Depth Modulation: Composing motion in immersive audiovisual spaces

EWA TRĘBACZ

Center for Digital Arts and Experimental Media (DXARTS), Box 353414, University of Washington, Seattle, WA 98195-3680, USA
E-mail: trebacze@uw.edu, trebacz.ewa@gmail.com

The field of electroacoustic music has witnessed years of extensive exploration of aural spatial perception and an abundance of spatialisation techniques. Today the growing ubiquity of visual 3D technologies gives artists a similar opportunity in the realm of visual music. With the use of stereoscopic video we now have the ability to compose individual depth cues independently. The process of continuous change of the perceived depth of the audiovisual space over time is being referred to as *depth modulation*, and can only be fully appreciated through motion.

What can be achieved through the separation and manipulation of visual and sonic spatial cues? What can we learn about the way we perceive space if the basic components building our understanding of the surrounding environment are artificially split and re-arranged?

Visual music appears to be a perfect field for such experimentation. Strata of visual and aural depth cues can be used to create audiovisual counterpoints in three-dimensional spaces. The choice of abstract imagery and the lack of obvious narrative storylines allow us to focus our perception on the evolution of the immersive audiovisual space itself. A new language of an immersive audiovisual medium should emerge as a delicate, ever-changing balance between all previously separated and altered components.

1. INTRODUCTION

The field of electroacoustic music has witnessed many years of extensive exploration of aural spatial perception. Research based on the separation and manipulation of monaural and binaural cues resulted in an abundance of sound spatialisation techniques and approaches. Today the growing ubiquity of visual 3D technologies gives artists a similar opportunity in the realm of visual music: with the use of stereoscopic video we now have the ability to compose individual depth cues independently. This ability to manipulate depth cues independently opens new areas of artistic exploration.

Visual music appears to be a perfect field for such experimentation. Strata of depth cues can be used to create audiovisual counterpoints in three-dimensional spaces. Using abstract imagery enables artists to focus their creative attention on the evolution of the audiovisual spaces themselves. This is still largely unexplored territory, which gives artists the opportunity to pose some fundamental questions regarding the nature of human perception of space and its boundaries. More importantly, such questions would be posed directly through the language of art, while at the same time enriching and transforming the traditional vocabulary of visual music.

I will approach these questions through an experiment in art, introducing the term *depth modulation* to describe the process of continuous change of the perceived depth of audiovisual space *over time*.

My own work *Errai*, which premiered at the 52nd International Festival of Contemporary Music 'Warsaw Autumn' (2009), represents an early effort to explore this rich territory. It was realised with the use of stereoscopic video and ambisonic sound, combined with live music performance and computer-controlled light animation. *Errai* will be used to give examples of how depth modulation can be applied in practice, with a particular focus on the visual realm. However, the scope of this article is much broader than just the applications of these techniques in my own work, and will hopefully open a discussion about the potential for the creative use of depth cues as building blocks for relations and structures in the audiovisual space.

2. SENSORY OVERLOAD OR THE FABRIC OF ART OF THE FUTURE?

The late nineteenth century synaesthetical concept of arts initiated a process of blending the boundaries between traditional art disciplines. As a result of an emphasis on the holistic concept of perception, the strict divisions between time-based and space-based arts began to dissolve, with this process continuing and intensifying through the entire twentieth century. The phenomenological philosopher Maurice Merleau-Ponty claims that all visible objects are palpable objects at the same time, while *tactile* and *visible* are the two demonstrations of the same being (quoted in de la Motte-Haber 2002: 32). If a tactile object is simply an embodiment of a visible object, should *space* be understood as a sequence of tangible objects? Visual and tactile spaces show a mutual relationship. Also the visible and the audible appear to be strongly interconnected. In experimental works of *musique*

Organised Sound **17**(2): 156–162 © Cambridge University Press, 2012. doi:10.1017/S1355771812000088

https://doi.org/10.1017/S1355771812000088 Published online by Cambridge University Press

concrète (in the 1940s and 1950s) sound is treated as an object that can be characterised in a way resembling descriptions so far reserved for the visible, such as having a volume or density.

