Sound Spatialisation

D.G. Malham
Department of Music, University of York, England
dgm2@york.ac.uk

Abstract

Towards the end of the nineteenth century two inventions, the telephone and the phonograph, appeared which were to change the way music was dealt with. Prior to the developments they brought about, every musical performance was indivisible from its place in time and space. Their appearance meant that music could be presented remotely in both time and space from its origin. This has inevitably resulted in various forms of distortion - nonlinear, spectral, temporal, spatial, - of the original. Whilst it has proved relatively easy to deal satisfactorily with the first three so that we can now present "remoted" music which is excellent in all of those three aspects, removing the distortions in the spatial presentation has proved far more intractable. Even the best systems in use today for sound "spatialisation" are relatively crude, allowing for little more than the creation of an illusion, sometimes very good, more often poor.

However, despite this, the creative use of sound spatialisation is becoming more and more important, whether for serious avant-garde composers, computer game designers, in cinema, television and multimedia productions or in audio recording. It is anticipated that the demand will escalate even more with the appearance of DVD with its multiple audio channel capability. This tutorial paper briefly covers the basic directional hearing mechanisms of the human brain before examining in more detail the various different ways of dealing with sound spatialisation, starting with headphone related technologies such as binaural and transaural. Loudspeaker-based systems will then be covered, starting with conventional stereo followed by cinema style surround sound systems. Finally a true 3-d system, Ambisonics, will be examined. The advantages and limitations of all the systems, both aurally and in terms of difficulty of implementation or control, will be covered. It is hoped to give a number of demonstrations.

1 Introduction

Our hearing provides us with our only fully three dimensional information about remote, ie non-contact, events. Through the medium of sound, we are able to perceive where acoustic sources are in the space around us, including above and below, whether they are moving or stationary; under the right circumstances we can also estimate the distance of the sound producing object as well as get some idea of its nature and size. For most of recorded history and for the entire period life existed on the Earth before that, those creatures blessed with directional hearing have existed in a 3-D sonic environment from which they could not escape since, unlike sight, hearing cannot be cut off by anything as simple as night or a blindfold. It was only with the appearance of technology in the form first of the telephone, then of the phonograph, that sonic events could be shorn of their three-dimensional nature. Audio and recording engineers, composers and sound artist have been striving ever since to find ways of capturing and re-creating 3-D *soundfields* or even, of course, creating artificial ones. Generically, the body of techniques that has been developed has come to be known as *sound spatialisation*. This paper will examine some of the issues involved in sound spatialisation, concentrating mainly on methodologies which can meet the needs of the creative artist. In order to appreciate the nature of the problem, we need to have some understanding of two important areas, namely, the way sound behaves in space and the way the human directional hearing system functions

2 Sound in Space

As is well known, sound is carried through air as longitudinal pressure waves. These expand outwards from whatever is emitting them, reducing in level as they spread, reflecting off or being diffracting by objects that they encounter, their spectral contents changing as they interact with the physical properties of these objects in ways which may also change with the angle of incidence. They also interact with the air they travel through, losing higher frequencies progressively with distance as a result of absorption by humidity in the air. Even for the hypothetical *point source* which emits simple spherical wavefronts, the soundfield produced in anything other than *free space* rapidly becomes very complex both spatially and timbrally. This is illustrated in *Figures 1* and *2*, which were produced by Damian Murphy's 'WaveVerb' multi channel spatial simulation system currently being developed as part of his Dphil research with the Music Technology Group at York (Murphy, 1998). For a normal, real world, sound source this picture is further complicated by the behaviour of its extended emitting surface, since this will usually vary in a non-simple manner with both position and frequency. For instance, sounds with wavelengths larger than the size of the body will behave much as if they had been emitted from a point source, so their intensity will drop by the familiar 6dB per doubling of distance. For sounds with wavelengths much less than the size of the object, this may well reduce to 3dB per doubling.

