Images as spatial sound maps

Etienne Deleflie
School of Creative Arts. Sonic Arts Research Network.
Faculty of Creative Arts
University of Wollongong
ed386@uow.edu.au

Greg Schiemer
School of Creative Arts. Sonic Arts Research Network.
Faculty of Creative Arts
University of Wollongong
schiemer@uow.edu.au

ABSTRACT

The tools for spatial composition typically model just a small subset of the spatial audio cues known to researchers. As composers explore this medium it has become evident that the nature of spatial sound perception is complex. Yet interfaces for spatial composition are often simplistic and the end results can be disappointing. This paper presents an interface that is designed to liberate the composer from thinking of spatialised sound as points in space. Instead, visual images are used to define sound in terms of shape, size and location. Images can be sequenced into video, thereby creating rich and complex temporal soundscapes. The interface offers both the ability to craft soundscapes and also compose their evolution in time.

Keywords

Spatial audio, surround sound, ambisonics, granular synthesis, decorrelation, diffusion.

1. INTRODUCTION

Composers working in spatial composition are often limited by the paucity of the tools available. Many of these tools tend to be focused on point sources of sound in space. Ongoing research highlights that the perception of spatial audio includes a range of cues [1, 2] many of which are not activated by the directional encoding of point sources. The physical shape of a sound-emitting object, its size, the direction in which it is pointing, its interaction with the physical environment and the current atmospheric conditions are all factors that contribute to the spatial character of sound [3, 4, 5].

The complexity of modeling these cues partly explains the limited availability of tools, but also highlights the challenge of designing appropriate interfaces.

1.1 Apparent sound source extent

One of the more important spatial audio cues is the perception of apparent sound source extent [2]. This implies our ability to perceive the physical size of a sound-emitting object. Examples of real-world scenes where this cue would come into play include wind blowing through trees, waves breaking on sand, a large group of people chatting, thunder, heavy traffic or the distant ambience of a city.

Various techniques have been proposed to synthesize apparent sound source extent [6, 7, 8, 9, 10] but to date, none of these have had much consideration within the context of interface design.

2. THE SOUND MAP

The technique presented here uses images as spatial sound maps. Consider an image, scaled up and laid flat like a map, with the listener standing in the centre, as shown in Figure 1.

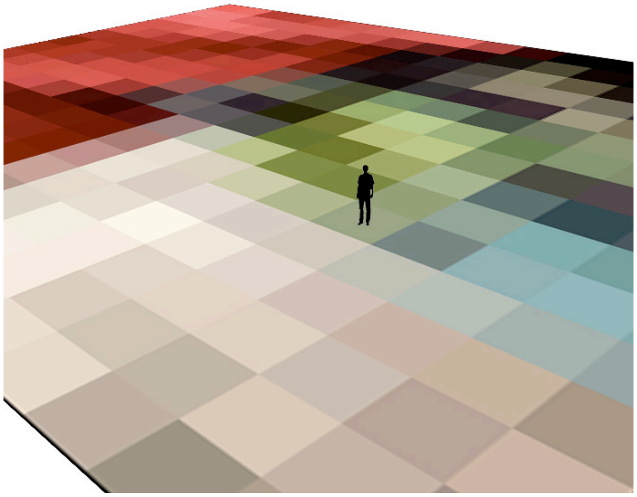


Figure 1. An image can represent a spatial sonic map. The image is laid flat and greatly scaled up revealing its pixelated nature. The listener stands in the centre.

Each pixel represents a single independent sound source. The colour of the pixel represents its sonic properties such as level, pitch and timbre. Each sound source is spatially encoded with reference to the listener.

The challenge in this representation is the re-assembly of atomized sound events into a larger sound whose physical size can be perceived.

2.1 Synthesising apparent sound source extent

The perception of sound source size relies on complex mechanisms one of which can be expressed in general terms as the 'fine dissimilarity between a sound at the two ears' [11]. This dissimilarity is quantified as the Inter Aural Cross Correlation (IACC) coefficient [2], where a lower IACC results in the increased perception of sound source size.