The concept of the intersensory nature of human sensory perception inspired artists to cross boundaries between the traditional arts, which resulted in a variety of hybrid forms. Similar tendencies have been present in parallel in the visual arts and avant-garde music. The most obvious manifestation of this process is found in installations, in which sculptural elements are expanded by sound. The simultaneous character of a visual experience and the sequential character of aural experience create together a new concept of space, understood as a network of temporal relations. Physical (architectural) spaces are often enriched by sound or video projection. In the late twentieth century and the first decade of the twenty-first century architectural spaces were often modified and disconnected from their utilitarian functions by video projection. Also, experiments in virtual and augmented reality attempt to blend the frontier between the real and the illusionary.

The years since the late 1990s have witnessed a real explosion in various 3D technologies, both visual and sonic, and the true re-birth of stereoscopy as a legitimate approach to filmmaking. Digital cinema successfully brought large audiences back to theatres, reviving the interest in three-dimensional movies for the first time since 1956. Yet, despite the large commercial success of selected films, 3D cinema is still awaiting the creation of its own language, a language that would truly express the uniqueness of this medium. Without such a language, 3D films will remain reduced to mere entertainment with their sporadic commercial successes based on overwhelming the senses rather than producing works that would equal the masterpieces of the traditional cinematic art. Experimental art, cautiously but steadily approaching 3D technologies, is in danger of following a similar fate. If interest in such techniques is not met by the emergence of a unique language, capable of expressing the nature of immersive media, the field may continue to languish despite the ready availability of the underlying technologies.

The phenomenon of sensory overload has become so obvious that we seem to have overlooked the moment when it seemingly turned into our everyday experience. Not only are we providing our senses with amounts and combinations of signals that are reaching the physical limits of our perception, but, with the rapid development of new technologies, we are now able to extend our eyes and ears beyond the audible and the visible. As we approach the physical boundaries of human perception, the question of the role of sensory experience in perceiving a work of art returns with a vengeance. If the physical limitations of our senses are the final frontier, then further development in art may appear only as a possibility rather than something that is granted.

This inevitably raises a very well-known question about the relationship between the direct experience of art and the work of art itself. Is direct sensory experience necessary to comprehend a work of art as a whole? Can a work of art function at all if its fundamental aspects are inaccessible through direct experience? Is intellectual compensation for our imperfect senses sufficient for an artwork to be perceived as intended by an artist? Finally – is it possible that the further development of art might be bound to the alteration of human perceptual skills?

These and many similar questions have inspired me to pursue projects in the realm of audiovisual time-based art, focusing on the intersensory perception of space. I have been particularly interested in the following problems:

- What can we learn about the way we perceive space if the basic components building our understanding of the surrounding environment are artificially split and rearranged?
- What can be achieved through the separation and manipulation of visual and sonic spatial cues?
- How do we perceive phenomena not available through direct experience in the real world that can be artificially created, for instance a continuous change of image depth over time?
- In such an artificial environment, how would we perceive the relationship between various media and how would we interpret the information perceived by various senses? Would we be able to seamlessly switch between concurrent lines of movement?

Before we begin to examine these intriguing problems, it will be necessary to investigate the role of individual depth cues and relations among them in the perception of audiovisual spaces. What follows is a brief overview of visual depth cues, and a discussion of so-called *direct depth modulation*.

3. DIRECT DEPTH MODULATION AND ITS PERCEIVED EFFECTS

If *depth modulation* is defined as a process of continuous change of the perceived depth of the audiovisual space over time, then *direct* depth modulation is such a change using binocular or binaural cues only. Direct depth modulation is a phenomenon unknown to real-world experience. It is however available to us in the visual realm through the direct modification of visual cues using stereoscopic video, and in my opinion this process shows great potential for artistic experimentation.

Perceiving space is one of the essential skills guiding animals and humans through the environments

https://doi.org/10.1017/S1355771812000088 Published online by Cambridge University Press

they live in. Humans develop a full set of spatial perceptual skills by the first year of life through an extensive learning process. It is worth noting that in human infants the full ability to perceive spatial relations develops much earlier than the full set of motor skills. Information about depth and distances is strongly related to information about motion, and is gathered through combinations of numerous sources.

