TUTORIAL ARTICLE

Approaches to spatialisation

D. G. MALHAM

Department of Music, University of York, York YO10 5DD, UK
E-mail: dgm2@york.ac.uk

This article describes some of the basic principles of acoustics and psychoacoustics related to the spatialisation of sound. It introduces recording and diffusion technologies, including binaural, stereo and surround-sound techniques.

1. SOUND IN SPACE

Sound is transmitted through air as longitudinal pressure waves. These expand outwards from their source and reduce in level as they spread. The objects they encounter will either absorb, reflect or diffract them. Usually some combination of these processes occurs, resulting in the spectrum of the sound wave changing due to interaction with the physical properties of the objects. The nature of the interaction will change with the angle of the encounter. Sound waves also interact with the air they travel through, losing higher frequencies progressively with distance as a result of absorption by water vapour in the air. Even for the simplest sound-emitting object, the purely hypothetical *point source* which emits simple spherical wavefronts, the soundfield produced in a space in which there is one or more other objects rapidly becomes very complex both spatially and timbrally. Even in *free space*, where there is nothing else to interact with, real sound sources which have extended sound-emitting surfaces have a more complicated behaviour, since the radiation of sound will vary in a non-simple manner with both position and frequency. Intuitively, we tend to expect this complex behaviour from objects which themselves are mechanically complex, such as a violin, but it is also true for simpler objects like a vibrating flat sheet. To understand why this should be so, note that sounds with wavelengths larger than the size of the body which emits them will behave much as if they had been emitted from a point source, with the result that their intensity will drop by 6 dB per doubling of distance. This is a result of the energy emitted by the source being spread over the increasing area of the expanding spherical wavefront (figure 1). In contrast, an emitting surface which is very much larger than the wavelength of the sound emitted produces a wavefront which is more like a plane wave, especially for larger ratios of surface size to sound wavelength. A true plane wave does not spread out with increasing distance, so it does not decrease in loudness with distance. In practice, true plane-wave sources do not exist, any more than true point sources. All sound emitters actually consist of a combination of the two which varies with frequency and direction. For instance, consider the difference between a busy road with many cars on it, which approximates to a line source of sound, and a quiet country road with only one car on it, which is closer to a point source. According to Begault (Begault 1993: 87), the loudness of the busy road will decrease with distance at half the rate of that of the quiet road. This is not, however, the whole story. Consider a flat rectangular sheet of metal, mounted with the short edge closest to the listener. As can be seen from figure 2, when this is struck at one end, vibrations passing along the sheet will have an increasing radius of curvature, so will approximate a plane wave when they arrive at the far edge. The sound emitted from that edge will therefore appear to be a line source, like the busy road. There will, of course, also be radiation of sound from the other edges and from the main surfaces. This will also be heard by the listener, but fractionally after the edge sound, since the vibrations in a metal sheet travel faster than soundwaves in air. These differential delays will result in significant spectral modifications caused by cancellation or reinforcement of components of the sound as a result of the differential delays. For a listener directly facing one of the two major surfaces of the sheet, these effects will be much less obvious, but the plane wave emitted from the surface will not be at right angles to the surface but tilted (figure 3) due to the finite speed at which the wavefront crosses the sheet. In this simplified analysis, the effects of the discontinuity at the sheet edges has been ignored and it is assumed that the wave in the sheet terminates at the edges. In practice, the wave is reflected back into the sheet which further complicates the emission behaviour. This effect was exploited in the early reverberation systems known as *plate reverbs*, which originally consisted of a suspended sheet of steel, about 2×3 m in size, later superseded by a much smaller sheet of gold foil. The device would be fitted with transducers for injecting

Organised Sound **3**(2): 167–77 © 1998 Cambridge University Press. Printed in the United Kingdom.