For a group of pixels, of arbitrary shape, to be perceived as a single, wide, sound-emitting object their virtual-speakers must emit the same audio content but collectively have a low IACC. This can be achieved by decorrelating the output of each speaker.

Decorrelation happens to be one of the bi-products of granular synthesis [12]. Therefore if granular synthesis is used to sonify each pixel, there is the opportunity for groups of pixels to be perceived as a single, wide sound source.

The use of granular synthesis to imbue the perception of physical size through sound has been explored in the context of diffusion environments. Diffusion is a form of audio spatialisation where typically stereo sources are projected over multiple speakers placed amongst the audience [13]. Barry Truax writes:

> *"The volume, or perceived magnitude, of a sound depends on its spectral richness, duration, and the presence of unsynchronised temporal components, such as those produced by the acoustic choral effect and reverberation. Electroacoustic techniques expand the range of methods by which the volume of a sound may be shaped. Granular time-stretching is perhaps the single most effective approach, as it contributes to all three of the variables just described. It prolongs the sound in time and overlays several unsynchronised streams of simultaneous grains derived from the source such that prominent spectral components are enhanced. It should be noted that delays of only a few milliseconds are sufficient to decorrelate the different grains streams and thus increase their sense of volume."* [13]

These ideas can be similarly applied to ambisonics, the spatialisation technology used in this paper. Whilst diffusion and ambisonics are very different techniques, there is an interesting parallel between Truax's use of decorrelation and what is attempted here.

In our implementation each pixel represents a virtual speaker as shown in figure 2. This effectively models a sound diffusion environment implemented with a very large speaker array. At the same time it can be accurately implemented with a 3rd order ambisonics signal decoded over as little as 8 speakers.

This approach potentially offers diffusionists a means of diffusing sound through a very large number of speakers without the need to physically provide them. There are a number of advantages, one being that virtual speakers may be positioned outside the physical limits of an auditorium

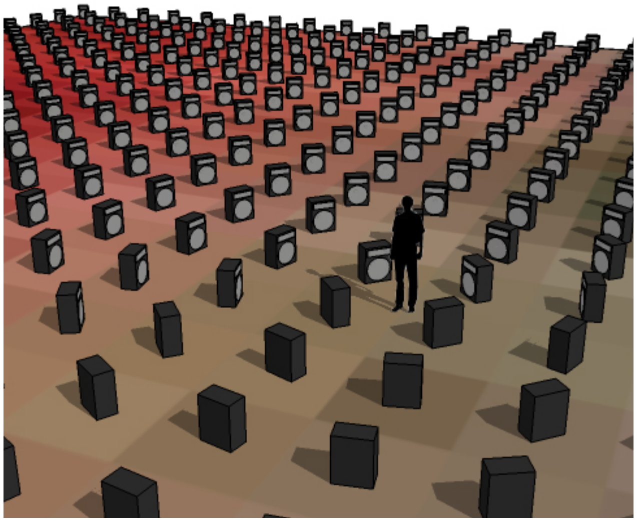


Figure 2. Diagram shows how each pixel could be represented as an independent virtual speaker in a multi-channel diffusion environment. In this illustration, 900 speakers are displayed.

2.2 Spatial encoding of virtual speakers

Amongst the different spatial audio encoding technologies commonly used, ambisonics has proven its ability to faithfully recreate recorded soundfields [3]. However, Ambisonic encoding only caters for the rendering of each virtual speakers' position relative to the central listener.

Distance cues need to integrated outside of ambisonic encoding. Here the diffuse field (reverberation), first reflections, loss of energy due to sound radiation and the effect of air absorption, where far sounds are 'muffled', have all been modeled.