Humans visually perceive three-dimensional space by interpreting the two-dimensional images in each eye using a network of visual cues. A long list of these cues can be found in the literature and is usually divided into two categories: *monocular* cues and *binocular* cues. Monocular depth cues can be *kinetic* or *pictorial*. A classic example of a kinetic cue is motion parallax: as we move, objects that are located further away seem to be 'moving' slower relative to closer objects. Examples of pictorial cues include the relative or familiar size of objects, shadows, elevation, orientation, occlusion, texture gradient, colour, linear perspective and atmospheric perspective.

Binocular depth cues include disparity (binocular parallax) and convergence. Disparity indicates that the retina of each eye records a slightly different image. Convergence is related to an ability of the eyes to turn towards each other in order to zoom and focus on objects and preserve binocular vision (Solso 1994).

Computer graphics software and modern immersive technologies have finally given us the freedom to compose depth cues independently in various types of virtual environments. Using 3D animation software or specially rigged stereo cameras we can now directly manipulate binocular depth cues. Direct depth modulation is accomplished by changing the distance between the stereoscopic cameras over time. As the distance between the cameras changes, so do the depth cues associated with binocular vision.

At any given time the human left and right eye converge on a point in space, known as the point of fixation. This convergence, and the angle it creates, is used to determine distances. The difference between angles of fixation between different objects determines the relative depths of these objects within a stereoscopic image.

Changes in the distance between the two virtual cameras change the depth cues that rely on stereopsis, the sense of relative distance based on binocular disparity (Gillam 1995: 41). As the cameras move towards each other, the perceived depth of an image flattens as the relative difference between angles created by points of fixation grow smaller. While other depth cues, such as those for perspective, may remain, those associated with binocular vision are lost. When the cameras occupy the same space, the left and right stereo images become identical, in effect creating a normal 2D image. Further, the cameras can exchange positions by passing through each other, which results in transforming the image from stereoscopic, to flat, to pseudo-stereoscopic.

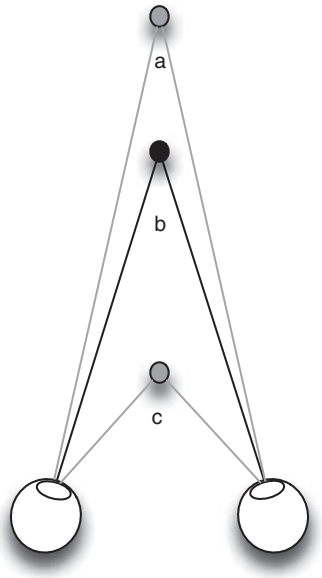


Figure 1. Angular disparity and perceived depth of image.

Figure 1 shows how, as the distance changes, so does the angle. This angle determines the relative depth of an object. The greater the angle, the closer the object appears.

Figure 2 shows how changing separation of the virtual cameras can change this angle without changing the actual distance of the point of fixation. This creates the illusion of a change in depth. Angles a and c from above can be achieved simply by changing the distance between cameras.

3.1. Perceived depth

Direct depth modulation has two interesting effects on how one perceives the image. The first and most obvious is how it changes the perceived distance of objects from the viewer. As the cameras move closer together, stereoscopic depth cues fade and objects appear to flatten towards the depth of the physical screen. Objects in the background are drawn forward, objects in the foreground move back, and objects with perceivable depth get flatter.

Figures 3a–c illustrate a fairly simple scenario, with three unmoving objects. In these figures we see a top-down view of a three-dimensional space showing the perceived position of these objects in space along with the positions of the cameras. Objects A and B are considered to be two-dimensional objects with no depth cues of their own; object C has depth.

Because the cameras in this example are parallel, rather than convergent, there is no predetermined point of fixation. The viewer determines the actual

https://doi.org/10.1017/S1355771812000088 Published online by Cambridge University Press

point of fixation based on the object he or she is tracking. Without other depth cues, the physical screen becomes a natural point of fixation for the viewer. Therefore as objects lose their depth cues they appear to converge towards the screen.

