W-PANNING AND O-FORMAT, TOOLS FOR OBJECT SPATIALIZATION

Dylan Menzies

Zenprobe Technologies
dylan@zenprobe.com

ABSTRACT

Real acoustic objects have spatial width and characteristic radiation patterns. Techniques are described for efficiently synthesizing these qualities, by encoding with spherical harmonics. This approach naturally lends itself to Ambisonic reproduction, although it can be usefully applied to other forms of reproduction.

1. INTRODUCTION

The primary problem in spatialization is creating and controlling the perception of sound from different directions, since then in principle any soundfield can be simulated by mixing appropriate sounds in different directions. [1, 2, 3, 8, 10] are all focused on this objective. By contrast, little has been written about the secondary problem of simulating composite structures that generate a variety of sound from different directions over an extended space. The need for doing this arises very naturally from the sound produced by composite real objects, whether solid or fluid in nature. For instance, a passing vehicle emits a variety of different sounds from different parts, projected in various directions. A swarm of bees flying past the listener even creates a soundfield that completely envelopes the listener. One approach to modelling such complex audio scenes is to break it into a set of independent point sources. The game audio technology provider Sensaura [13] does this in its “zoomfx” function. To keep the sources well defined they are generated from simple decorrelated copies of the input source. This has the immediate problem of computational cost, which is relevant when a scene of many extended objects must be synthesized in real-time, as in virtual reality or live musical applications. Also, there are no natural high level parameters immediately available. This is particularly important when working in an aesthetic or musical context, where the ability to shape the sound intuitively is very useful.

The first technique to be described, *W-panning,* is a way of directly encoding the sound for a simplified extended object, without breaking it up into points. As a bonus, W-panning illuminates the psychoacoustics of object perception in relation to rendering processes.

We should also like to simulate the radiation *pattern* or direction-dependence from each source point. As an object rotates relative to the listener, the received signal varies. Computer game developers typically do this by using *sound cones,* as specified in the Windows DirectX API [16], which provide a simple direction dependent variation of a single signal. A realistic radiation pattern must be composed of many such sound cones pointing in different directions. *O-format* is a more efficient and natural way to encode complex source radiation patterns, with simple rendering properties.

2. B-FORMAT AND AMBISONICS

2.1. Encoding

Before describing W-panning and O-format in detail, it is worth briefly reviewing the basic principles of spherical harmonic encoding and Ambisonics. Further information can be found in [1, 2, 3].

A *B-Format* signal is a spherical harmonic encoding up to 1st order. The *W* signal is 0 order and can be thought of as deriving from an omni-directional microphone response applied to the soundfield. The *X,Y,Z* signals are 1st order and correspond to signals from figure-of-eight microphones aligned to orthogonal axis. Figure 1 is a 3d sketch of these responses separately and together.

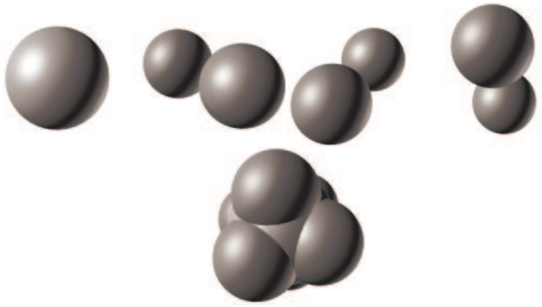

Figure 1: W,X,Y,Z directional responses.

The W signal is weighted such that the B-format signal for a single point source, $\mathbf{S}$, is given by:

$$\mathbf{W} = \mathrm{W}\mathbf{S}, \mathbf{X} = \mathrm{X}\mathbf{S}, \mathbf{Y} = \mathrm{Y}\mathbf{S}, \mathbf{Z} = \mathrm{Z}\mathbf{S} \qquad (1)$$

where

$$\begin{aligned} &\mathrm{W} = 2^{-1/2}\beta, \quad \mathrm{X} = n_x\beta, \\ &\mathrm{Y} = n_y\beta, \quad \mathrm{Z} = n_z\beta \end{aligned} \qquad (2)$$

and n_x, n_y, n_z are the Cartesian components of the normalized direction vector to the point source. β is an overall gain factor. The coefficients satisfy

$$2\mathrm{W}^2 = \mathrm{X}^2 + \mathrm{Y}^2 + \mathrm{Z}^2 \qquad (3)$$

The B-format signal for a scene consisting of a collection of point sources is just the linear sum of their individual B-format signals.