The speed of sound is also modelled prior to ambisonic encoding. A time delay is applied to each virtual speaker's output. The amount of delay applied models how long a sound emitted from that speaker would take to reach the listener. Sounds generated at the extremity of the image (the pixels in the corner are 424m away from the central listening position) will take over 1 second to reach the listener. In other words, a single image might contain 2 identical dots (one in the centre and one at the edge) but when translated to spatial audio, the 2 dots will be heard with a 1.25 second delay between them.

The Doppler effect, where movement causes apparent pitch changes in sounds, is not modelled. Whilst this is a strong psychoacoustic cue, its implementation has been deferred in the interest of exploring less common spatial cues.

There is no modeling of room character such as reverberation.

3. SYNTHESIS FROM GRAPHIC NOTATION

Using graphic-notational images as a compositional tool is not new. Xenakis' UPIC system allowed a composer to draw structures to be sonified. The sonification of images has also been explored for such things as vision substitution applications for the blind [14].

Granular Synthesis as well as Formant Synthesis are processing techniques often targeted by experimental musical interfaces [15]. How the interface's output is mapped or translated to granular parameters will vary depending on the interface. Video analysis algorithms, whether they implement motion tracking or other, have been used to map vectoral information into granular parameters [16]. Similarly to ideas presented here, granular parameters have been driven *directly* from pixel values thereby foregoing the computational cost of any intermediary analysis of video frames [17]. In this paper, a further step has been taken in that a parallel has been drawn between the space represented on an image, and sonic space. This parallel forms the crux of the interface presented. The image *is* the spatial sound map.

If one considers each image as a simple matrix of values used to drive granular processes, then this comes very close to a technique used by Xenakis to organize the macro-structure of granular clouds [18]. A system of three dimensional 'screens' allowed him to define frequency, amplitude and time parameters for controlling clouds of grains. A sequence of 'screens' was put together into a 'book' to introduce a temporal structure [19]. In this work, a 'screen' can be seen as synonymous with an image and in the same way a 'book' can be seen as synonymous with a video.

4. SOUND FROM PIXELS

Granular synthesis has been chosen as the main audio synthesis technique because of its ability to decorrelate multiple streams of sound. This enables the perception of sound source size in groups of like-coloured pixels. What remains to be decided is how the pixel's colour components are mapped to granular parameters.

Given that each pixel has around 16 million colour combinations, the composer has access to a broad range of auditory parameters.

4.1 Source image size and mapped dimension.

Each image processed is 60 x 60 pixels in size producing a total of 3600 granular voices.

For the purpose of locating each virtual speaker in space, each pixel is deemed to be 10m in size. In other words the 3600 granular voices are spaced 10m apart covering an area of 600 square metres. With the listener's reference position set in the centre of the image, this means that the most distant sound, represented by a pixel in the corner of the image, is 424m away.

4.2 Granular Synthesis parameters

Granular synthesis has been used extensively by different composers working in different ways. Here, granular synthesis is used in such a way that sound grains are extruded and overlapped similarly to time stretching techniques.

4.2.1 Fixed granular parameters

Granulation can occur over any chosen source audio file. Each pixel represents one granular voice, or one stream. All granular voices for all pixels will use the same source audio file. Grain size and grain overlap (or density) will be constant and consistent.

For each granular voice, grains of 24msec duration will be triggered every 4msec. This represents a granular density of 250 grains per second. The chosen image size of 3600 pixels results in a final grain density of 9,000,000 (250 x 3600) grains per second.

4.2.2 Map-able granular parameters

The granular parameters which can be driven by pixel values are grain position, grain volume and grain speed (or pitch).

4.3 Pixel-to-granular-parameter maps

4.3.1 Pixel Colour Spaces

The auditory space is represented by JPG images where each pixel is a combination of three values between 0 and 255. Common image formats typically use the RGB (Red Green Blue) colour space. RGB is a format designed for graphics display hardware and therefore has little bearing on how we perceive light. However, RGB values can easily be translated into alternative colour spaces. Both the HSL (Hue Saturation Lightness) and HSV (Hue Saturation and Value) colour spaces are more representative of human perception and are therefore perhaps more appropriate for driving granular parameters.