Figure 3a shows the cameras at their designated starting distance, with object A in front, object B at the same depth as the physical screen, and object C starting at the depth of the physical screen and projecting into the background.

Figure 3b illustrates the effects of moving the cameras towards each other. As the cameras move towards each other, it appears as if object A is approaching closer to the screen line and as if object C is losing depth. Because object B is already at screen depth there is no perceived change. When the cameras are coincident, there are no binocular depth cues.

As the cameras exchange positions by passing through each other, with the left camera on the right side and the right camera on the left, the binocular depth cues are reversed, creating a pseudo-stereoscopic image. In the example given in Figure 3c, as the cameras move away from each other, object A appears to move to the background and object C appears to project towards the viewer.

It should be noted that this does not change perspective cues nor does it change how the objects are ordered. Objects should be carefully chosen to avoid both perspective problems and problems where an object is occluded by another object.

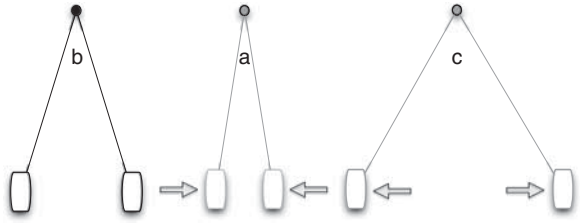


Figure 2. Angular disparity and separation between virtual cameras.

3.2. Perceived size

Along with changes in perceived depth, depth modulation also influences the perceived size of objects in the image. This effect is created by the clash between two depth cues, binocular disparity and perspective.

Figure 4a shows the real-world conditions: as an object moves towards the viewer it fills a greater portion of the field of view, and as it moves away from the viewer it fills less of the field of view. This is an effect known as *looming* (Gillam 1995: 26). The angles represent the field of view commanded by the object as it changes depth.

When objects move towards and away from the viewer, perspective and binocular disparity are properly synchronised, matching our real-world expectations. However when only depth modulation is applied with no changes to the objects themselves, only binocular disparity is affected. Perspective is unchanged. Even while the object appears to be

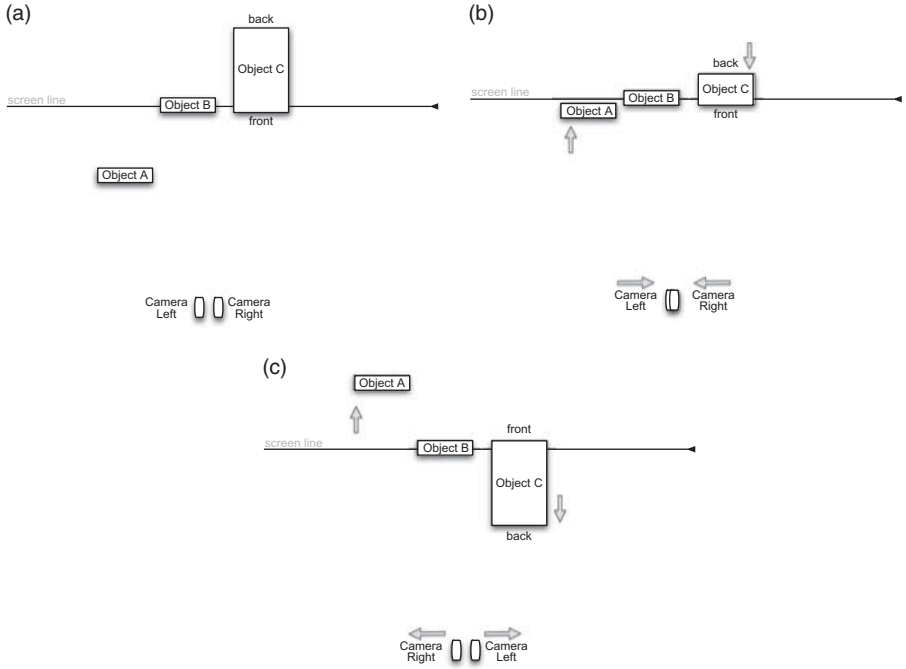


Figure 3. (a) Virtual camera movement and perceived depth: example A (static distance). (b) Virtual camera movement and perceived depth: example B. (c) Virtual camera movement and perceived depth: example C.

https://doi.org/10.1017/S1355771812000088 Published online by Cambridge University Press

moving backward or forward in space, there is no looming.

