources for the required spectral data. These files were sourced in the Csound Github[127] repository. This opcode outputs a stereo signal that calculates the position of the sound source on the horizontal plane. The left and right output signals are then divided by *kDist* at lines 12 and 13. This value is calculated at line 06 using the opcode *portk*. This opcode applies a portamento between distance values to smooth out the signal. This measure was taken to avoid audio discontinuities when the distance changed. The left and right signals are then scaled and sent to the stereo output at line 18.

[127] https://github.com/csound/csound/tree/develop/samples (accessed 11/01/2022)

ii) The output from the *Granular Instrument* is sent to the *Spectral Analysis Instrument* shown in Ex. A.10.

```
00 fSig pvsanal gaGranularOut, iFFTSize, iOverlap, iWinSize, iWinShape
01
02 iOverLap, inBins, iWindowSize, iFormat pvsinfo fSig
03
04 iFreqTableftgen        0, 0, inBins, 2, 0
05 iAmpTable ftgen        0, 0, inBins, 2, 0
06
07 kFlag     pvsftw fSig, iAmpTable, iFreqTable
08
09  if kFlag == 0 goto contin
10
11 kCount = 0
12 loop:
13    kAmp   tablekt      kCount, iAmpTable
14
15    S_channelNamesprintfk "fftAmpBin%d", kCount
16                        chnset kAmp, S_channelName
17
18    loop_lt             kCount, 1, inBins, loop
19
20 contin:
21 endin
```

Example A.10. Spectral analysis instrument.

At line 00 *pvsanal* analyses the signal *gaGranularOut* and transforms it to the frequency domain using an FFT. The frequency domain signal *fSig* is then passed to *pvsinfo* at line 02. This opcode outputs information about the FFT such as *inBins* which is the number of analysis bins used in the transform. This value is then used at lines 04 and 05 to determine the size of the tables used to hold frequency and amplitude data. At line 07 the opcode *pvsftw* is used to write the amplitude and frequency data contained in *fSig* to the tables *iFreqTable* and *iAmpTable*. This opcode outputs *kFlag* which has the value 1 when there is new data available and 0 otherwise. This is used at line 09 to determine whether the loop runs or not. If *kFlag* is 0 the loop is skipped and execution of the code goes straight to *endin* which is the end of the instrument. If *kFlag* is 1 then the loop is executed. The opcode *tablekt* reads values from *iAmpTable* at control rate and outputs the variable *kAmp*. This value is then sent back to *Studio::Update()* using *chnset* at line 16. At line 15 *sprintfk* sets the output channel names assigning one channel per analysis bin. The values are then processed in *Studio::Update()*.

```
00 double lowFreqVals = 0.0f;
01
02 // fft frequency bin amplitude values from Csound
03 for(int i = 0; i < NUM_FFT_BINS * 0.66; i++)
04 {
05     lowFreqVals += *m_pFftAmpBinOut[i];
06 }
07
08 m_dLowFreqAvg = lowFreqVals / (NUM_FFT_BINS * 0.66);
09
10 // smooth lowFreqVals data stream
11 if(m_dLowFreqAvg > 0)
12 {
13     float currentLowFreqVal = static_cast<float>(m_dLowFreqAvg);
14     float lerpFraction = 0.8f;
15     m_fInterpolatedLowFreqVal = currentLowFreqVal + lerpFraction *
(m_fPrevLowFreqVal - currentLowFreqVal);
16     m_fPrevLowFreqVal = currentLowFreqVal;
17 }
```

Example A.11. FFT values processed in the update loop.

In Ex. A.11 above, the *for loop* at line 03 iterates through the first two thirds of the FFT bins and adds all the amplitude values together in the float *lowFreqVal*. The higher frequency bins were omitted in order to visually emphasise lower frequency sounds. The reasoning was that lower frequency sounds create larger vibrations so there was a desire to reflect this in the visual movement. At line 08 the value across the lower two thirds of the bins is averaged. It was then felt that the values needed further smoothing between frames. From lines 11 to 17 an interpolation between the previous average value and the current average value was implemented. The interpolated value *m_fInterpolatedLowFreqVal* is then sent as a uniform to the fragment shader.

