THE PERCEPTION, EVALUATION AND CREATIVE APPLICATION OF HIGH ORDER AMBISONICS IN CONTEMPORARY MUSIC PRACTICE

Ircam Composer in Research Report 2012
Natasha Barrett
Composer
nlb@natashabarrett.org

ABSTRACT

The aims of this research residency were to test and evaluate the practical, perceptual and compositional use of higher order ambisonics. The work explored three main technical areas: very high order synthesised ambisonics sound-fields in 2D up to 12th order and 3D up to 9th order, near-field compensated higher order ambisonics and higher-order recorded sound-fields using the EigenMike. The following musical areas were investigated: distance information, enhanced spatial precision, detailed recorded sound-fields to capture accurate real-world spatial relationships and the control of sound points, motions and masses in relation to both timbre and 'extra-musical' meaning. The work was carried out in IRCAM's Studio-1 installed with a 24-loudspeaker hemisphere, and in the Espace de Projection, installed with a 75-loudspeaker hemisphere. Ambsionics synthesis was carried out using existing and specially implemented methods in IRCAM's Spatialisateur. The residency period was for three months, carried out part-time over seven months. A concert work called *Hidden Values* was also composed.

The residency was funded by IRCAM, Oslo Cultural Grant for International collaboration and the Norwegian Cultural Council. *Hidden Values* was composed with support from the Norwegian Composers' Fund.

1. INTRODUCTION

The music research residency 'The Perception, Evaluation and Creative Application of High Order Ambisonics in Contemporary Music Practice' explored technical and artistic realms. Ambisonics theory is already well published and will not be repeated here (Daniel et al. [1], [2]). Previously, my compositional work with ambisonics dating back to 2000 can be summarised as exploring the following (Barrett [3], [4], [5]):

- HOA up to 5th order: exploring the advantages of increased spatial resolution and specifically addressing spatial counterpoint within a compositional language.
- Creative solutions using existing tools: combining commercial and non-commercial software solutions.
- Hybrid solutions to the problem of creating and controlling vertical information under variable concert conditions: a layered vertical distribution of horizontal information.
- Hybrid solutions addressing the control of sound points, sound masses and audience coverage: combining HOA with first-order Soundfield recordings and mixed decoding solutions.
- The pros and cons of encoded (fixed) spatial information and traditional sound diffusion performance practice.
- Hyper-real enhancements of spatial information: to articulate more clearly sound without visual causation.

To take these ideas further three assumptions were derived from published theory and three main areas of investigation were established:

(a) The higher the order of ambisonics, the greater the angular precision and the larger the audience listening area. It was first necessary to test the perceptual truth of this assumption, explore where the thresholds lie, and then decide whether the results are compositionally useful. In the Espace de Projection, the 75 loudspeaker hemisphere containing 26 speakers in the lower ring and the decoding solutions implemented in the Spatialisateur (Spat) software allowed a maximum of 9th order full 3D (100 channels) and 12th order 2D (25 channels). This defined the technical limits.

(b) Near-field compensated higher-order ambisonics (NFC-HOA) can, in theory create a focused source inside the loudspeaker array and stabilise source location outside the array over a large listening area. Creating focused sources inside the loudspeaker array may also make more accurate our perception of distance and depth. The published NFC-HOA methods were implemented in Spat and tested. With different sounds types, the effect on our distance perception derived from NFC-HOA compared to other aspects of sound such as identity, gesture, reverberation and environmental cues was evaluated.

(c) Higher order recorded scenes can reproduce more accurate real environments in terms of the relationship between elements of the spatial scene: An EigenMike, which contains 32 microphone capsules located on a sphere can capture sound-fields that can be encoded as 4th order ambisonics, then decoded with a standard 4th order decoder. Recordings made with the EigenMike were cross-tested with those of the Soundfield microphone (1st order microphone) in different indoor and outdoor spaces.

Sounds examples and compositions reside on the HOA-Espro computer and in the IRCAM archive.

2. TECHNICAL AND ARTISTIC RESULTS

The work was carried out with collaboration from the EAC team: Markus Noisternig, Thibaut Carpentier and Olivier Warusfel.