Whilst HSL and HSV attempt similar perception-centric representations of colour, their implementations are slightly different. Figure 4 shows that component 'L' is perhaps more representative of pixel brightness than component 'V' and may therefore be a good mapping to sound loudness.

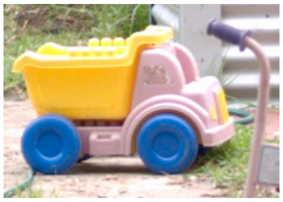


Figure 3. A coloured image.

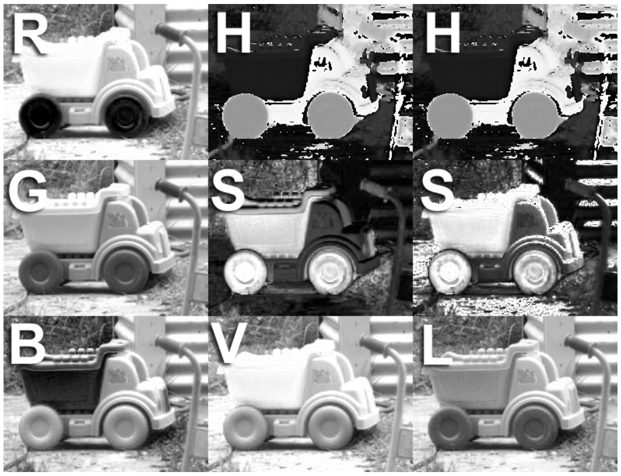


Figure 4. The image shown in Figure 3 is represented by 3 different colour spaces. The first column shows the image decomposed into RGB component channels, the second shows HSV component channels, and the third shows HSL component channels.

4.3.2 Maps

Table 1. details some of the mapping strategies explored. Different components from the various colour spaces have been used to drive different granular parameters. In some cases not all of the image information is used.

The most intuitive link between colour and sound was found to be mapping pixel lightness to grain volume. Bright colours are louder. General conclusions are discussed below (see Evalutation).

Table 1. Four pixel-to-GS parameter maps.

Map	GS parameter	Colour component
1	Grain position	HSV Hue
	Volume of grain	HSL Lightness
	Pitch	-
2	Grain position	HSV Saturation
	Volume of grain	HSL Lightness
	Pitch	-
3	Grain position	HSV Hue
	Volume of grain	HSL Lightness
	Pitch	HSV Saturation
4	Grain position	HSV Lightness
	Volume of grain	HSL Saturation
	Pitch	-

5. IMPLEMENTATION

The implementation was done in SuperCollider. SuperCollider offers a sophisticated Object Oriented DSP environment tightly bound to Graphical User Interface widgets. SuperCollider also has strong support for granular synthesis techniques.

Due to the high volume of granular processing required to spatialise images (9,000,000 grains per second) real time rendering is not possible on current desktop computers. On a Macbook Pro Core2duo, the rendering of one second of audio (using 60x60 pixel images) takes around a minute. Again, SuperCollider caters well for the non-real-time scenario through its Score objects.

A GUI developed in SuperCollider consists of the image spatialiser and an ambisonic player as shown in Figure 5. However, this represents only the last mile of the compositional process since image and video editing software is also used to author and edit the sequence of images spatialised.