Appendix B: *Ag Fás Ar Ais Arís* Code

This appendix contains some of the code used in the development of *Ag Fás Ar Ais Arís*. A brief description of the code functionality is given with each code snippet.

B.1 Audio Processing

i) In the file *agFasArAisAris.csd*, the *ClickPopStatic* instrument is triggered automatically by the *ClickPopStaticTrigger*. The *ClickPopStaticTrigger* generates score events and sets some parameters that are then used by the *ClickPopStatic* instrument.

```
0 kFileSpeed =      5.0
1 kGrainDurFactor   =      90.02
3 kGaussVal         gauss         6.0
4 kGaussVal2 gauss         100
5
6 seed 0
7 kRand             random 2.2, 10.8
8 kRand2            random 3, 20
9 kRand3          random      1, 0.1
```

Example B.1. Parameters controlled by ClickPopStaticTrigger.

The *ClickPopStaticTrigger* sets values for *kFileSpeed* and *kGrainDurFactor* at lines 0 and 1 in Ex. B.1. These variables eventually affect the sample position and the grain size respectively. The *gauss*[128] opcode is used to generate a gaussian distribution of values centred around 0.0. These are *kGaussVal* which is in the range -6.0 to 6.0, and *kGaussVal2* which is in the range -100.0 to 100.0. The *random*[129] opcode is also used to generate values with a pseudo-random distribution. These are stored in *kRand*, which is in the range 2.2 to 10.8,*kRand2*, which is in the range 3 to 20 and *kRand3*, which is in the range 1 to 0.1. The *seed*[130] opcode on line 6 means that *random* will use the system clock as a seed which allows it to output different values each time it runs. These various random control variables are used to introduce a sense of unpredictability when triggering the *ClickPopStatic* instrument.

[128] https://csound.com/docs/manual/gauss.html (accessed 26/03/2022).
[129] https://csound.com/docs/manual//random.html (accessed 26/03/2022).
[130] https://csound.com/docs/manual/seed.html (accessed 26/03/2022).

```
00 kMetVal0       metro       0.33, 0.00000001
01 kTrigRateVal   samphold    kRand3, kMetVal0
02
03 kMetVal        metro       kTrigRateVal, 0.00000001
04 kTrigVal       samphold    kRand, kMetVal
05 kTrigger         metro        kTrigVal
06
07
08 kMinTim          = 0
09 kMaxNum          = 5
10 kInsNum          = 6
11 kWhen            = 0
12 gkDur            = kRand2
13 kSpeed           = kFileSpeed + kGaussVal
14 kGrainFreq       = p4 + kGaussVal
15 kGrainDurFactor= kGrainDurFactor + kGaussVal2
16 kCentCalc = kGrainFreq + kGaussVal
17 kPosRand  = 100 + kGaussVal
18 kCentRand = kCentCalc + kGaussVal
19 kPanCalc         = 1
20 kDist            = 0.7
21
22 schedkwhenkTrigger, kMinTim, kMaxNum, kInsNum, kWhen, gkDur, kSpeed,
kGrainFreq, kGrainDurFactor, kCentCalc, kPosRand, kCentRand, kPanCalc,
kDist
```

Example B.2. ClickPopStaticTrigger main body.

As shown in Ex. B.2, at line 00, *kMetVal0* is used as a constant trigger that fires every three seconds. Each time it sends out a 1, *samphold*[131], at line 01, takes the current value of *kRand3* and sends it repeatedly out to *kTrigRateVal*. This value is used on line 03 to set the rate at which *metro*[132] sends out a 1 to *kMetVal*. At line 04, *samphold* uses *kMetVal* as the gate variable. Each time *kMetVal* is 1, the current value of *kRand* is sent to *kTrigVal*. This value is then used as the frequency parameter for *metro* on line 05. *kTrigger* is then used as the first parameter for *schedkwhen*. This process has the effect of randomising the rate at which *schedkwhen* generates new score events.