2.2. Decoding

The B-format signal can be decoded by an Ambisonic decoder to generate speaker feeds for a variety of speaker arrays· For example a simple decoder for a cubic speaker rig has the 8 speaker feeds $\mathbf{W} \pm \mathbf{X} \pm \mathbf{Y} \pm \mathbf{Z}$ **.**

Binaural and hence transaural feeds can also be derived from B-format in a natural way, by integrating the HRTFs over the harmonics. Specifically, by constructing the binaural signals of a 1st order soundfield we immediately obtain,

$$\begin{aligned} &F_i = \int d\Omega\, R_i(\theta,\phi) HRTF_L(\theta,\phi),\ i = W, X, Y, Z \\ &B_L = F_W(W) + F_X(X) + F_Y(Y) + F_Z(Z) \\ &B_R = F_W(W) + F_X(X) + F_Y(Y) - F_Z(Z) \end{aligned}$$

Where F_i are convenient pre-calculated filters, R_i are the harmonic functions, B_L and B_R are the left and right binaural signals, and the Z harmonic is aligned along the direction between the ears. Note that this is a rendering of a B-format signal, *not* the source un-approximated soundfield. The inherent limitations of B-format are present, but also the advantages. In particular a single source would not be rendered as sharply as possible by direct HRTF encoding, but only four filters are required. Excellent results have been obtained this way, for example by Lake DSP [14], where synthesized B-format room responses have been used to create B-format scenes that are then converted to HRTF. Dolby Laboratories has adopted this technology for its “Dolby headphones” [15].

Reproduction by vector-based panning is less straight forward and discussion will be delayed to section 2.4.

2.3. Transformation

An important class of transformations acting on B-format signals are linear operations that preserve Equation 3 for spatialized signals. By linearity these operations preserve the property for a signal composed of many separately spatialized signals, which essentially means that the transformed B-format signals still ‘make sense’ to the listener. Rotations and reflections in the x,y,z axis are obvious examples. Less obvious are the *dominance* operations parameterised by direction and a special factor. Depending on whether the factor is positive or negative, the operation will “focus” spatialized vectors towards or away from the direction, with associated rise or fall in gain. More details can be found in [1, 6].

2.4. Summary

B-format is only an approximation to a soundfield, but this is compensated by its compactness and symmetry in many applications. It excels at describing a soundfield that is evenly balanced around the listener, without a strong focus in one direction. This is particularly useful in a fully immersive environment, in which the listener’s attention may be drawn to any direction, for instance in sophistocated entertainment and virtual reality systems incorporating surround video as well as surround sound.

Rendering B-format with Ambisonics provides a holographic image, meaning the 3-dimensional image transforms correctly as the listener’s head rotates, in contrast to HRTF techniques, which require the listener’s head to be fixed, unless cumbersome head tracking technology is employed. The freedom of head movement provides reinforcing location cues that compensate, at least in part, for the inherent directional limitations of B-format. No special HRTF information is required since the listener's head shapes the soundfield as they would the original soundfield, to an approximation.

For the balance we note one area where B-format and Ambisonic technology falls short. To render a point-like source that is very close, for instance under the chin, requires head filtering, and there is no escape from HRTF techniques of some kind. Essentially, the head has become too entangled with the source to allow construction of the soundfield from distant loudspeakers such that the speaker feeds and source position can be kept constant while the head is rotated. Sensaura implement this case with special near-HRTF functions called "macrofx" [13].

In summary, B-format provides an efficient delivery-independent encoding format, suitable both for synthesis and recording. It is biased towards use with Ambisonic reproduction, which has advantages over HRTF based rendering of holography, speaker-array scalability, HRTF independence.

3. W-PANNING

3.1. Simplified Source Model

The direct approach to synthesising an extended object is to immediately calculate its B-format signal. Two extreme cases help to clarify the situation. At great distance an object appears as a point source. Then when the object surrounds the listener, sound comes from all directions. Figure 2 illustrates this idea.