2.1 HOA 3D

In the Espro we can decode up to 9th order 3D using the mode matching decoder [6] implemented in Spat. However, the 9th order file contains 100 channels, and each minute of audio at a resolution of 24bits 44.1 Khz is 1.2GB in size (the WAV / AIFF file size limit is just over 4GB). So it is useful explore, using the decoding methods available, whether our perception can hear a difference between a 4th order (25 channel file), 7th order (64 channel file) and 9th order.

Before a fair evaluation, the decoder method needs to be decided. The settings should reproduce a faithful transfer of spatial information from composing space to concert space, optimising for the best representation of motion, fixed location and dynamic contrast as encoded in the source. The options explained in [7] are summarised as:

- Decoder method: projection, pseudoinverse, energy preserving.
- Decoder type: basic, inphase, maxre, basicmaxre, maxreinphase.
- Crossover: frequency point for dual-band decoding (max-r_E / max-r_V).

In 4th, 7th and 9th order, informal listening tests of sources containing complex motion, fixed points and general enveloping sounds over a wide spectral range, concluded that the following produced the most realistic reproduction over a suitable sweet spot (50% the width / depth of the room):

- Decoder method: energy preserving.
- Decoder type: max-r_E in phase.
- Crossover frequency 200 Hz.

Depending on the frequency content of the source material and the intended sense of envelopment, the crossover frequency could be increased up to 400 Hz despite published recommendations being of a lower value. In addition, Noisternig recommended a full spherical decoding where the first and second rings of virtual loudspeakers (theoretical loudspeakers below the equator) are mirrored up into the real hemisphere. This optimises the decoding and regains any energy lost in the virtual speakers.

2.1.1 HOA 3D Results

The difference between 7th and 9th order was unnoticeable in blind tests. The reasons for this are unclear, but may be due to room acoustics, less optimal decoding (loudspeakers not in locations for a perfect decoding), slight differences in the loudspeaker responses and the resolution of our hearing.

The difference between 4th and 7th order was more noticeable, not only in terms of angular precision but with a more 'open' or 'transparent' spatial envelopment. As 4th order was the decoding optimum in studio 1, cross tests between rooms were made easy. Working at 7th order also allows easy decoding for other concert spaces where only a 2D set-up is available, requiring only 16 loudspeakers.

2.2 HOA 2D

It was immediately clear in both studio 1 and the Espro that, although full 3D decoding appeared more transparent in space than 2D (i.e. with complete independence from the loudspeakers as physical objects making sound), there was an inherent problem of the sound appearing higher than its actual spatialised elevation. This is not surprising because the decoder places energy across vertical loudspeakers even for horizontally spatialised sounds. A 2D decoding removes this vertical energy and also creates a useful contrast when in combination with a 3D decoding of other source material.

2.2.1 HOA 2D results

5th, 7th and 12th order 2D decodings were tested. The perceptual difference between 5th and 7th order were clear. Between 7th and 12th the differences were subtle, but for complex scenes there was a marginal improvement in source separation.

2.3 Near-field encoding

As far we are aware, this is the first time (spring 2012) that NFC-HOA has been tested in a large space using a variety of sound types. The standard published methods for encoding NFC-HOA ([1] [2]) were implemented in Spat[1] by Carpentier. Even though we were aware of various problems with the current theory, it was useful to explore the extent of these problems in practical application:

- Higher encoding-decoding orders achieve more convincing NFC results. However, increasing the order increases the bass boost problem as the source approaches the centre of the listening area. High-pass compensation filters are then necessary to protect listener and equipment from the dangers of sudden bass increase, but in doing so we experience that the NFC effect was also removed.

[1] More recent work on NFC-HOA was also considered [8] but CPU constraints of real-time implementation made the method for now impractical. It may be advantageous to test the method in a non-real time application, but this was outside the scope of the team.

- Both NFC and the high-pass filters were found to be extremely CPU intensive, especially for moving sources.
- The loudspeaker distances for the decoding stage need to be specified in the encoding stage, which means encoding and decoding stages are no longer independent. As a work around a loudspeaker distance adaptor was implemented [2].