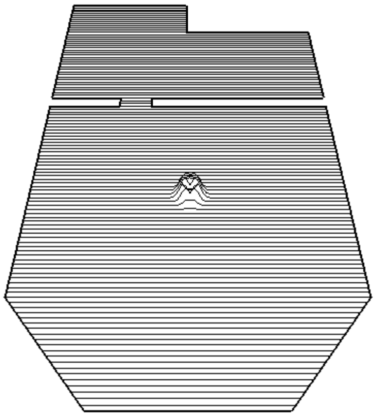

Figure 1. Impulse initiated

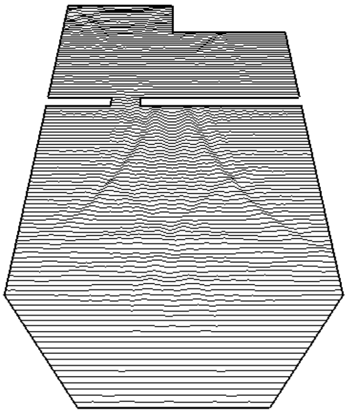

Figure 2. Impulse after several reflections

It is very tempting when dealing with sound and hearing to employ visual analogs. Such analogs are of some use, but they should be treated with extreme caution since, in practise, the differences between the two outweigh the similarities. Firstly, although the audible part of the acoustic spectrum, normally taken as 20Hz to 20KHz, spans some ten octaves or so, the visible part of the electromagnetic spectrum covers only around one, from 400 to 770 nanometres. Secondly, for all but a minority of everyday situations (such as soap bubbles and oil films) the structures with which visible light interacts are vastly larger in extent than the wavelength of the light itself. Soundwaves, on the other hand, frequently have wavelengths that are larger than the environments in which they are heard. As a result, constructing artificial soundfields using the simple methods commonly employed for electromagnetic simulations, such as *image modelling* or *ray tracing,* can give disappointing results, especially at the lower frequencies. Despite this, the significant computational penalties of more accurate methods such as *finite element* or *boundary element* modelling, (Begault, 1994:187) mean that the simpler methods are still widely used, especially when working in real time. Fortunately, as we shall see later, our hearing mechanism has evolved to deal with ambiguous or incomplete information so a simulation which is reasonably well made and provides a number of sufficiently good *hearing cues*, will be accepted by the ear as *true to life.*

3 Directional Hearing in Humans

When immersed the sort of complex soundfield discussed in the previous section, a number of cues are available to the directional part of our hearing mechanism. It should be noted, however, that because of the complexity of most soundfields, it is possible for these cues to be ambiguous or conflicting. This is especially common in artificial soundfields, whether synthesised or recorded, but it also happens in real world situations. Under these circumstances, the perceived direction and distance of a sound source may not match the actual direction and/or distance. It should be noted that in making these judgements, it appears that the brain weights each cue according to its apparent degree of unambiguity. It is this factor which enables us to construct useable artificial soundfields. Whilst the advent of digital technology and the computer has greatly increased what we can do, we cannot at present recreate exactly an original soundfield (or construct an artificial one of a similar complexity) if it extends over any significant area, though we can do so for a small number of points. By concentrating on a subset of the possible cues and trying to make them as unambiguous as possible, artificial acoustic environments with acceptable performance can be made using relatively simple equipment.

Although there are maybe other, more subtle mechanisms, we can define the main cues used to determine the position of a sound source as follows;

- The time of arrival of the wavefront of a sound event at the ears, or more specifically, the difference in arrival times between our two ears. A sound source anywhere on a line from due front, through due above to due back (the median plane) will have its wavefront arrive at the two ears simultaneously. Move the source away from this line and one ear will begin to receive the wavefront after the other. This is known as the *Interaural Time Delay* or ITD. Note that this is really related to the phase difference between the two ears, rather than the actual time difference, since if the brain time differences directly, this would need to be capable of discriminating time spans in the microsecond range.

- Sound from a source to the left of the head, for example, will arrive directly at the left ear, but will be diffracted round the head to get to the right ear. Its amplitude will be less at the right ear than the left, both as a result of the screening effect of the head and, to a lesser extent, due to the extra distance travelled. This is referred to as the ILD (*Interaural Level Difference*).

- The shape of the head and the external part of the ears results in a frequency dependant response which varies with sound position and is, in general, different for each ear. Although this is often referred to as the *Head Related Transfer Function* (HRTF), strictly speaking HRTF's also include ILD and ITD's. For this reason, this should really be referred to as *Head Related Frequency Response* (HRFR). For positions where ILD's or ITD's give ambiguous or nonexistent differences between ear signals (such as median plane signals) this is the main positional sensing mechanism. It is also one of the two main mechanisms for distinguishing frontal sound sources from rear ones.