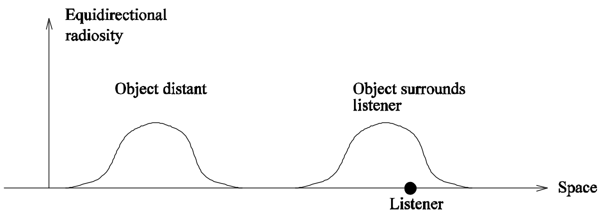

Figure 2: Diffuse source distant and close.

In the simplest case where the object emits the same signal from all parts, and symmetrically in all directions, the resulting B-format signal is non-zero only in the W component[1]. Thus as the object approaches the listener the W component increases relative to the X,Y,Z components. In the literature of Ambisonics, this variation of W is called the *interior effect*. Analog panning devices could produce the interior effect but it was not controlled in such a way as to synthesize object width, [4]. For instance the central position of a joystick control might be a signal zero.

The B-format signal resulting from any placement of the object could be calculated explicitly from the distributed radiation pattern. However, from a control viewpoint it is more instructive to find intuitive high level parameters, and vary these. The extent of the interior effect can be quantified by introducing a parameter γ into (3) ,

$$2\gamma \mathrm{W}^2 = \mathrm{X}^2 + \mathrm{Y}^2 + \mathrm{Z}^2 \quad 0 \le \gamma \le 1 \tag{4}$$

An analogue of (2) is required that incorporates γ. Additionally we should like to control the overall energy of the signal independently of γ. For the second parameter we redefine β as a pressure gain factor determined by the sum of the harmonic energies,

$$2\mathrm{W}^2 + \mathrm{X}^2 + \mathrm{Y}^2 + \mathrm{Z}^2 = \beta^2 \tag{5}$$

From (4) and (5) the following coefficients are found,

$$\begin{gathered} \mathrm{W} = \beta(2+2\gamma)^{-1/2}, \\ \mathrm{X} = an_x\ , \quad \mathrm{Y} = an_y\ , \quad \mathrm{Z} = an_z \\ a = \begin{cases} \beta(1+\gamma^{-1})^{-1/2} & 0 < \gamma \le 1 \\ 0 & \gamma = 0 \end{cases} \end{gathered} \tag{6}$$

It remains to determine how the parameters γ, β should vary with distance r. This could be calculated exactly by integration, but from a practical and aesthetic viewpoint it is better to show what form the variation should take and then allow modification. At distance γ tends to 1, and is 0 and smooth at $r = 0$. The scale of transition near zero is of the order of the linear size of the object. Its shape depends on the radiation distribution from the object. An example function incorporating this behaviour is

[1] Note that the W representation is very convenient for expressing this.

$$\gamma = \frac{r^2}{r^2 + w^2} \tag{7}$$

where w is the object width scale.

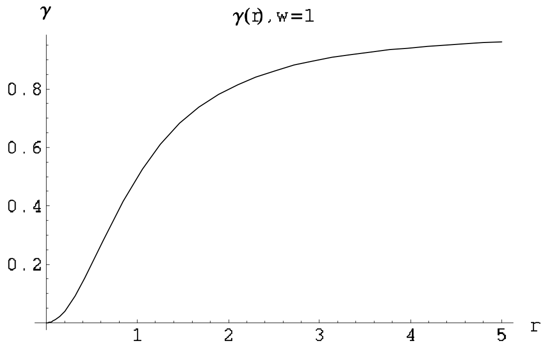

Figure 3: γ radius profile.

The approximation can be refined further by finding an asymptotic law for γ at large r, determined by the bulk width, but in practice this is not very useful, and will not be considered further here.