The variables *kTrigger, kMinTim, kMaxNum, kInsNum, kWhen* and *gkDur* are used as the standard parameters for *schedkwhen*. They define the event trigger, minimum time between events, maximum number of simultaneous instances of the instrument, instrument number, start time of the triggered event and duration of the event respectively. These are the parameters that control the meta-

[131] https://csound.com/docs/manual/samphold.html (accessed 26/03/2022).
[132] https://www.csounds.com/manual/html/metro.html (accessed 26/03/2022).

characteristics of each *ClickPopStatic* note. The parameters listed after these are all treated by the *ClickPopStatic* instrument as *p-fields* would be treated in a normal score file.[133] That means that they are specific to the *ClickPopStatic* instrument and are used to process data within it. These parameters correspond to *p-fields* 4 to 11 in the *ClickPopStatic* instrument. The various randomisation variables are used between lines 07 and 13 to try to ensure that the *ClickPopStatic* textures do not become too static. In Ex. B.3, you can see how the parameters are passed as *p* numbers within *ClickPopStatic*.

```
00 /*score parameters*/
01 iSpeed          = p4         ; 1 = original speed
02 iGrainFreq= p5         ; grain rate
03 iGrainDurFactor= p6          ; grain size in ms
04 iCent           = p7         ; transposition in cent
05 iPosRand        = p8         ; time position randomness (offset) of the
pointer in ms
06 iCentRand = p9         ; transposition randomness in cents
07 iPan            = p10        ; panning narrow (0) to wide (1)
08 iDist           = p11        ; grain distribution (0=periodic,
1=scattered)
```

Example B.3. Parameters passed as p-fields.

Of note here is that the parameters in the *ClickPopStaticTrigger* are all *k-rate*. They are then interpreted as *i-rate* by the *ClickPopStatic* instrument as they do not change over the course of the triggered event. They change between events. The *ClickPopStatic* instrument uses the variables sent from the *ClickPopStaticTrigger* in various processes to calculate the final parameters that are sent to *partikkel*[134].

The instrument uses *24cellRow.wav* as the source waveform. This is loaded into a function table using the GEN01[135] routine. It is then referenced through the global name *giFile*. This is passed to *partikkel* through the *kWaveForm1, kWaveForm2, kWaveForm3* and *kWaveForm4* variables. The parameter *iWaveAmpTab* is set to the default -1. This gives an equal mix of all four waveforms. There is some random deviation applied to the position of each sample followed by transposition of each source wave. As shown in Ex. B.4, the first wave form is not transposed. The second waveform is read at half speed, the third waveform is read at 1.32 times the original speed and the fourth waveform is read at 0.66 times the original speed. At line 05, the maximum amount of grains is set to a value of 3000.

[133] http://write.flossmanuals.net/csound/methods-of-writing-csound-scores/ (accessed 10/07/2020).
[134] https://csound.com/docs/manual/partikkel.html (accessed 26/03/2022).
[135] https://csound.com/docs/manual/GEN01.html (accessed 26/03/2022).

```
00 kWaveKey1 = 1
01 kWaveKey2 = 0.5
02 kWaveKey3 = 1.32
03 kWaveKey4 = 0.66
04
05 iMaxGrains       = 3000
```

Example B.4. Wave pitch adjustment and maximum grain definition.

In Ex. B.5, instead of sending *kGrainDurFactor* directly to *partikkel*, it is used to modulate *iGrainFreq*. The resulting value is sent to *partikkel* as *kDuration*. This is interpreted as the grain duration in milliseconds.

```
00 kDuration = (0.5 / iGrainFreq) * iGrainDurFactor
```

Example B.5. Using parameters to affect duration.