- Our ability to change the position of our head to minimise the ITD, ILD and the difference between the HRFR's at the two ears. This is, or should be, the point at which we are directly facing towards (or away from) the sound source. This is also the other and possibly main, mechanism for *front-back discrimination*, which is accomplished by observing whether inter-aural differences are increasing or decreasing for a particular direction of head movement

The main cues for the distance of a sound source are;

- The ratio of direct to reverberant sound - in a reverberant environment, the energy in the reverberant field stays more or less constant for all combinations of listener/source positioning, (so for a given source level, the reverberation loudness remains the same) whereas the source loudness drops off with increasing distance from the listener.

- The pattern of directions and delays for the first few (ie the early) reflections off surfaces in the environment changes in a manner which is dependent on both source and listener positions.

- Higher frequencies drop progressively with distance, due to absorption by moisture in the atmosphere

- The fall of loudness with distance

The interpretation of the last two is heavily dependent on acquired knowledge of both the spectra and loudness of the sound source. In particular, reliance on loudness for distance cueing is known to be of very doubtful value, since experiments in anechoic chambers have shown errors of more than two to one when subjects were asked to estimate the distance of a sound source. (Nielsen, 1993)

We should note here that these are not the only ways that the body perceives sound, nor indeed are the other perceptual mechanisms unable to provide directional cues. Unfortunately, because of the difficulty of working experimentally on, say, chest cavity pickup or bone conduction mechanisms (and the corresponding ease with which headphone-based measurements can be made) little work has been done on these means of perception and their directional discrimination capabilities. Suffice to say that, as a result of informal experimentation, we are convinced that such mechanisms should be taken seriously. In particular, we believe that the chest cavity may play a roll in low frequency directional discrimination and that the commonly held belief that we cannot determine the direction of sources in the very low bass, where the phase difference between the ears becomes very low, may only hold fully true for headphone presentation. This has an important bearing on the choice of on headphone versus loudspeaker presentation methodologies and on the use of separate LFE (Low Frequency Effects) channels or supposedly "non-directional" subwoofers.

4 The Techniques of Sound Spatialisation

Sound can be spatialisation in essentially two ways. Firstly, the system can attempt to provide signals, usually but not always via headphones, *directly at the ear cnanl entrance* similar to those which would have occurred had there been real sound sources in the intended positions. Alternatively, the system can be designed to produce a soundfield in a more extended space which the listener will be able to interpret correctly to produce the desired results.

4.1 Headphone-based systems

In this section we will consider systems that are intended for listening to over headphones or which use the same approach but are modified so that loudspeakers can be used[1].

This is perhaps the most obviously "correct" way of approaching the problem of full 3-d spatialisation of sound. Exact duplication of what the ear would hear in a natural situation should indeed produce the best results. In fact, under a certain limited set of circumstances, this is true, there can theoretically be no closer approach to verisimilitude. There are, however, some very real problems.

For recording natural soundfields, a model head known as a *dummy head* is used with microphones inserted in its ears. This approach was, as far as can be ascertained, first adopted as early as the 1920's by Dr. Harvey Fletcher and his team at Bell Labs (Sanal, 1976:832) and has been used in various forms ever since. However, when a synthetically constructed soundscape has to be produced in a computing system, each sound source must be treated using the appropriate HRTF's for the *source to left ear* and the *source to right ear* paths. The HRTF's, which naturally have to be different for each different source position to ear path, can be produced in a number of different ways. They can be

1. measured on the individual listener (*individualised*)
2. the average of many different listeners HRTF's (*generalised*)
3. measured on a dummy head, which will itself usually have been generalised from the measurement of many individuals
4. calculated from a mathematical model (*synthesised*).