At large distances, in open 3-dimensional space, β tends to br^{-1} for some constant b, since the signal is proportional to the *sound pressure level*. This can be generalised slightly to br^{-k}, allowing for decay rates associated with 2 and 1 dimensional spaces, where k is 0.5 and 0 respectively. At $r = 0$, β has a smooth maximum, a, determined mainly by the object width: A more compact object gives a relatively large maximum. An example fit for this behaviour is

$$\beta = \frac{1+r}{a^{-1}(1+r) + b^{-1}r^{1+k}} \quad k > 0, r \geq 0 \tag{8}$$

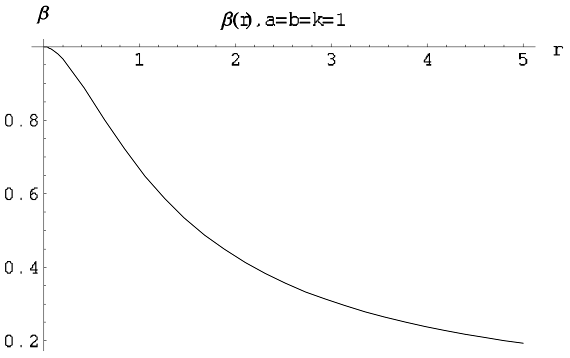

Figure 4: β radius profile.

3.2. A More Realistic Source

The main assumption was that the source radiated the same signal uniformly and isotropically, leading to no directional components at $r = 0$. A more realistic model would allow for some variation in the frequency content of the signal. This can be introduced by filtering the source signal into two or more complementary signals, bleeding these into the X,Y,Z signals, and reducing the energy to the W signal. This is related to the 'monospreader' described by Gerzon in [5], for spreading a single source around 360 degrees. In this case the W component remains unboosted however. Note that this is distinct from the vector base 'spreading' discussed in [11], where the frequency content is uniform, and the primary concern is to maintain an isotropic image width.

Although distance gain variation has been introduced as an integral part of W-panning, nothing has been said about other distance cues, such as high frequency attenuation and reverberation. These can be incorporated independently, for instance as described in [8].

3.3. Encoding width into HRTFs

The underlying idea in W-panning, of width encoding, can be applied more directly to non-Ambisonic reproduction methods, allowing for the possibility of greater resolution. For example with HRTFs, width modified HRTFs could be precalculated for objects of different apparent width (including surround). Intermediate widths could be obtained by interpolating HRTFs in the similar manner to directional interpolation. One attraction of using intermediate spherical harmonics for encoding is that only the four integrated HRTFs for the harmonics are needed rather than the much larger collection of sets over direction and width.

3.4. Evaluation

No formal listening tests have been conducted on this approach. Quantifying width and immersion experimentally is inherently more difficult and prone to subjective variation than directionality. However, some simple supporting observations can be made. Apart from the properties of efficiency and controllability previously mentioned, one major advantage over the many-body approach is that movement of the object through the centre is very smooth. With many-bodies the amplitude and directional emphasis will peak and swing as different bodies happen to come close to the centre, tending to disrupt the sensation of a unified object.

W-panning has been deployed in software intended for electroacoustic music creation. The “LAmb” (which received a prize at the Bourges competition in 1997) is available for download [12]. It can give very good results on a square four speaker array, the simplest possible for Ambisonic reproduction. It includes controls for decorrelation, distance eq, global rotations and dominance, and live hard-disk recording. Interestingly, decorrelation has little effect in a medium to large concert setting, although for a small room it can be useful. The consensus amongst composers from around the world that have used W-panning in this software, is that it provides an effective way to create and control the impression of object width and immersion.

From experience two factors contribute to the realism of W-panning. Object movement is very important, particularly when moving through the centre. A static central object can switch perceptually to an ‘in head’ sound. Also some room acoustic helps reduce this effect. This is likely to be due to the creation of early reflections, which adds realism by providing correlated cues. Using the decorrelation-spreading technique described in Section 3.2 performs a similar function. Note that the maintaining a consistent energy profile by establishing (5) is vital to the integrity of the effect. Any energy ‘dips’ or ‘bumps’ in the central region are very noticeable.

The image created for a central object, by a dominating W signal, is effective in creating the impression of nearness by logical implication of being immersed in a continuous medium. However it cannot create the very-near, or *macro* effect described in Section 1.4.