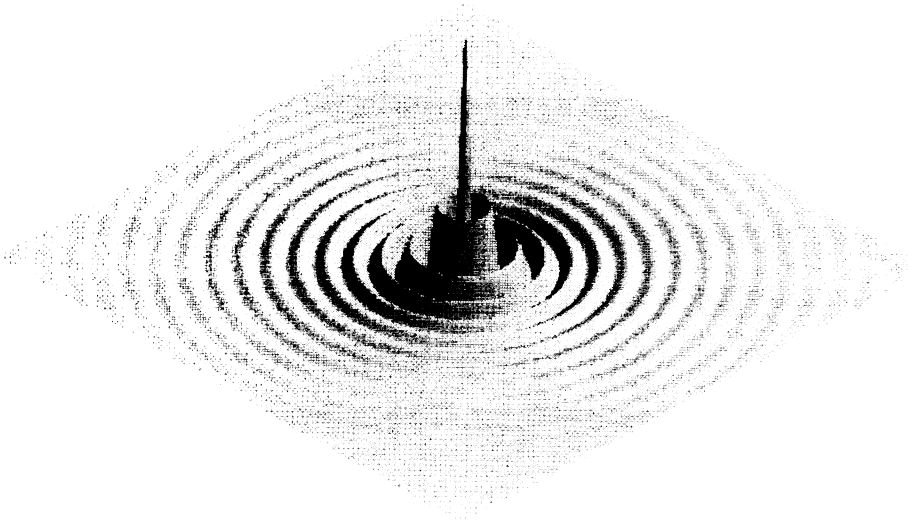

Figure 1. Simulation of sinewave radiation from a near point source. Scaled logarithmically to match human loudness perception.

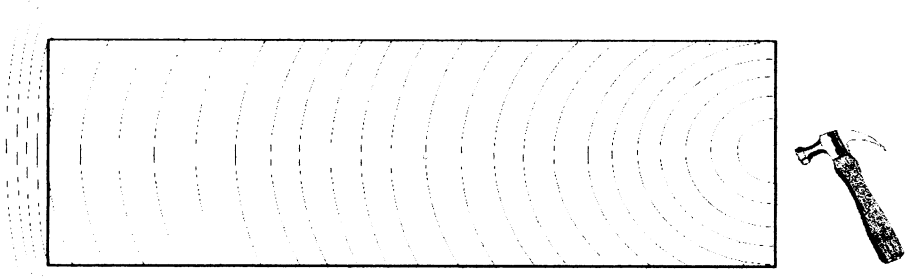

Figure 2. A flat sheet struck on one edge radiates like a line source from the far edge.

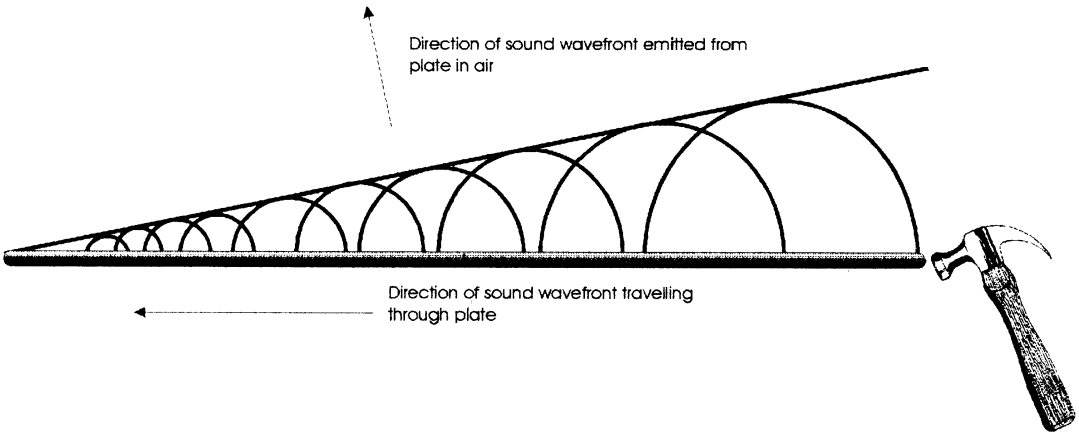

Figure 3. A flat radiator struck at one end will produce a tilted wavefront from its surface due to the differences in the speed of sound between the sheet and the air.

the audio and microphones or some other means of picking up the resulting vibrations in the sheet. Careful adjustment of the tension on the sheet and the position of an acoustic damper resulted in a controllable artificial reverberation device which produced quite realistic reverberation, albeit not corresponding to any real acoustic environment.

The preceding discussion is intended to show how complex the behaviour of any real-world sound source is, both spatial and timbrally. When attempting to make a sound object appear 'real', i.e. plausible, it is esential to bear these facts in mind when employing spatial manipulations. Simplistic spatial representations are unlikely to produce sound objects which sound real in this sense. A lack of spatial complexity is one of the reasons for it being so difficult to develop fully effective synthesised versions of acoustic instruments, or even reproducing recordings of real instruments so as to be indistinguishable from the original.