As may easily be deduced, the individualised approach works best - indeed it can potentially produce reality-equivalent results - but the difficulty of measuring every possible HRTF for every possible user of a system means that this is currently only used in research systems. For normal, everyday use, generalised HRTF sets are used but, unfortunately, these cause some very real problems. Whilst the mismatch between an individual's ILD or ITD cues and those of a generalised set are likely to be small and lead to correspondingly small angular source position errors, the differences between individual and generalised HRFR's can be quite gross, especially at higher frequencies. Because of the importance of these cues for front-back discrimination, *front-back reversal errors* become much more common and even complete failure to perceive any sounds as being at the front (or rear) is not unknown. However, if the position of the listeners' head can be tracked and used to select the appropriate set of HRTF's, head rotation-based cues can be used for front-back discrimination, greatly reducing the number of such errors. With head tracking, even seriously mismatched HRTF's can become usable, although the effect is only present during head movements and it can be quite disconcerting continually having the image swapping between correct presentation (during movement) and incorrect (when still). Without head tracking, even using personalised HRTF's, problems occur because the soundfield is fixed with respect to the head, rather than the exterior world. This causes significant problems for, say, recordings that might be listened to from walkman type systems, where the listener is moving, but is less of a problem for situations where the listeners head is normally less like to be mobile, such as when working with a computer or watching television.

So far, we have been discussing the use of such HRTF based, or *binaural*, systems in a fairly theoretical manner. In practise, there are further significant limitations to this approach. The computational burden of the binaural approach is high, even for a single sound source. HRTF's are usually stored and processed as *impulse responses.* These typically comprise, at the CD sampling rate of 44.1KHz, of some 512 samples for each of the HRTF's source-ear paths, although various data reduction techniques can be applied (Begault, 1993:158) to reduce these numbers. The application of these HRTF's to the sound from each source is done with *Finite Impulse Response* filters so each sample of any one sound source will require some 1024 multiple-accumulate cycles in order to produce the two ear signals, although there are techniques for reducing this burden, such as those used in the Lake DSP Huron box. For simple sound imagery, this is not a problem on modern hardware and indeed almost every soundcard found in current PCs has some variant of this technology built into it. The best of these can, and do, produce good results, provided one is dealing with simple sound imagery. As soon as the imagery starts to get close to that of a real world soundfield, the computational burden becomes excessive preventing generation in real-time, even

[1] Such loudspeaker presentations of headphone type material are generally known as *transaural* systems. This term was originated by Duane Cooper and Jerry Bauck (Cooper and Bauck 1989) and was registered by them as a trademark.

using massively parallel supercomputers. The extra burdens of manipulating and interpolating between multiple sets of HRTF's result in this limit being reached much earlier when head tracking needs to be used. For the foreseeable future, soundfields of near real-world complexity, at least those synthesised using the direct HRTF approach, will only be realisable off-line and without the option to apply head tracking. A further disadvantage is that it is currently impossible to use a binaural recording of a natural soundfield in a head tracked system as there is no known way of changing the HRTF's applied to each sound source during the recording so that there are new ones applied corresponding to the changed soundfield/head orientation since there are too many unknowns involved. The same limitation applies to the output from off-line full complexity HRTF-based soundfield synthesis programs. This does not, of course, eliminate the possibility of precomputing a high complexity background soundfield against which a smaller number of active sources could be positioned, since a number of such soundfields containing the same sonic sequences but with different orientations could be generated prior to realtime use. Interpolation between the nearest precomputed orientations can be used to generate all possible intermediate head positions thus placing far smaller computational loads on the realtime system. It does, however, impact significantly on the data storage requirements of the system. With the appearance of large capacity, low cost storage media, such as DVD, this may be less of a consideration.

As mentioned earlier, direct headphone presentation is not the only option for binaural material. It can also be used within the context of loudspeaker-based systems. In such systems, there is a degree of *crosstalk* between the signal streams intended for the two ears as they are no longer seperated by the headphones. Instead, the right ear receives not only its own signal as emitted from the righthand speaker, but also the one intended for the left ear emitted from the lefthand speaker. The same thing happens for the opposite ear path. It is possible to cancel sufficient of this out using a system known as *interaural crosstalk cancellation* (Cooper and Bauck 1989) where a cancelling signal for the crosstalk from the left ear signal is emitted from the right-hand speaker and vice versa. Theoretically, if a transaural system is to work to the fullest extent, the orientation and location of the listener relative to the speakers must be precisely known but this is generally out of the control of the designer. Nevertheless, careful design has enabled the production of marketable systems, as evidenced by the number of two speaker 3-D surround sound options now available on PC cards, TV's etc.