It was found that using *iGrainDurFactor* directly as the grain duration gave a deeper texture. By using it as a modulating factor for *iGrainFreq* the desired pointillistic texture was found.

```
00 giWin            ftgen  0, 0, 4096, 20, 6, 1
01 giAttack         ftgen  0, 0, 513, 5, 0.0001, 512, 1
02 giDecay          ftgen  0, 0, 513, 5, 1, 512, 0.0001
03
04 /*grain envelope*/
05 kEnv2Amt         = 0.5
06 iEnv2Tab         = giWin
07 iEnvAttack= giAttack
08 iEnvDecay = giDecay
09 kSustainAmount   = 0.8
10 kA_D_Ratio= 0.75
```

Example B.6. Envelope generators controlling grain shape.

The code snippet in Ex. B.6 shows the use of envelope generators to control the grain shape. Each grain can be shaped using envelopes consisting of 'independently specified attack and decay' sections with a 'sustain portion in the middle' (Brandtsegg, Saue and Johansen 2011: 40). This means each grain will have an attack-sustain-decay (ASD) profile. *iEnvAttack* and *iEnvDecay* are both assigned envelopes represented by *giAttack* and *giDecay*. These tables are generated using *ftgen*[136] and the *GEN05*[137] routine. This routine creates functions using exponential curves. *giAttack* starts at 0.0001 and ends at 1 whereas *giDecay* begins at 1 and curves down to 0.0001. The parameter *kSustainAmount* is fixed at 0.8. This means the sustain section of the envelope will be 80% of the grain duration. The

[136] https://csound.com/docs/manual/ftgen.html (accessed 26/03/2022).
[137] https://csound.com/docs/manual/GEN05.html (accessed 26/03/2022).

remaining 20% is split between the attack and decay sections. The parameter *kA_D_Ratio* is fixed at 0.75. This means that the attack section is given 75% and the decay section will be given 25% of the remaining duration. This represents an attack section of 15% and a decay section of 5% of the overall duration.

giWin is assigned to *iEnv2Tab* at line 06. This is an envelope generated by *ftgen* and the *GEN20*[138] routine at line 00. It is in the form of a Gaussian window function[139]. *iEnv2Tab* is a secondary envelope that is applied to the grain after the main ASD envelope described above. The *kEnv2Amt* parameter on line 05 dictates how much of the secondary envelope is used. It is set to 0.5 here, which means that the resulting envelope will be interpolated between a square window and the Gaussian window specified in *iEnv2Tab*. The secondary envelope is used here to try to fine-tune the main ASD envelope. The signal output from *partikkel* is finally multiplied with an ADSR envelope before being sent to a global output channel to be used by the *SpectralAnalysis* and *SoundLocaliser* instruments.

ii) The *ModalSynth* instrument is fully autonomous and runs for six seconds when the scene is launched.. The *wgbow* opcode excites a bank of *mode* filters. The values used to generate the *wgbow* and *mode* signals are calculated through a process of randomisation and Gaussian distribution. The use of Gaussian distribution was intended as a way to exert a type of loose control over the random value ranges.

```
0 kRangeMin  gauss  kGaussRange
1 kRangeMin += 65
2 kRangeMax  gauss  kGaussRange
3 kRangeMax += 80
4 kcpsMin    gauss  kGaussRange * 0.1
5 kcpsMin += 4
6 kcpsMax    gauss  kGaussRange * 0.1
7 kcpsMax += 6
8
9 kFreqScale rspline      kRangeMin,   kRangeMax,   kcpsMin,     kcpsMax
```

Example B.7. Randomisation of parameters for rspline and wgbow.