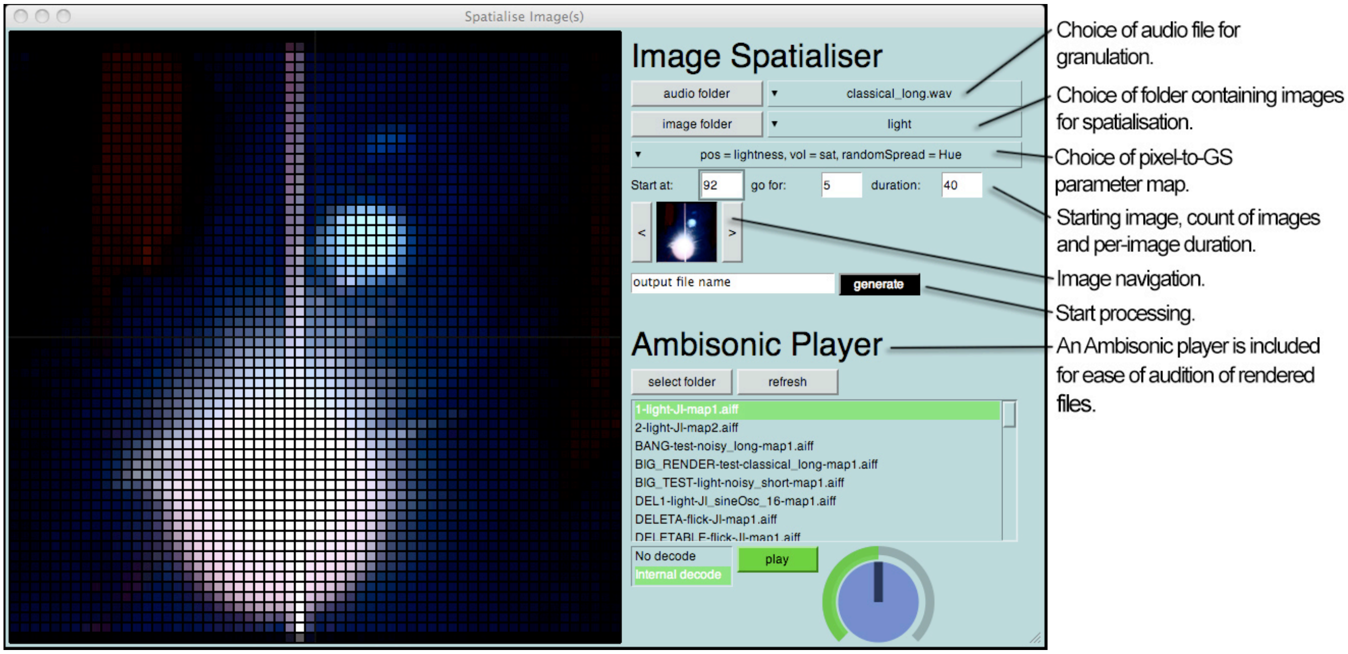
6. EVALUATION

A number of tests were conducted to evaluate the resultant spatial and sonic interest of the different mapping strategies. These tests used different source audio files and different image sequences.

The spatialisation in Figure 5 shows a video of a torch moving around. The resultant audio can be heard in 5.1 surround here: http://soundofspace.com/a/62. The same sound can be heard mixed down to stereo but accompanying its source video here: http://vimeo.com/9202956.

It was quickly apparent that the most significant characteristic of the spatial soundscapes produced was caused by the modeling of the speed of sound. Since pixels are spaced 10m apart the farthest pixels in any image are around 420m away from a centrally located listener. This distance represents a time-of-arrival delay of around 1.25 seconds.

Figure 6 illustrates how one shape, representing a sounding object 300m wide, will produce sound where the farthest parts of



Spatialise Image(s)

Image Spatialiser

audio folder: classical_long.wav

image folder: light

pos = lightness, vol = sat, randomSpread = Hue

Start at: 92 go for: 5 duration: 40

output file name

Ambisonic Player

1-light-JI-map1.aiff
2-light-JI-map2.aiff
BANG-test-noisy_long-map1.aiff
BIG_RENDER-test-classical_long-map1.aiff
BIG_TEST-light-noisy_short-map1.aiff
DEL1-light-JI_sineOsc_16-map1.aiff
DELETA-flick-JI-map1.aiff
DELETABLE-flick-JI-map1.aiff

No decode
Internal decode

Choice of audio file for granulation.

Choice of folder containing images for spatialisation.

Choice of pixel-to-GS parameter map.