4.2 Loudspeaker Based Systems

This terminology is used to denote systems which in some way attempt to create the illusion of the existence of a real soundfield directly within the listening space. The term *illusion* is used advisedly, since, despite claims to the contrary, it is impracticable with current technology to recreate fully the three-dimensional soundfield over any significant area, owing to the large number of information channels that would be necessary[2]. Nevertheless, there are several ways which a much more severely limited number of channels can be used to create a subset of the soundfield containing within it a set of cues of a sufficiently unambiguous nature that the listener is provided with an acceptable illusion. In the limiting case, where only two channels are available, these can be used to provide either a *stereo* image of the kind familiar for the last four or five decades[3] or a partial, usually horizontal plane only, surround image. Note that here we are dealing with *transmission* channels, not with *reproduction,* ie loudspeaker drive, signals which may be significantly larger in number. For the purpose of this paper, we will limit discussion to three main types of system, namely stereo, *cinema style* surround and full 3-d surround based on *Ambisonic* technology.

4.2.1 Stereo

Strictly speaking, stereo means 'solid' so any sound reproduction system other than a pure, single speaker, *monophonic* one can be described as stereo, but industry usage has limited its meaning to systems using two channels of audio to drive two speakers placed so as to cover a small arc, usually around sixty degrees wide, in front of the listener. Occasionally this is extended to two or three channels driving three or more loudspeakers but, in order to simplify matters, we will only discuss two channel, two speaker systems here. Within the context of this definition, the distinguishing feature of a stereo system is that, unlike the surround sound systems we shall look at later and the binaural systems we looked

[2] The exact figures given in various sources differ, but all agree that based on information theory arguments, it would take many hundreds of thousands of channels and speakers to fully recreate the soundfield within a 2m diameter sphere over the entire range of audible frequencies. More limited reconstructions are, however, possible in specific circumstances using *sparsely sampled* arrays of speakers. See, for example Boone, 1995.

[3] Although it dates to much earlier than that (Askew, 1981) (Fox, 1982)

at earlier, it only attempts to cover a limited *sound stage*, usually in front of the listener.

There are two main ways of producing a stereo image. Either amplitude differences or time differences between the two speakers can be used. The first is by far the most common approach, since the *pan* control as found on almost every mixing console is an example of an amplitude difference based system. The many recordings made with *coincident pairs* of directional microphones as their main or even only stereo source are also examples of this intensity panning approach. There are relatively few cases in which time differences are used in synthetically generated stereo, though it is the main mechanism for image generation in recordings made with *spaced pairs* of non-directional microphones.

At low frequencies (below around 700Hz) an amplitude difference of between 15 and 19dB is sufficient to move the sound fully into the loudest speaker, assuming a subtended angle of 60 degrees between the speaker as viewed from the listening position. At or below that frequency the variation of position with amplitude follows the well known *stereophonic law of sines*;

$$\sin\alpha = \frac{L-R}{L+R}\sin\theta$$

where α is the apparent position of the source, **L** and **R** are the signals fed to the speakers and θ is the angle subtended by the speakers at the listening position (Bennett, 1985:315). Above 700Hz the apparent angular source location produced by this rule increases although it has been found (Clark, 1957:108) that multiplying the (L-R) component by 0.7 above this frequency could partially compensate for this. A more complete rule for this compensation is given in Bennett (1985). Even though this requirement has been known about since Blumlein's work in the 1930's (Blumlein, 1931), this multiband compensation is rarely used, but fortunately there are sufficiently strong cues produced in the lower band for most people to obtain good results from stereo without it. This strong cueing is a result of the fact that the vectorial additions of the signals from both loudspeakers at each ear results in signals with the correct phase differences appearing at both ears - in essence the original wavefront is simulated for central listeners. Curiously, for intensity stereo, the crosstalk which causes difficulties for loudspeaker presentation of binaural material is actually what makes the system work, at least at low frequencies. At higher frequencies, head shadowing comes into play rather than these phase differences (Clark, 1957:109) which is the reason for the difference in apparent source location. A more comprehensive coverage of this is given in Bennett (1985) and in Gerzon (1994).

Stereo has a number of limitations, the main ones being

- Its limited, front only, soundstage, caused by the fact that the image positions central to the pair of loudspeakers, being *phantom*, are inherently less stable than those produced nearer the speaker positions so speaker separations of more than 60 degrees become unacceptable.