In Ex. B.7 the variable *kGaussRange* is set to 10. On line 0, the *gauss*[140] opcode takes *kGaussRange* as an input and outputs a *k-rate* (control rate) variable called *kRangeMin*. This value will be within the range *-kGaussRange* to *kGaussRange*, and is weighted by a gaussian distribution centred at0. The

[138] https://csound.com/docs/manual/GEN20.html (accessed 26/03/2022).
[139] http://www.csounds.com/manualOLPC/MiscWindows.html (accessed 11/07/2020).
[140] https://csound.com/docs/manual/gauss.html (accessed 28/03/2022).

value 65 is then added to *kRangeMin* which shifts the mean centre of distribution from 0 to 65. This means *kRangeMin* will be in the range of *65.0 - kGaussRange* to *65.0 + kGaussRange*. This procedure is then repeated with different mean values for *kRangeMax*, *kcpsMin* and *kcpsMax*. These values are then used as parameters for *rspline*[141] on line 9. This opcode creates a random cubic spline curve using points generated within the ranges given to it through its parameters. This process of continually introducing elements of randomisation was intended to give a dynamic character to the audio signal.

As shown in line 1 of Ex. B.8, the value of *kFreqScale* is assigned to *kFreq*, which is then used as a parameter for *wgbow* on line 7. Each of the five parameters for *wgbow* were calculated using the above system of randomisation.

```
0 kAmp = ampdbfs(kWgbowAmpVal)
1 kFreq = kFreqScale
2 kPres = kWgbowPressureVal
3 kRat = kWgbowPosVal
4 kVibF = kWgbowVibF
5 kVAmp = ampdbfs(kWgbowVibAmp)
6
7 aExc wgbow        kAmp, kFreq, kPres, kRat, kVibF, kVAmp
```

Example B.8. Wgbow parameters.

The *wgbow* parameters correspond to:

- amplitude of the produced note (*kAmp*)
- frequency of the note (*kFreq*)
- bow pressure on the string (*kPres*)
- position of the bow on the string (*kRat*)
- frequency of vibrato (*kVibF*)
- amplitude of the vibrato (*kVAmp*)

[141] https://csound.com/docs/manual/rspline.html (accessed 25/03/2022).

Ex. B.9 shows the bank of *mode* filters used in this instrument.

```
0 aRes1      mode   aExc, 100 + kGaussRange, 220 + kGaussRange
1 aRes2      mode   aExc, 142 + kGaussRange, 280 + kGaussRange
2 aRes3      mode   aExc, 211 + kGaussRange, 200 + kGaussRange
3 aRes4      mode   aExc, 247 + kGaussRange, 220 + kGaussRange
4 aRes5      mode   aExc, 467.9, 140 + kGaussRange * (472.7 * 0.1)
5 aRes6      mode   aExc, 935.8, 140
6
7 aRes sum   aRes1, aRes2, aRes3, aRes4, aRes5, aRes6
```

Example B.9. Bank of mode filters.

The audio rate variable *aExc* is used as an input signal to the bank of resonating filters. The resonating frequency of the first four filters is then modulated by *kGaussRange.* The quality factor for filters 1 to 5 is modulated again by *kGaussRange*. The resonant frequencies of the first four filters were chosen according to modal frequency ratios in Lindroth (2020). The ratio of frequencies between these filters correspond to a Douglas Fir wood plate. Two further filters were added on lines 4 and 5 to mix in some higher frequencies. On line 7, the signals are then summed to create the final output signal *aRes.*