3.5. Application to vector-based panning

Some discussion will be given here to the application of W-panning to the vector base panning (VBAP) technique for spatial sound reproduction introduced in [10]. Essentially a VBAP supposes we have a speaker array in which each speaker can be controlled independently by the spatialization process. The B-format signal could be rendered by using multiple vector bases, or ‘MDAPs’ to fit an appropriate Ambisonic decoding, as indicated in [11], but nothing new is gained from this. Alternatively, the more directed B-format components could be treated with fewer bases. A distant object would be handled by panning between at most three speakers. Note however, that it is possible for objects located at the side of the listener in this manner to be very poorly localized by the listener, as reconfirmed in [10]. This after all, was part of the original reason why Gerzon and others pursued alternative methods of spatialization to quadraphonics, which was a simple form of vector panning on a four-speaker array.

If one has a speaker array with independent control of all the feeds, then other forms of mapping B-format, or higher order encodings of the object, may be considered other than VBAP. For a more in depth comparison of VBAP and Ambisonics see [7]. The name “Ambisonics” reflects the fact that natural signals are frequently not from just one direction.

4. O-FORMAT

4.1. Sperical Harmonic Radiation

B-format is the use of low order spherical harmonics to approximate a soundfield at a listening point, with *inward* moving waves. The same idea is applied now in reverse for sound radiated from an object. Hence the *outward* moving sound is approximated by spherical harmonics. To first order we shall call this approximation *O-format*. O-format carries all the advantages of B-format. It can convincingly capture complex distribution of sound over the full sphere, and can be manipulated well mathematically.

4.2. Encoding O-Format

As with B-format, there are two options. Either the O-format is recorded or synthesized. Recording can be done by placing microphones around an object, and then processing to generate harmonics. An alternative ‘creative’ approach is to just use a B-format recording as O-format, which can be thought of as ‘turning the signal inside-out’. Synthesis is the same as for B-format, and is achieved by adding several mono signals that have been each spatialized in different directions.

This could include signals encoded with width using W-panning.

4.3. Rendering O-Format

In the simplest case we have a point source with radiation described by an O-format signal. The listener only receives radiation arriving along the line of sight from the source, as shown in Figure 5.

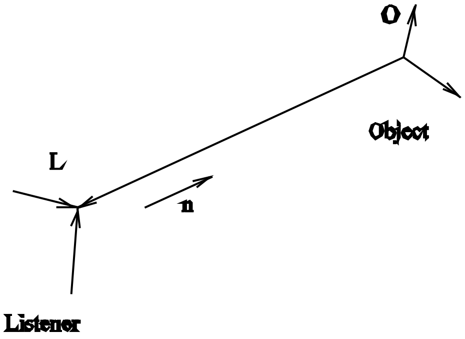

Figure 5: Audio ray from object to listener.

We wish to sample the object radiation in this direction and spatialize the resulting signal in the direction using the procedure given by (1),(2). To formalize this, write the O-format signal as a 4-vector,

$$O = \begin{pmatrix} W \\ X \\ Y \\ Z \end{pmatrix} \tag{8}$$

and for convenience define the following two 4-vectors in terms of the direction normal (3-vector), **n**.

$$N = \begin{pmatrix} \sqrt{2} \\ -n \end{pmatrix} \tag{9}$$

$$\overline{N} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ n \end{pmatrix} \tag{10}$$

The signal received at L in the direction **n** is then given by the sum of the signals from the spherical harmonic components in this direction, which is written compactly as the dot product

$$O.N \tag{11}$$

The spatialized B-format version of this signal, to use for rendering, is

$$(O.N)\overline{N} \tag{11}$$

This operation is linear from O-format to B-format and preserves the relationship (3) by construction. From Section 1.3 it is known that such an operation in general consists of rotation, reflection and dominance. Since the resulting signal is encoded in a single direction rather than spread, we infer that this is a *maximum dominance* operation.

4.4. Adding width to O-format

Consider now an object with width that radiates according to a single O-format signal from each point on it. At any time, the listener receives signals over a range of directions centred about the direction to the object centre. This spread of signals can be calculated approximately by reducing the dominance factor slightly from maximum. A lower dominance value gives more spread, corresponding to a closer/larger object. Clearly this technique breaks down for large spreads, and it is only an approximation, but it does provide a very simple way of extending the rendering procedure.