2.3.1 12th order NFC-HOA Results

- The loudspeaker distance adaptor worked within a variance of approximately 50-200%. In other words, if we encoded for a loudspeaker array of 4-meter radius, we can safely translate the results to a loudspeaker array of anything between two meters and eight meters in radius.
- Based on an 8-meter radius, placing the source approximately two meters inside the loudspeaker array produces the sensation of a focused source inside the array (without bass compensation filters). Closer than this we hear the start of the dramatic bass boost and extreme care should be taken.
- The real-time implementation was tested on a Mac Pro (quad core Intel Xeon 2.4 Mhz). Even for one moving source, with Spat spread across all four cores, CPU usage was at the maximum. To put this in perspective, the same computer could calculate an encoded and decoded sound-field of 30 7th HOA moving sources.
- Sources appeared more stable in relation to a moving listening position.

If we compare 12th order NFC-HOA with wave-field synthesis (WFS), the latter clearly creates a more accurate focused source. But we also need to remember that the results are too different for fair comparison:

- For an audience, WFS produces many small but accurate 'sweet-spots' where you have to be in the correct space to experience the reality of the near-field sound, while HOA produces one larger listening area.
- WFS (at IRCAM) uses 64 loudspeakers for each side array, a total of 256 loudspeakers, while 12th order ambisonics uses only 26 loudspeakers for all four sides.
- 'A little' inside the loudspeaker array is enough to address a more intimate sense of proximity. In my own tests, there was no clear perceptual difference between a focused source located 3 meters or 6 meters away. In this light it is more useful to consider distance and contrast of NFC and normal HOA rather than absolute distance.

We can evaluate NFC-HOA in the context of other distance cues, where NFC's source stability and 'closer' effect add to this more robust information:

- The relative mix of direct sound and the realistic reverberant field.
- Sound identity and spectral balance in the source.
- Spatial gesture, how the sound moves in the space and whether it's apparent size and spectral cues imply movement towards or away from the listener.
- Sound in its environment without considering sound identity (other sources occurring in the same space provide a spatial context for the source in question).

For first movement of *Hidden Values – The Umbrella*, a version was created that experimented with material projected over the WFS arrays. Some sound material was positioned for inside array focused sources and added to the main HOA material. This spatial extension was interesting in that it invaded the perimeter of the listener's intimate sphere. After the performance some listeners commented on this new experience. Ideally, an improved NFC-HOA method could replace the WFS layer.

2.4 Higher order recorded scenes

For decades, ambisonic recordings have commonly be made with the Soundfield microphone. This microphone captures the complete sound-field using four cardioid response capsules located as close as possible in a tetrahedron geometry. The four microphone signals are then combined in a simple linear matrix to convert from 'A-format' to the 1st order spherical harmonics of 'B-format'. With correct decoding the results capture a sense of space and environmental envelopment, but are limited by inherent low resolution. To more thoroughly investigate complex real-world scenes (possibly using 'beamforming' to isolate sounds, easily explained in [9]), a higher order microphone can be investigated.

The EigenMike can in theory produce a 4th order sound-field and was used in practical recording sessions. The locations included IRCAM's Studio 5 (a dry recording space), the Centre Pompidou foyer (a large and open public space) and outdoors close to the Centre Pompidou and IRCAM.

2.4.1 Comparison between the Soundfield microphone and the EigenMike.

Sources recorded with Soundfield microphones (SPS200 and ST250) and the EigenMike where decoded for comparisons in Studio 1 and in the Espro. The studio 1 set-up allows for a 4th order decoding, which is perfect for the EigenMike. However, using normal decoding methods 1st order ambisonics should be constrained to a 4-channel decoding. To cross-test fairly, two methods suited for decoding 1st order over a 24-loudspeaker array were used: the 1st order decoder implemented in Spat and Harpex [10].

2.4.2 Recorded scenes Results

The EigenMike produced a greater 'openness', but spatial accuracy in terms of source azimuth was not greatly improved. Further, the difference we hear between 1st and 4th order synthesised sound-fields are far more pronounced that the differences we hear between Soundfield and EigenMike recordings. This simply confirms what Daniel always shows in [13] - that the spatial encoding of the recorded sound field is far from the theoretical encoding functions.