iii) The *SoundLocaliser* instrument receives the values for each sound source using the *chnget*[142] opcode. The opcode *hrtfmove2*[143] takes an audio signal, an azimuth value in degrees and an elevation value in degrees. It then uses head-related-transfer-function (HRTF) filters to calculate the direction of the sound source in the 3D environment. As with *Obj_#3*, the data files containing HRTF measurements are based on the MIT database.[144]

```
00 kPortTime        linseg 0.0, 0.001, 0.05
01
02 kDistVals[0]     portk  kDistances[0], kPortTime
03
04 aLeftSigs[0], aRightSigs[0]   hrtfmove2    aInstSigs[0], kAzimuths[0],
kElevation[0], "hrtf-48000-left.dat", "hrtf-48000-right.dat", 4, 9.0, 48000
05 kDistSquared2 pow kDistVals[2], 2
06 aLeftSigs[0]     =      aLeftSigs[0] / kDistSquared2
07 aRightSigs[0]    =      aRightSigs[0] / kDistSquared2
08
09 aL         sum    aLeftSigs[0], aLeftSigs[1], aLeftSigs[2]
10 aR         sum    aRightSigs[0], aRightSigs[1], aRightSigs[2]
11
12 outs             aL, aR
```

Example B.10. Source location instrument.

[142] https://csound.com/docs/manual/chnget.html (accessed 28/03/2022).
[143] http://csounds.com/manual/html/hrtfmove2.html (accessed 28/03/2022).
[144] https://sound.media.mit.edu/resources/KEMAR.html (accessed 12/07/2020).

At line 02 in Ex. B.10, the distance value received from *Studio()*, *kDistances[02]*, is first processed using *portk*. This applies a portamento to the incoming signal. This was necessary becauseI discovered I needed to apply this processing as the raw distance value was creating discontinuities in the sound. At lines 06 and 07 in, the left and right output signals are divided by the squared distance value. At lines 08 and 09 all audio signals are summed before being sent to the stereo output channel.

B.2. Visual Rendering

i) The simplest structural component in this piece is the sphere. The SDF is shown using GLSL code in Ex. B.11.

```
00 float sphereSDF(vec3 p, float radius)
01 {
02     return abs(length(p) - radius);
03 }
```

Example B.11. SphereSDF function.

This simply takes the point *p* and a float giving the *radius* as arguments. The length between *p* and the origin is given by *length(p)*. The *radius* value is then subtracted from this. The result of this calculation is positive if the point is outside the radius of the sphere, negative if the point is inside the radius of the sphere and 0 if the point is exactly on the surface. The *abs()* function returns the absolute value of the calculation which means that the sphere will be hollow and the camera can travel inside the structure. If *abs()* was not used the sphere would be a solid mass. This SDF is adapted from Quilez (2020a).

ii) The plane SDF, also adapted from Quilez (2020a) is shown in Ex. B.12. This is the code that creates the upper-plane and the floating fragments that seem to break off it.

```
00 #define ITERATIONS 10
01 #define FOLD_CUTOFF 50
02 #define SCALE 2.0
03 #define OFFSET 2.0
04
05 float planeSDF(vec3 p, vec4 normal)
06 {
07     int n = 0;
08     while(n < ITERATIONS)
09     {
10           if(p.x + p.y < 0.0) p.xy = -p.yx + (fbmVal_left * 0.25); //
fold 1
11           if(p.x + p.z < 0.0) p.xz = -p.zx + (fbmVal_left * 0.25); //
fold 2
12           if(p.y + p.z < 0.0) p.zy = -p.yz + (fbmVal_left * 0.25); //
fold 3
13
14           p = p * SCALE - OFFSET * (SCALE - 1.0);
15
16           if(length(p) > float(FOLD_CUTOFF)) break;
17
18           n++
19     }
20
21     return dot(p, normal.xyz) + normal.w;
22 }
```

Example B.12. PlaneSDF with folds.

The basic SDF calculation is shown at line 21. The dot product between *p* and the *xyz* components of *normal* is calculated. This will return 0 if *p* is perpendicular to *normal.xyz*. This would mean that *p* is on the plane. The *w* component of *normal* determines the distance of the plane from the origin.