Figure 4b shows the field of view taken up by the object. Without looming the amount of space the object takes up in our field of view, represented by the angle subtended by the edges of the object, is unchanged.

4. STRATA OF DEPTH CUES

Direct depth modulation is only the most obvious form of depth modulation. In my preliminary experiments with stereoscopic projection, my major focus was on creating a particular impression of pulsation of the depth dimension. The illusion seemed quite powerful and showed a lot of yet unexplored potential. Later I noticed that this effect was even more intriguing when direct manipulation of depth was independent from monocular cues, usually associated with depth perception. For instance, subtle changes of focus, colour or contrast could be used to provide conflicting cues between the video layers. Similarly, selected monocular cues can be used in synchronisation with the direct depth modulation to reinforce the illusion. One can also imagine situations in which the entire audiovisual space is evolving in one direction, but the timing of change varies between individual monocular and binocular depth cues.

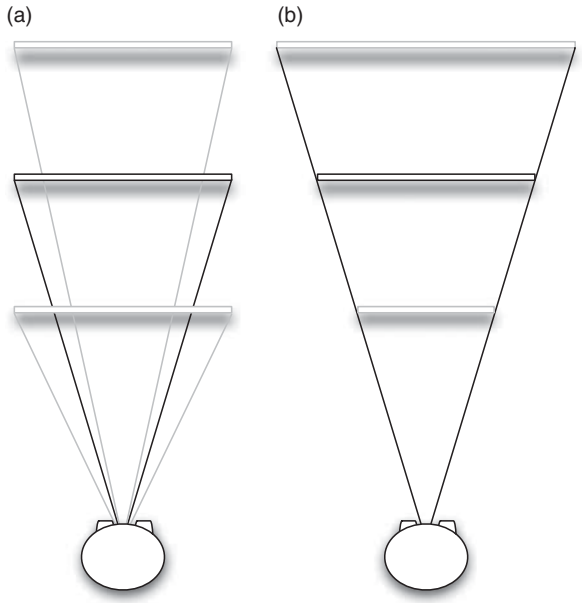


Figure 4. (a) Distance changes and perceived size. (b) Direct depth modulation and perceived size.

These experiments with stereoscopic video projection, as well as my ongoing interest in sound spatialisation, led me in 2009 to a larger work, in which I attempted to treat an immersive audiovisual space as a single whole. *Errai* (2009) combined computer-animated stereoscopic video and ambisonic sound with live computer-controlled animated lighting and live music performance. During my work on *Errai* I gradually realised that depth modulation can be approached on several simultaneous layers, and these layers can be organised in hierarchical way, which I later called *strata of depth cues*. I will describe each of the layers or strata and provide brief examples based on my own work. However, the concept of strata of depth cues is much broader than its implementation in *Errai*.

To better demonstrate the nearly unlimited potential for creating powerful illusions and audiovisual counterpoints by composing layers of independent depth cues, I propose the representation shown in figure 5. Complex relations can be created between elements of each layer and between the layers themselves.

4.1. Direct depth modulation

The first layer represents the ability to modulate binocular and binaural depth cues only. This is the most direct example of depth modulation: we are modulating the differences between each eye and each ear.

STRATA OF DEPTH CUES
visual music on the Z-axis

Direct depth modulation: binocular and binaural depth cues
Image: disparity and convergence
Sound: ITD, ILD

Indirect depth modulation – monaural and monocular depth cues
Image: perspective, shadows, texture gradient, contrast, occlusion, etc.
Sound: reverberation, impression of loudness, amount of high frequencies, etc.