A key element in EigenMike recordings is the transcoder from microphone signals to spherical harmonics. The transcoder used in this project was that implemented by Daniel accompanying the EigenMike on loan from Orange Labs.

3. COMPOSITION OF 'HIDDEN VALUES'

Initially, *Hidden Values* was intended as a work with live performers and spatial-timbral composition. Our visual perception interacts with our aural perception in complicated ways, making for me, live electroacoustic music a multi-media experience. How our visual perception affects our spatial hearing is an area I have been concerned with for some time. However, I decided that the complexities of visual interaction were more than that possible to explore on during the three-month residence. Subsequently, *Hidden Values* became a pure 'acousmatic' composition fusing musical and dramatic ideas: a musical-drama played in space and an abstract spatial-timbral composition.

3.1 Artistic themes

I chose themes and musical concepts that would yield to the compositional use of space, the projection of near and far information and the transformation between sound masses, sound scenes and precise spatial points. I was inspired by the observation that every year, new inventions push the boundaries of science and enrich our understanding of the natural world. Ancient and seemingly minor inventions have also shaped our societies and affect our everyday in a multitude of ways. A single object can connect to the history of the world, yet the utility of these simple devices go unnoticed. The artistic themes explore directly, dramatically and through metaphor, three of these inventions, resulting in three movements of the work *Hidden Values: The Umbrella, The Lock* and sight correction in the third part called *Optical Tubes.*

Part I: *The Umbrella.*
An umbrella protects from the environment - protects from the rain, snow, sun and to some extent the wind. As a metaphor it protects and saves, defends and deflects, cover and disguise, but maybe its just unnecessary baggage. The Umbrella explores a real umbrella and a real environment, but also the metaphor found in a short poem by Jorge Luis Borges – "Instantes", "… I was one of those who never goes anywhere without a thermometer, without a hot-water bottle, and without an umbrella and without a parachute… If I could live again, I would travel lighter. If I could live again, I would begin to walk barefoot from the beginning of spring and I would continue barefoot until autumn ends."

Part II: *The Lock*
The invention of the lock and key can be traced back over 4000 years. The theme of the lock and key and its metaphors, have been used throughout literary and dramatic history. Locked doors provide safety in a modern world. A lock hides secrets from prying eyes, locks people in, locks people out, represents power and ownership. The Lock plays out a drama between two forces: one represented by the female voice, the other by percussion instruments.

Part III: *Optical Tubes*
'Optical tubes', apparently invented by Descartes, were glass tubes that touched the eyeball like contact lenses, but with the unfortunate side effect that you could not blink! Over 50% of the adult population wear glasses to correct their vision. Seeing the world in focus or through a haze is something we can choose to do. In *Optical Tubes*, imagining how it would have been for objects to appear in focus as you moved towards or away from them is a central musical idea.

3.2 Choice of source material: Voice and percussion.

Voiced sounds are clear in their source. Through sound transformations the identity of the person may change, but only in extreme temporal-spectral distortion will we be able to disguise the vocal nature of the source. Voiced sounds are therefore interesting in terms of our understanding distance cues that relate to sound identity and spectrum: changing the spectrum changes our understanding of source distance rather than source identity.

Percussion sounds are in many ways the opposite of the voice. For this work I chose a practical selection of metal, wood and skin instruments portable enough to move between the chosen recording locations. Even before computer manipulation a percussion sound may already confuse the listener of it's identity and instead we hear the gesture and the energy behind the articulation. Performed gesture and energy, along with ease of transformation and abstraction were important to the work in terms of spatial gestures and to the continuum between abstract sound masses and concrete points.

3.3 The working process

Musical sketches, descriptions of gestures, personas or behaviours and were prepared for the two performers (soprano Evdokija Danajloska and percussionist Gilles Durot). A first recording session took place in IRCAM's studio 5, which is a dry recording space. Materials were worked on in preliminary compositional sketches and a second development recorded in the Centre Pompidou foyer. Further materials were recorded outdoors and additional close microphone recordings of percussion instruments were made. For all recording sessions the following microphone set-up was used:

- EigenMike in a central location.
- Soundfield microphone in a central location.
- Spot microphones on percussion.