- The increasingly poor performance as the listener moves off axis

- Difficulties with image stability under head rotation with, in the limit where the listener is parallel to the speakers rather than facing then, it is impossible to generate stable central phantom images (Thiele, 1987)

4.2.2 Cinema Style surround systems

In Cinema Style surround systems, as typified by Dolby 5.1, additional channels are added to the standard front stereo pair. Firstly, a central loudspeaker channel is used between the front pair. This system has long been used in cinemas as a means of "locking" the dialogue to the screen and for improving the performance for off-centre listeners. Secondly, a pair of channels are devoted to surround speakers, placed on the rear half of the side walls and sometimes also the back wall of the cinema. These are rarely used directly in conjunction with the front speakers because of problems caused by the wide spread of the typical film audience and are (usually) subjected to a delay by the replay system. This is intended to ensure that those seated near the rear of the cinema do not receive sound from the surround channels prior to that arriving from the front. This is done in order to prevent attention being drawn away from the screen. The 0.1 refers to the presence of a *Low Frequency Effects*[4] (LFE) channel which, in most cases, is used to drive a separate subwoofer. Although this system is being pressed more and more into use for music recording and composition, it was not really designed for the purpose. It can be argued that the ideal system for music would be one in which the image of the reproduced soundfield, whether recorded or a synthesised, was both homogeneous and coherent[5]. By deliberate design, Cinema Style surround does not meet these criteria although it is

[4] Also known as *Low Frequency Enhancement*

[5] In a *homogeneous* system, no direction is preferentially treated. In a *coherent* system, the image remains stable for different listening positions (though the image may change as, indeed, a natural soundfield does)

possible to circumvent this to a greater or lesser extent in the recording studio or by using computer processing[6]. It is possible, by careful tailoring of the speaker feeds, to approach the homogeneous/coherent criteria within the context of a particular systems' actual layout. However, as has long been recognized (Weiland, 1975), for it to work as well on any other system say, for instance, a home surround setup, similarity of layout is essential. Unfortunately there is no effective standardisation of loudspeaker locations for Cinema Style systems, just a rather vague set of guidelines.

An ideal system would therefore need to have matching arrays of speakers in the originating and reproducing locations or, if differences in number or position of loudspeakers were to be allowed, there would need to be a defined *transformation matrix* between the two layouts. Since, in the real world, only a minority of listeners to existing stereo systems have their speakers correctly set up it seems unlikely that the standardisation approach would work, so it makes sense to go for a transform based system. A good example of this approach is the principle behind the Ambisonic system devised in the 1970's by Michael Gerzon, Peter Fellgett, Peter Craven and Geoffrey Barton (Gerzon, 1972,1975)(Fellgett, 1975) and independently developed by Cooper and Shiga (Cooper, 1972). In the Ambisonic system, the sounds complete with their directional components are encoded vectorially in a set of spherical harmonics, of which, in the simplest fully three dimensional case, there are four. By applying a suitable *transformation matrix* (or *decoder*) to these four signals, known collectively as the *B Format* signals, almost any regular, three-dimensional array of speakers placed around a listener can be driven. The results over the whole of the sphere round the listener can approach what stereo is capable of over a mere 60^{O} arc in front of the listener. The nature of a B format encoded soundfield is such that it can be treated computationally as a single entity. This applies whether it contains a single sound source or a multiplicity of them with a multiplicity of different positions. It can be subject to transformations, such as rotation, tilting, tumbling, or mirroring exactly as if manipulating a graphical object. Many different transforms can be applied simultaneously to an arbitrarily complex B format coded soundfield using just one multiplication of the 4x1 input signal matrix with a 4x4 matrix of coefficients. The computing power required to do so in real time, even on better than CD quality audio, is easily within the reach of most modern PC's or workstations. The approach can even be used to form the basis of a *spatial computing engine* within systems intended to output binaural sound to headphones or to speakers using transaural algorithms (Malham, 1993). This approach is now in use in the Lake DSP Huron processor as it reduces the computational loading problems which, as discussed earlier, are associated with pure binaural systems which have to deal with realistic or near-realistic soundscapes. By placing all the sound sources in a B format soundfield including, if required, full complexity natural soundfields recorded with a *Soundfield* microphone (Gerzon, 1975, Farrah, 1977), the processing involved in rotating, tilting, etc. the full soundfield is much simpler than if performed at the HRTF stage. The B format signals can then be decoded to virtual speaker feed signals and only these need to be passed through HRTF's. Since these are limited to a single fixed set of HRTF's, it is possible to do all necessary operations on standard hardware, even when full head tracking is in use.