Sound and image: *audiovisual spatial counterpoints*

Competing spaces: modulating presence in *virtual* and *physical* spaces

Figure 5. Strata of depth cues.

https://doi.org/10.1017/S1355771812000088 Published online by Cambridge University Press

Using an example from my own work, the stereoscopic video in *Errai* explores a full spectrum of continuous change of perceived depth over time. This was achieved by directly modulating distances between pairs of virtual cameras in a 3D animation environment with several independently evolving video layers. As a result, these layers evolve simultaneously and independently from each other: from true stereoscopic – through flat 2D – to pseudo-stereoscopic images. At times the only perceivable movement is an expansion or contraction of the visual space itself.

Partial expansion or contraction of the perceivable depth occurs on various layers throughout the entire work, but a full cycle with over a 2-minute-long transition from a pseudo- to a true stereoscopic image is presented only once, becoming the extended climax of the work. While this transition fully reveals the strongest formative principle of *Errai*, it is at the same time the most obvious demonstration of how relative our perception can be. The exact time when one becomes aware of the change happening in the depth dimension will vary greatly between individual viewers. It also depends on the decision of which parts of the image a particular viewer would decide to track at any given moment. Such a choice has significant consequences: some elements of the image that are neglected while being in the background, suddenly and surprisingly become relevant after their position was shifted towards the foreground. Since we cannot observe similar processes in the real world, it is also somewhat difficult to recognise the exact moment when depth starts contracting or expanding. As observed during several tests, it will vary not only between viewers but also between different experiences of the work by the same viewer.

In case of the audio component of *Errai*, the situation is a little different. Unlike the video, the initial audio material was not generated by computer software. Instead it was recorded in ambisonics with the simultaneous use of two Soundfield microphones placed at different elevations and distances from the sources. The recording sites were carefully chosen for their unique acoustic properties, with the assumption that their spatial qualities will play a key role in defining not only the character of the musical narration, but also ultimately the formal shape of the work. This material was later processed, spatially manipulated in B-format, and composed into a sequence of soundscapes. As a result, I did not independently compose binaural depth cues, but rather attempted to create continuous transitions between three-dimensional sound images, captured simultaneously by these two microphones. While there is some analogy to direct depth modulation in the video layers, this process may not exactly fit its strict definition, and calls for further investigations.

4.2. Indirect depth modulation

The second layer represents monocular and monaural depth cues. For instance, binocular cues could indicate that the space is expanding, while monocular cues would suggest contraction. Alternatively, a combination of monocular and binocular cues can be used to create a stronger illusion of a continuous depth change over time.

In *Errai* I often used selected monocular depth cues to reinforce the effects of direct depth modulation. The visual environment of *Errai* was constructed in such a way that the entire image could be easily flipped, choosing imagery with semi-transparent and luminescent objects in order to avoid problems with occlusion. Background could become foreground and vice versa, with continuous transitions possible between those two states. More importantly – during such a transition or shift, different elements of the image or different movements within a selected layer are exposed, allowing the viewers different interpretations of the stereoscopic video. Selected elements of the image (such as objects or movement) gradually become revealed during the transition, as the visual layers change their perceived depth. These elements can be seen as more relevant or dominant depending on the stage of the depth modulation process. These changes of depth are at times related to subtle changes of colour, contrast and movement on each layer. Combined with direct depth modulation, this results in changing of hierarchy between layers: elements gain and lose importance as depth evolves.

The use of monaural depth cues is very obvious in the audio layer. In *Errai*, single sound objects are 'placed' in space at various distances from the listener, becoming components of larger entities, but rarely changing their location independently. These three-dimensional sound images are then spatially manipulated as a whole. However, the perceived distance of sound objects from the listener is a very important element of the structure of this work, and the differences between these locations are at times quite extreme. This is achieved by the direct application of monaural depth cues, such as amount of reverberation, amount of high frequencies in the sound signal and its perceived loudness.

4.3. Sound and image

On the top of that we can design relationships between three-dimensional image and three-dimensional sound. This can be achieved by creating and resolving tension using conflicting depth cues between the sound and image, and by creating contrapuntal relations between parallel layers of movement.

The audio and video layers of *Errai* were for the most part created independently, but they are not separate entities. They share the same common pace

https://doi.org/10.1017/S1355771812000088 Published online by Cambridge University Press

and oscillate around the same formal plan. No strict, note-to-note synchronisation is required for the audio and video layers, but they do meet within each of the ten sections of the piece.