Starting image, count of images and per-image duration.

Image navigation.

Start processing.

An Ambisonic player is included for ease of audition of rendered files.

Figure 5. Screenshot of the SuperCollider GUI showing a 60x60 pixel image, and spatial rendering parameters.

the sound will arrive at the listener considerably later than the closer parts. This resembles reverberation.

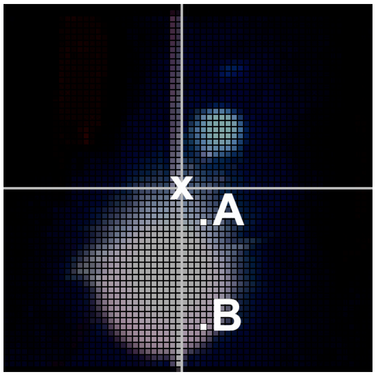


Figure 6. Image shows a lens flare, from a digital camera, used as a sonic map. Due to the speed of sound, the listener (at x) will hear sound emitted from pixel B 500msec after sound emitted from pixel A. This causes a sort of unidirectional reverberation effect.

However, unlike room-modelled reverberation, where both first reflections and diffuse reflections typically arrive from every direction, this reverberation is somewhat unidirectional. There are no real reflections, all sounds arrive directly from the object. Sounds emitted from a far part of the object will arrive much later. The spatialisation of this image can be heard at http://soundofspace.com/a/61.

This unidirectional reverberation might be a psychoacoustic cue related to the perception of size in extremely large sounding objects.

The image in Figure 7, showing a sequence of discs 100m apart, was designed to confirm the accuracy of the spatial acoustic modeling. It highlights three spatial audio characteristics: the speed of sound; each disc will be heard separately with a small (150msec) time gap between them; direction: each sound will move incrementally to the right of the listener; and atmospheric sound absorption: each disc will lose volume sequentially in the higher frequencies. The spatialisation of this image can be heard at http://soundofspace.com/a/60.

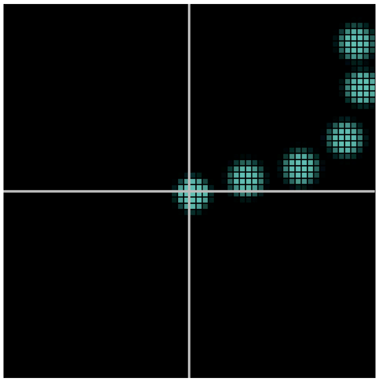


Figure 7. A synthetic image designed to confirm the spatial acoustic modelling of the interface developed.

Figure 8 shows an image designed to test one pixel-to-granular-parameter mapping strategy; pixel hue is mapped to the grain position in the source audio file. Each colour will activate the playing of different grains. The spatialisation of this image can be heard at http://soundofspace.com/a/59.

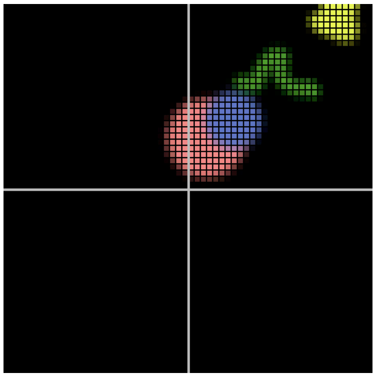


Figure 8. A synthetic image designed to confirm that hue has been accurately mapped to granular position.

7. CONCLUSION

In this paper, an interface for spatial audio composition has been presented. The interface allows composers to work with spatial audio cues usually not modelled in software environments.

The technique offers some interesting insights into the challenges of building interfaces for spatial audio composition. It demonstrates that images can be used to quickly and easily author location and size. It also highlights that a spatial audio interface must somehow negotiate the effects produced by the speed of sound. Sounds triggered simultaneously will not necessarily be heard simultaneously. Should the spatial composer think of sound in terms of when it is triggered, or when it is heard?