4.5. Resonance models

In the context of interactive virtual worlds, an appealing application of O-format is to use it in the calculation of output from resonance models, thereby adding rich spatial qualities to the resonance model. For instance, the impulse response of an object could be stored in O-format and used to generate an O-format response to general input. Alternatively an O-format resonance algorithm, for example using a delay feedback structure, could be used to synthesize output without using a stored impulse.

4.6. Alternatives to B-format

Having obtained the signal from the source, $O.N$, it may be desirable to bypass the B-format encoding and go direct to the rendering system. For instance the signal could be rendered with HRTFs. O-format retains its usefulness as a compact radiation pattern encoding, but we loose the dominance spread control.

4.7. Evaluation

O-format is more amenable to testing than W-panning, but again only informal tests have been performed. The principle question is whether O-format encoding improves synthesis over using sound-cone based methods, and here it is certainly successful. It offers the possibility of recording object radiation, creating complex radiation patterns compactly, and rendering with low costs. O-format may be criticised for not being able to encode sharply directed signals. In practise such signals are unusual and do not provide reliable cues for object recognition. They occur "naturally" when high frequency signals interfere, for instance from machinery such as hard-drives. If needed they can be generated with additional soundcones.[2]

To appreciate an O-format object it is vital that the listener and object rotate relative to each other. This can happen simply by relative motion, or by explicit rotation of the object in the world frame.

5. CONCLUSION

W-panning and O-format are complementary techniques. W-panning gives the impression of width, and by implication for central objects, nearness, while O-format encodes variation of radiation with direction. Together they provide an inexpensive way to enrich the synthesized sound world.

In the introduction the need to avoid many point sources was given as a reason for pursuing W-panning. However, it makes sense to group a few W-panned objects to create an extended object with radiation varying over space. The overall result is that we have replaced a large number of point objects with a few W-panned objects. The correlation inherent in the motion of such an extended object can further increase the sense of immersion. Similarly O-format objects can be grouped to create extended objects, and even combined with W-panned objects for greater flexibility and variation.

REFERENCES

[1] M. Gerzon "Practical Periphony: The Reproduction of Full-Sphere Sound" in *Lecture 65th Audio Engineering Society Convention.*

[2] M. Gerzon, "General Metatheory of Auditory Localisation" *92nd Audio Engineering Society Convention*, 1992.

[3] Malham, D.G. and Myatt, A. "3-D Sound Spatialization using Ambisonic Techniques" *Computer Music Journal*, 19;4, pp 58-70, Winter 1995

[4] M. Gerzon, *NRDC Ambisonic Technology Report 3, Pan Pot and Sound Field Controls*, 1975.

[5] M. Gerzon, NRDC Ambisonic Technology Report 4, Artificial Reverberation and Spreader Devices, 1975.

[6] M. Gerzon, G. Barton, "Ambisonic Decoders for HDTV", *Preprint 3345 of the 92nd Audio Engineering Society Convention*, 1992

[7] J.-M. Jot, V. Larcher, J.-M. Pernaux, "A Comparative Study of 3-D Audio Encoding and Rendering Techniques", *16th International conference of the Audio Engineering Society*, April 1999 Apr

[8] J.-M. Jot, O. Warusfel, "A Real-Time Spatial Sound Processor for Music and Virtual Reality Applications", *International computer music conference*, Sept 1995

[9] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing virtual sound source attributes using a binaural auditory model", *Journal of the Audio Engineering Society*, 47(4) pp. 203-217 April 1999

[10] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning", *Journal of the Audio Engineering Society*, vol 45, no 6 1997 June

[11] V. Pulkki, "Uniform spreading of amplitude panned virtual sources," in *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1999, pp. 187-190.

[12] D. Menzies, "LAmb, an Introduction and Tutorial", *Bourges Synthese, www.imeb.net*, June 1997.

[13] Sensaura, "3-dimensional positional audio", *www.sensaura.net*

[14] Lake DSP, "Spatial audio processing", *www.lakedsp.com*

[15] Dolby Laboratories, "Dolby Headphones", *www.dolby.com*

[2] O-format could be extended to higher order harmonics for more resolution, but the added costs may out weigh its usefulness, as with higher order Ambisonics.

[16] Microsoft Developers Network Library, "Soundcones", *msdn.microsoft.com/library/*