- Lapel microphone and medium distance microphone (approximately 1 meter away) for the soprano.
- In addition, in the Centre Pompidou four omni-directional microphones were located in the corners of the enormous space. In Oslo I also recorded some percussion instruments with a very close Soundfield microphone.

Although it is possible to isolate individual sounds in a complex sound-field recorded by the EigenMike with the beamforming software, the sound quality is compromised. For higher quality sources I complimented the results with spot microphones. This allowed flexible spatial and spectral control in composition. The four distant microphones allowed me to explore distant space, as if the performers were heard from far away, along with capturing high levels of ambient sounds. The close Soundfield microphone sources provided rich and enveloping sound-fields.

3.3.1 Everyday work: Blind encoding?

We cannot daily use a concert room as the composition space. How can we compose in 7th order ambisonics and be sure of our results, while monitoring in as low resolution as 1st order? My own compositional workflow was spread between 1st order, 4th order, 7th order and binaural listening environments. I found that shorter periods monitoring in the higher resolution environment allows the development of listening expectations (similar to those acquired when moving between stereo monitoring and sound diffusion performance). When moving to 4th order and worse to 1st order there is of course much guess work, mainly in terms of the individual clarity of multiple sources. Checking a suitable binaural rendering (my own choice being Harpex's binaural rendering [11]) using either the IRCAM or the MIT HRTF sets) can help clarify ambiguities.

3.3.2 'Lower order concert' and musical information.

If we are to develop a musical work that relies on spatial clarity, what happens to that work if a lower order concert decoding is necessary? The sweet spot dramatically reduces in size and audience size needs similar scaling. In addition, even in the shrunken sweet spot spatial clarity will be blurred. However, some spatialisation techniques appear preserved: dynamic motions through and round the space and rapid, highly separated spatial articulations.

3.4 Counterpoint

One clear advantage of working with very high order is that the increased spatial separation means we can hear and control high numbers of simultaneous sounds and develop a more complex spatial counterpoint. By spatial counterpoint I'm referring not simply to multiple sources, but multiple motions and articulations whose individual behaviour is bound up in the gesture of sources, which work together and against each other forming the spatial language.

3.5 Sound images

In previous work I elaborate on the problems of spatialisation tools encouraging composers to consider sound sources a single points [4]. In the current work I was interested in transformations from sound points to sound mass, or from a sound object to a sound envelopment. Amongst these methods were the use of statistically regulated point clouds, control over group motion and lower order, inherently less focused recorded or convolved sources decoded in 'incorrect' ways (for example using *max*-r_V decoding over incorrect loudspeaker numbers). However, as we are working in real-time, and each additional sound adds a CPU load, it is useful to explore the minimum number of points necessary to achieve a specific result in the highest order encoding we choose to work. A few conclusions can be made:

Expanding and contracting between a single point, a sound appearing to have size and shape and complete audience envelopment requires 6-12 instances of the sound, depending on the spectral content. As expected, noisier and textured sounds require more points than harmonic tones. Points need to be partly incoherent in spectral-temporal content. In-phase and identical sounds result in an expanding mono image rather than the implication of an image with shape or dimension. If using less than six points, as the image width expands, the points tend to disassociate and instead we hear multiple mono sources. The results are also somewhat dependent on the decoding order. However, 4th order decoding in studio 1 and a 7th order decoding in the Espro were surprisingly consistent when listening in the sweet-spot. Another useful technique is fast, random motion of a few sounds, which can create the illusion of a cloud. As with the first method, spectral content has much to say for the number of sources needed to create a sound mass rather than a cloud of individual mono points. All of these techniques can be heard in the third part of *Hidden values: Optical Tubes*.

3.6 Illusions of distance and depth

3.6.1 Source recording

As suggested in the hybrid recording technique above, lower order recorded scenes can be rearticulated with high order encodings where point sources recorded from the real scene are layered back into the total sound picture. This simple technique adds the illusion of depth and distance in the scene – extending between close up sources to distance sound-fields - while allowing space-timbre expansions relating to, and departing from, acoustic settings. Although the 'close-mic' dry sounds are not the same as an inside loudspeaker array near-field encoding, when in contrast to distance ambiences the musical-perceptual results can be rather similar.