It is too early to say whether this realistically-modeled time delay is an inhibitor to the spatial composer or a benefit in that it forces the composer to think of sound in space in real terms. Perhaps designing such spatial sound artifacts into compositional structure holds the key to a spatial audio aesthetic.

The perception of sound source size is achieved through the decorrelation of multiple sound streams. However, the predominant characteristic of a large sounding object is the resultant unidirectional reverberation produced as a result of the same sound being emitted from different distances.

As our understanding of spatial audio cues deepens and software tools evolve, the spatial composer's bag of tricks will expand. This expansion will need be equally met with efforts to develop appropriate interfaces that both accurately model spatial cues but also liberate the composer to find expression for their ideas.

REFERENCES

[1] A.S. Bregman, *Auditory scene analysis : the perceptual organization of sound*, Cambridge, Mass.: MIT Press, 1990.

[2] J. Blauert, *Spatial hearing : the psychophysics of human sound localization*, Cambridge, Mass.: MIT Press, 1997.

[3] D.G. Malham. "Approaches to spatialisation," in *Organised Sound*, vol. 3, no. 02, pp. 167-177, 1998.

[4] T. Myatt. "Sound in space," in *Organised Sound*, vol. 3, no. 02, pp. 91-92, 1998.

[5] D. Worral. "Space in sound: sound of space," in *Organised Sound*, vol. 3, no. 02, pp. 93-99, 1998.

[6] D. Malham, "Spherical Harmonic Coding of Sound Objects - the Ambisonic 'O' Format," *AES 19th International Conference*, 2001, pp. 54-57.

[7] D. Menzies, "W-panning and O-format, tools for object spatialisation.," *AES 22nd International Conference*, 2002.

[8] G. Potard and I. Burnett, "Decorrelation Techniques for the Rendering of Apparent Sound Source Width in 3D Audio Displays," *Conference on Digital Audio Effects (DAFx'04)*, 2004.

[9] D. Menzies and M. Al-Akaidi. "Ambisonic Synthesis of Complex Sources," in *Audio Engineering Society*, vol. 55, no. 10, pp. 864-876, 2007.

[10] E. Deleflie and G. Schiemer, "Spatial Grains: Imbuing Granular Particles With Spatial-Domain Information," *ACMC09, Improvise, The Australasian Computer Music Conference*, 2009.

[11] D. Cabrera. "The Size of Sound: Auditor Volume Reassessed," in *MikroPolyphonie 5. Online journal*, vol. 5, 1999.

[12] C. Rolfe and D. Keller, "Decorrelation as a By-Product of Granular Synthesis," *The XIII Colloquium on Musical Informatics*, 2000.

[13] B. Truax. "Composition and diffusion: space in sound in space," in *Organised Sound*, vol. 3, no. 02, pp. 141-146, 1998.

[14] G. Bologna, B. Deville, T. Pun and M. Vinckenbosch. "Transforming 3D coloured pixels into musical instrument notes for vision substitution applications," in *Image Video Process.*, no. 2, pp. 8-8, 2007.

[15] V. Maniatakos and C. Jacquemin, "Towards an affective gesture interface for expressive music performance casapaganini.org," *Proc. of the 8th Conf. on New Interfaces for Musical Expression (NIME)*, 2008.

[16] C. Kiefer, N. Collins and G. Fitzpatrick, "Phalanger: Controlling Music Software With Hand Movement Using A Computer Vision and Machine Learning Approach," *Proc. of the 9th Conf. on New Interfaces for Musical Expression (NIME)*, 2009.

[17] J. Jakovich and K. Beilharz, "ParticleTecture: interactive granular soundspaces for architectural design," *Proc. of the 7th Conf. on New Interfaces for Musical Expression (NIME)*, 2007.

[18] I. Xenakis, *Formalized music : thought and mathematics in composition*, Stuyvesant, NY: Pendragon Press, 1992.

[19] C. Roads. "Xenakis and granular synthesis," in *Quaderni*, 2005.