The floating visual artefacts that emerge from the plane are caused by a folding technique that is used to displace *p* before line 21. The *while* loop between lines 08 and 19 takes *p* and tests its location relative to the x, y and z planes passing through the origin. The point is then reflected through the relevant plane. At line 14, *p* is scaled and offset before being tested for breaking the fold cutoff limit at line 16. This is a limit introduced to optimise the real-time rendering as without it, the system slowed down.

iii) The SDF shown in Ex. B.13 is adapted from Christensen (2011a) and is used to generate the foreground fractal structure.

```
00
01 #define SCALE 2.0
02 #define OFFSET 2.0
03 #define FOLD_CUTOFF 50
04
05 float kifSDF(vec3 p)
06 {
07     mat3 rot = rotationMatrix(vec3(0.5, 1.0, 0.0), fractalAngle + (fbmVal
* 0.25));
08     orbit = vec4(10.0, 10.0, 10.0, 1.0);
09     int n = 0;
10     while(n < iterVal)
11     {
12           p = rot * p;
13
14           if(p.x + p.y < 0.0) p.xy = -p.yx; // fold 1
15           if(p.x + p.z < 0.0) p.xz = -p.zx; // fold 2
16           if(p.y + p.z < 0.0) p.zy = -p.yz; // fold 3
17
18           p = rot * p;
19           p = p * scaleVal - offsetVal * (scaleVal - 1.0);
20
21           orbit = min(orbit, vec4(abs(p), 1.0));
22
23           if(length(p) > float(FOLD_CUTOFF)) break;
24           n++;
25     }
26     return length(p) * pow(scaleVal * (1.0 + (granRainAmp * 0.25)), -
float(n));
27 }
```

Example B.13. KIFS SDF function adapted from Christensen (2011a).

At line 07 a rotation matrix is added that is continuously updated according to *fractalAngle* which is sent as a uniform from *Studio()*. The *fractalAngle* value is also modulated by *fbmVal* which is an fBm noise value. The variable *fractalAngle* is controlled by the AV-participant through the right-hand neural network. The resulting matrix, *rot*, is multiplied by *p* at lines 12 and 18. This causes the movement of the fractal structure. These rotations are placed before and after the folding operations at lines 14 to 16. These are the same folding operations used above in the *planeSDF*. The point *p* is then scaled and offset at line 19 to create the Sierpinski structure. There is a cutoff test at line 23 to break out of the *while* loop. Finally the value returned at line 27 estimates the distance from the surface as a product of the length of *p* multiplied by the *scaleVal* factor raised to the negative power of *n*, which is the number of iterations *p* went through. This value is also modulated by *granRainAmp* which is the spectral amplitude of the signal output by the *GranulatedRain* instrument.

At line 08 a *vec4* is assigned to the global vector *orbit*. At line 21 *orbit* is compared to *vec4(abs(p))* and the value 1.0. The component with the smallest value is then assigned to *orbit*. The vector *orbit* is then used later in the processing to create surface patterns on the structures.

iv) After the distance to the nearest surface has been estimated, colouring and shading is applied. The same colouring procedure is applied across the scene. The material colour is generated according to a series of smoothing calculations, colour mixtures and use of the orbit value calculated earlier.

```
00
01 #define SCALE 2.0
02
03 float sq = float(iterVal) * float(iterVal);
04 float smootherVal = float(index) + log(log(sq)) / log(SCALE) -
log(log(dot(pos, pos))) / log(SCALE);
05 vec3 matCol1 = vec3(pow(redVal, log(smootherVal)), pow(greenVal,
log(smootherVal)), pow(blueVal, log(smootherVal)));
06vec3 matCol2 = vec3(pow(0.333 * (modSamp_rmsOut * 80.0), 1.0 /
log(smootherVal)), pow(0.487 * (modSamp_rmsOut * 80.0), 1.0 /
log(smootherVal)), pow(0.184 * (modSamp_rmsOut * 80.0), 1.0 /
log(smootherVal)));
07 totMatCol = mix(matCol1, matCol2, clamp(6.0 * orbit.x, 0.0, 1.0));
```

Example B.14. Colour smoothing adapted from Christensen (2011c).