4.2.3 Ambisonic surround sound

A single sound source can be *Ambisonically* coded into B format, at least as far as its directional components go, by forming the four output signals from the single input signal thus;-

W = input signal * 0.707

X = input signal * Cos A * Cos B

Y = input signal * Sin A * Cos B

Z = input signal * Sin B

where A is the anticlockwise angle of rotation from the centre front and B is the angle of elevation from the horizontal plane. Note that the 0.707 multiplier on W is there as a result of engineering considerations related to getting a more even distribution of signal levels within the four channels when recording live sound from a Soundfield microphone.

The coding given above does not, however, provide any distance information. This must be added by controlling the various factors, such as loudness, direct to reverberant sound ratios, etc. as discussed earlier. This was not easily achievable when the technology was first developed but with current digital signal processing techniques there is little or no problem in implementing a good distance algorithm (Gerzon, 1992a).

A complete soundfield can easily be processed (say filtered, or controlled in volume) by processing all four signals equally without changing any of the directional elements. To change the directional elements, a transform must be applied to change the

[6] See, for instance, "Surround Sound Special" EQ Volume *, Issue 10, October 1997, pp 70-107 or Rumsey (1998)

original set of B format signals into a new one with modified elements. For instance, an angular rotation of the whole input soundfield to the left by an angle of C from the centre front coupled with a tilt of the soundfield by an angle D from the horizontal requires the following transformation

W' = W

X' = X * cos C -Y * sin C

Y' = X * sin C * cos D + Y * cos C * cos D - Z * sin D

Z' = X * sin C * sin D + Y * cos C * sin D + Z * cos D

where W', X', Y', Z' form the rotated and tilted soundfield. Note that this is all that has to be done, no matter how complex or simple the input soundfield is.

B format signals are not referenced to loudspeaker positions and loudspeaker layout need not be thought about at the production stage of any piece of music or soundscape which uses Ambisonically encoded sound. It is, however, important that there be a minimum number of speakers (4 speakers in a rectangle for 2-D work and 8 speakers in a cube for 3-D) although, in general, the more speakers, the better. However they must evenly spread around the central listening position, either on the perimeter of a circle for 2-d surround or on the surface of a sphere for 3-d. This latter rule can be ignored to some extent, so long as the speakers can be made to seem as if they are acoustically in the correct place by judicious use of delays and gain adjustments. The signal requirements for any particular layout and number of speakers can be met by suitable adjustments of the decoding algorithm. The decoding algorithm is the most complex part of the whole system, involving as it does the consideration of psychoacoustic factors relating to the differing directional perception mechanisms operating in different frequency bands for final optimisation. This optimisation is most appropriate for small (domestic room sized) loudspeaker arrays but for the larger arrays found in concert halls, simplified decoding algorithms are possible. Whilst this may seem contradictory, in the domestic situation, listeners are generally seated in smaller and better defined area, so more can be done to optimise performance. Two simple cases will illustrate what is needed for basic decoding.

Designating speaker positions with a letter code thus;

L - left
R - right
F - front
B - back
U - up
D - down

and assuming the pre-existing fact of 0.a horizontal square array of speakers can be driven with

LF = W + 0.5X + 0.5Y
RF = W + 0.5X - 0.5Y
LB = W - 0.5X + 0.5Y
RB = W - 0.5X - 0.5Y

and a cubic array with

LFU = W + 0.35X + 0.35Y + 0.5Z
RFU = W + 0.35X - 0.35Y + 0.5Z
LFD = W + 0.35X + 0.35Y - 0.5Z
RFD = W + 0.35X - 0.35Y - 0.5Z
LBU = W - 0.35X + 0.35Y + 0.5Z
RBU = W - 0.35X - 0.35Y + 0.5Z
LBD = W - 0.35X + 0.35Y - 0.5Z
RBD = W - 0.35X - 0.35Y - 0.5Z

For those interested in pursuing this further, an essentially complete analysis of the latest decoding technology, known colloquially as the *Vienna* technology, is in Gerzon (1992b) and also in US Patent No.5,757,927 “Surround Sound Apparatus”, also by Gerzon.