The video and audio layers can be considered as two groups of independent voices, constantly competing for the audience's attention. This relationship is most transparent while comparing climax points of audio and video layers. The effective climax of the video layer is a long transition from the pseudo-stereoscopic to the true stereoscopic image. This is immediately followed by a section that may be considered a climax of the audio layer, thus shifting the audience's attention between the visual and aural spaces.

4.4. Competing spaces

The last layer represents the relationship between the pre-recorded or artificially created audiovisual spaces, reproduced during the performance, and a physical space, in which the work is presented. The physical space is understood here as an architectural space with all its unique properties – like reverberation time, size and shape, lighting – and those properties can be referenced and enhanced by adding an element of live interaction, such as live music performance or lighting.

Errai for example used both live lighting (designed by Polish artist Robert Sowa) and live music performance (soprano and horn). The lighting in *Errai* creates a contrapuntal relationship to the stereoscopic video, becoming an equally important element of the dramaturgy. It is used to construct an immersive three-dimensional space, with continuity between the virtual world of the stereoscopic video projection, and the real world, represented by the physical space, in which the audience was enclosed.

The most important layer of the lighting design scheme is a direct interaction with the stereoscopic projection. Several short interventions of direct light were combined with the stereoscopic video by the use of masked animated lights pointed directly at the projection screen. This interaction constitutes one of the major formative principles organising the perceptual space of the piece. It was given the role as a bridge between the virtual world perceived through the 3D glasses and the semi-architectural space, created anew by the lighting of the physical space.

Similarly, the live audio performance (horn, soprano) creates a reference to the physical space of the performance hall, which is now considered to be the body of one extended musical instrument. Representing 'here and now', the live performance creates a dialogue with the pre-recorded soundscapes and the acoustics of the hall itself.

5. CONCLUSIONS

Earlier in this article, I posed several questions that inspired me to pursue projects in the realm of audio-visual immersive art. I introduced the term depth modulation, and attempted to show how this process can be applied in practice, giving examples from my work *Errai*. While *Errai* was not intended to be a scientific experiment, it has led me to several interesting observations regarding the visual perception of depth and the relationship between the various spatial components of the work. For example, following and comprehending a continuous change of depth over time seems to be a challenging task. Since we are not originally equipped to perceive such phenomena, it seemed like the perception of depth modulation had a discrete character. It was very difficult to follow the entire process without any interruptions or distractions. After perceiving the change at several time intervals, viewers were often able to realise that the process was continuous. But even after understanding this, the actual perception of this phenomenon remained stubbornly discrete.

While the examples I have given here may not have directly answered or exhausted any of the questions I posed, I hope they can open a broader discussion of this subject. With the development of new immersive technologies, new questions will inevitably appear and again challenge our understanding of human spatial perception. I strongly believe that the nature of human perception of space can and should be investigated through art. The potential for artistic exploration is nearly limitless. Hopefully investigations into human special perception will be continued not only by researchers, but also by artists, through their innovative work.

The realm of visual music, where individual depth cues can be used to compose motion in depth, seems to be a very promising area for such exploration. What we are ultimately looking for is a new language, not only enriching the existing vocabulary of visual music, but also capable of expressing the unique nature of immersive media. Such a new language should emerge as a delicate, ever-changing balance between all previously separated and altered components.

REFERENCES

Gillam, B. 1995. The Perception of Spatial Layout from Static Optical Information. In W. Epstein and S.J. Rogers (eds) *Perception of Space and Motion*. San Diego, CA: Academic Press, 23–67.

Motte-Haber, H. de la. 2002. Esthetic Perception in New Artistic Contexts: Aspects – hypotheses – unfinished thoughts. In B. Schulz (ed.) *Resonanzen/Resonances: Aspekte der Klangkunst/Aspects of Sound Art*. Heidelberg: Kehrer Verlag, 29–37.

Solso, R.L. 1994. *Cognition and the Visual Arts*. Cambridge, MA: The MIT Press.

https://doi.org/10.1017/S1355771812000088 Published online by Cambridge University Press