At line 04 in Ex. B.14, *smootherVal* is calculated according to a formula adapted from Christensen (2011).The components of the formula include *index*, which is the number of iterations the raymarching algorithm went through before exiting, *iterVal* which is the number of recursive iterations used to render the Sierpinski fractal, *SCALE* which is the scaling value used in the construction of the Sierpinski fractal and *pos* which is the point of intersection of the ray and the surface.

At line 05, *matCol1* creates the lines and veins running through the scene. The variables *redVal*, *greenVal* and *blueVal* are controlled by the AV-participant through the left-hand neural network. At line 06, *matCol2* creates the green/grey colours. The variable *modSamp_rmsOut* is mapped from the RMS value of the *ModalSampler* instrument. The vector *matCol1* raises the RGB components to the power of the log of *smootherVal* whereas *matCol2* raises them to the power of the inverse log of *smootherVal*. These values were found through experimenting with different combinations of *pow()*, *log()* and *smootherVal*.

Once the material colour is determined, the normal at the point of intersection is calculated. As shown in Ex. B.15, this can be found using a finite difference calculation.

```
00 #define EPSILON 0.02
01
02 vec3 norm(vec3 pos)
03 {
04     vec3 xDir = vec3(EPSILON, 0.0, 0.0);
05     vec3 yDir = vec3(0.0, EPSILON, 0.0);
06     vec3 zDir = vec3(0.0, 0.0, EPSILON);
07
08     return normalize(vec3( DE(pos + xDir) - DE(pos - xDir),
09                            DE(pos + yDir) - DE(pos - yDir),
10                            DE(pos + zDir) - DE(pos - zDir)));
11 }
```

Example B.15. Normal calculation adapted from Christensen (2011c).

This function is adapted from Christensen (2011). The technique is also described in Quilez (2015). It works by sampling points around *pos* and computing the gradient of that surface. According to Quilez, the gradient can be used as the normal because it is perpendicular to the surface.

v) Ex. B.16 shows the calculation relating to the global illumination setup.

```
00 #define SUN_DIR vec3(0.5, 0.8, 0.0)
01
02 float ambOcc = ao(pos, norm, 0.5, 5.0);
03 float sun = clamp(dot(norm, SUN_DIR), 0.0, 1.0);
04 float sky = clamp(0.5 + 0.5 * norm.y, 0.0, 1.0);
05 float ind = clamp(dot(norm, normalize(SUN_DIR * vec3(-1.0, -1.0, 0.0))),
0.0, 1.0);
06
07 vec3 lightRig = sun * vec3(1.64, 1.27, 0.99);
08 lightRig += sky * vec3(0.32, 0.4, 0.56) * ambOcc;
09 lightRig += ind * vec3(redVal + (1.0 * modSamp_rmsOut * 100.0), greenVal
+ (1.0 * modSamp_rmsOut * 100.0), blueVal + (1.0 * modSamp_rmsOut * 100.0))
* ambOcc;
10
11 colour = totMatCol * lightRig;
```

Example B.16. Lighting rig based on Quilez (2013).

At line 02, *ambOcc* is calculated and multiplied by the *sky* and *ind* lighting components on lines 08 and 09 to create some simple ambient occlusion effects. At line 03, *sun* is calculated using the dot product of the surface normal and the direction vector of the sun. This is the directional component of the sun light-source. At line 04, *sky* is calculated based on the *y* component of the surface normal.

The sky light-source simply points down across the whole scene. Finally,*ind* is calculated according to the inverse direction of the *sun* vector. This light is a reflection of the direct sunlight off the surface. Between lines 07 and 09 these components are then multiplied by colour vectors to balance the light sources throughout the scene. The indirect light-source is modulated by the *modSamp_rmsOut* value which is mapped from the RMS power of the *ModalSampler* instrument. The individual lighting components are summed and stored in *lightRig*. At line 11 this value is multiplied by the material colour, *totMatCol*, to give the final *colour* value.