Acknowledgements

My thanks to Damian Murphy for producing the screenshots from his Waveverb program which runs on a Silicon Graphics workstation The user can design 2-D acoustic spaces using a model based on digital waveguides. It is possible to observe the resulting wave propagation and output the resultant Room Impulse Response (RIR) for a particular source/receiver pair as either a mono, stereo or 2-D B-format signal.
Tools are provided for analysing these RIRs together with a software based Ambisonic decoder allowing arbitrary signals to be convolved and processed for audition, including B-format to Binaural conversion for headphone monitoring when a full surround sound array is not practical.

References

Askew, Anthony 1981 "The Amazing Clément Ader" *Studio Sound*, Volume 23, no's 9- 11, September-November 1981

Begault, D. R. 1994 "3-D Sound for Virtual Reality and Multimedia" AP Professional, Boston.

Bennett, J.C., Barker, K., Edeko, F.O. 1985 ”A New Approach to the Assessment of Stereophonic Sound

System Performance" *Journal of the Audio Engineering Society* (*JAES*) Vol. 33, No. 5 pp. 314-321

Blumlein, Alan D. 1931 British Patent Specification 394,325 reprinted in "Stereophonic Techniques", ed. John Eargle, Audio Engineering Society, New York 1986 pp32-40

Boone, Marinus M., Verheijen, Edwin N.G. and Van Tol, Peter F. 1995; "Spatial Sound-Field Reproduction by Wave-Field Synthesis" *JAES* Vol. 43, No.12 pp1003-1012

Cooper, D.H. and Bauck, J.L. 1989 "Prospects for Transaural Recording" *JAES*, Vol 37, No. 1/2, Jan/Feb 1989 pp. 3-19.

Cooper, D.H., and Shiga, T. 1972. "Discrete Matrix Multi-channel Stereo" *JAES*, Vol. 20, No. 5, June 1972 pp. 346-360

Clark, H. A. M., Dutton, G. F., Vanderlyn, P. B.,1958, "The 'Stereosonic' Recording and Reproducing System" *JAES*, Vol 6 No. 1, pp102-117

Farrah,K.,1979,"The SoundField Microphone" *Wireless World*, November 1979 pp. 99-103

Fellgett (1975) Fellgett, Peter, "Ambisonics. Part one: general system description" *Studio Sound*, August 1975 pp20-22 & 40.

Fox, Barry 1982 "Early Stereo Recording" *Studio Sound*, Vol 24, No. 5 May 1982, p36-42

Gerzon, Michael A. 1972 "Periphony: With-height Sound Reproduction" *JAES*, Vol. 21 No. 1 Jan/Feb 1972 pp.2-10

Gerzon, Michael A. 1975 " The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound" Preprint No. 20 50th Convention of the Audio Engineering Society, London, March 1975

Gerzon, Michael A. 1975 "Ambisonics. Part two: Studio Techniques" Studio Sound, August 1975 pp24-26, 28 & 30

Gerzon, Michael A. 1992a "The Design of Distance Panpots" Preprint No. 3308, 92nd. Convention of the Audio Engineering Society, 1992

Gerzon, Michael A. 1992b "Psychoacoustic Decoders for Multispeaker Stereo and Surround Sound" Preprint No. 3406, 92nd. Convention of the Audio Engineering Society, 1992

Gerzon, Michael A. 1994 "Applications of Blumlein Shuffling to Stereo Microphone Techniques" *JAES*, Vol 42 No. 6, pp. 435-453

Murphy, DT, and Howard, DM, 1998 "Modelling and directionally encoding the acoustics of a room", Electronics Letters, Vol 34, No 9, April 1998, pp 864-865

Nielsen, Spren H., 1993 "Auditory Distance Perception in Differenr Rooms" *JAES*, Vol. 41, No. 10 October 1993 pp 755-770

Rumsey, Francis 1998 "Microphone and mixing techniques for multichannel surround sound" *JAES*, Vol46, No. 4, April 1998, pp 354-358

Sanal, Arthur J. 1976 "Looking Backward", *JAES*, Vol. 24, No. 10 December 1976

Thiele, G. and Plenge, G. 1977 "Localisation of Lateral Phantom Sources" *JAES*, Vol. 25 No. 4, April 1977, pp 196-200

Weiland, F. Chr. 1975. "Electronic Music - Musical Aspects of the Electronic Medium" Institute of Sonology, Utrecht State University