3.6.2 Illusion of sources inside the loudspeaker array

Although contrasting various distance cues can create dramatic differences in the perceived distance of the source, two ambisonics related techniques can further give the illusion of the sound being located inside the loudspeaker array:

(a) W-channel manipulation.
Controlling the level of the W-channel is a known technique to imply the source is closer to the listener [12]. The W-channel is an omni-directional pressure field and as in real B-format recordings, the level of the W-channel gives the impression of proximity also for synthesised sources. This is especially true when scaled in relation to other distance cues mentioned above. By using 'Spat-oper', the level of the W-channel is automatically correctly scaled in relation to source distance. The user can further scale this parameter outside of spat-oper for more extreme results.

(b) Motion 'through' the space.
For moving sounds it is natural for our perception to find continuity. Similar to our visual perception where we invent the masked information that is actually invisible, I experience that aural perception of moving sounds works in a similar way. If all spatial cues (W-channel scaling, volume, filter change, image size change, reverb-dry source mix) are consistent with our understanding of real motion, then this motion is what we hear even if the actual spatialisation is somewhat discontinuous. For a sound moving 'through' the centre of the space, even though our ambisonics sound-field does not recreate such information, it is nevertheless what we will tend to hear.

(c) Listener space versus 'outside world'.
A sound decontextualised from its real-world scene, such as an acoustic instrument recorded in a dry studio, can be given a new spatial context. Played in the concert space without any other spatial cue than the real room reverberation places the sound in the listener's space. When in contrast to a recorded sound-field, the decontextualised sound appears inside the loudspeaker array, although of imprecise distance, while the recorded sound-field extends the listening space into an 'outer' experience.

All of these techniques can be heard in the second part of *Hidden values: The Lock.*

4. REFERENCES

[1] Daniel, J. Nicol, R. Moreau, S. "Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging". In *Audio Engineering Society Convention Paper 5788 Presented at the 114th Convention. 2003.*

[2] Daniel, J. Moreau, S. "Further Study of Sound Field Coding with Higher Order Ambisonics". In *Audio Engineering Society Convention Paper Presented at the 116th Convention May 8–11 Berlin, Germany.* 2004.

[3] Barrett, N. Spatio-musical composition strategies. In *Organised Sound,* 7(3). *pp 313-323.* 2002.

[4] Barrett, N. "Kernel Expansion: a three-dimensional spatial composition combining different ambisonics spatialisation techniques". In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics.* 2010.

[5] Barrett, N. "Ambisonics spatialisation and spatial ontology in an acousmatic context". In *Proc. from the Electroacoustic Music Studios conference.* http://www.ems-network.org/ems10/ 2010.

[6] Noisternig, M. Carpentier, T. Warusfel, O. "Espro 2.0-Implementation of a surrounding 350-loudspeaker array for sound field reproduction spatial audio in today's 3D world". In *AES 25th UK Conference.* 2011.

[7] F. Zotter, H. Pomberger, M. Noisternig. "Energy-Preserving Ambisonic Decoding". In *Acta Acustica United with Acustica Vol. 98 pp. 37 – 47.* 2012

[8] S. Favrot, J. M. Buchhol. "Reproduction of Nearby Sound Sources Using Higher-Order Ambisonics with Practical Loudspeaker Arrays". In *Acta Acustica United with Acustica Vol. 98 pp. 48 – 60.* 2012

[9] Brown, S. "Beamforming with the Eigenmike". *http://scoff.ee.unsw.edu.au/posters/posters2009/8_1.pdf. 2009.*

[10] Berge, S. Barrett. N. "High angular resolution planewave expansion". In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics.* 2010.

[11] Berge, S. Barrett, N. "A new method for B-format to binaural transcoding". In *AES 40th international conference, Tokyo, Japan, October 8–10.* 2010.

[12] Menzies, D. "W-Panning and O-Format, Tools for Object Spatialization". In *Proceedings of the International Con- ference on Auditory Display.* 2002.

[13] Daniel, J. "Evolving views on HOA: From Technological to pragmatic concerns". In *Ambisonics Symposium, June 25-27, Graz.* 2009