At this stage it is worth noting that whilst it is common practice to employ visual analogues when dealing with sound and hearing, they should be treated with extreme caution since the differences between light and sound far outweigh the similarities. Unfortunately, this is common amongst both composers and engineers. For instance, Varese often allows visual analogues to dominate his thinking, or at least his explanations of the way he conceived the structure of a piece. Here, when discussing *Integrales* he talks of geometrical figures being projected onto plane surfaces, movements of the two relative to each other resulting in

> . . . (the projection of) an apparently unpredictable image of a high degree of complexity; moreover, these qualities can be increased subsequently by permitting the form of the geometrical figure to vary as well as its speeds . . . (Varese 1959: 193)

The main problems with allowing such a visual dominance of our thinking about sound in space are that, firstly, although the audible part of the acoustic spectrum spans some ten octaves or so, the visible part of the electromagnetic spectrum covers only around one. Secondly, most of the structures with which visible light interacts are vastly larger in extent than the wavelength of the light itself. Contrast this with the situation for sound where, as has already been discussed, wavelengths are frequently larger than the structure involved in emission or even propagation.

As a result of these differences, the simpler methodologies which might be employed in the synthesis of soundfields are not as effective as in other fields, for instance radio wave propagation studies or RF antenna design. In those areas of work, the bandwidth and structure sizes are usually far more narrowly constrained than in acoustics. Despite this, however, the significant computational penalties of more accurate methods, such as *finite element* or *boundary element* modelling (Begault 1994: 187), mean that image modelling and ray tracing are still widely used, especially when working in real time. Fortunately, since the hearing mechanism has in fact evolved in circumstances where it frequently needs to deal with ambiguous or incomplete information, much can be done without constructing acoustically completely accurate soundfields.

2. HEARING MECHANISMS

Spatial perception of sound is based on the interpretation of a number of *cues* which are extracted from the soundfield surrounding the listener. As noted above, it is possible for these cues to be ambiguous or conflicting largely because of the complexity of most soundfields. This is especially common in artificial soundfields, whether synthesised or recorded, but can also happen 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 assigns a ranking to each of the cues according to their apparent ambiguity, and it is this factor which enables us to construct usable but simplified 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 it is possible to do so over a small area and to approximate it over a larger one. By concentrating on a subset of the possible cues and trying to make them as unambiguous as possible, relatively simple equipment can produce artificial acoustic environments with acceptable performance, at least in terms of 'naturalism'. Of course, for compositional purposes, the ambiguities may be even more interesting, but that is largely beyond the scope of the current paper.

We can describe the main cues used to determine the angular position of a sound source as follows (figure 4), although there are maybe other, more subtle mechanisms:

(1) 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* (ITD). The minimum difference in arrival times between the two ears which can be perceived is dependent on the nature of the

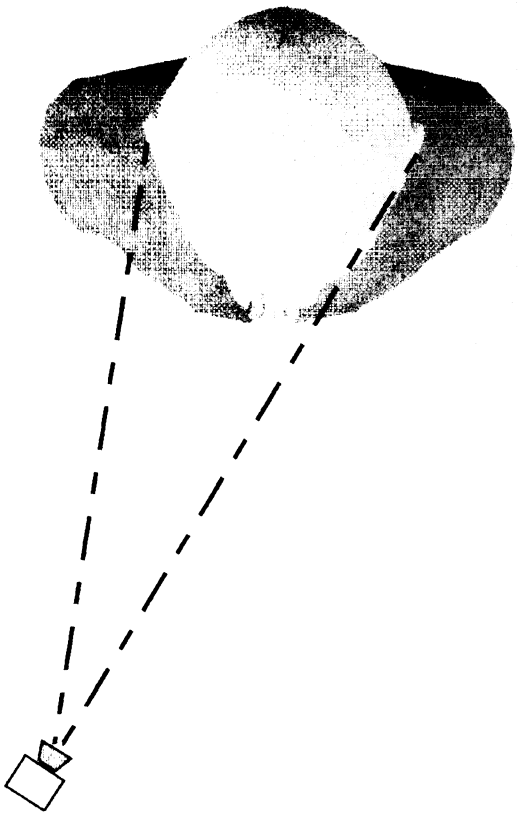

Figure 4. The main spatial perception mechanisms are based on path differences between the sounds which reach the right and left ears.

sound, varying between 5 μs and 1.5 ms (Begault 1994: 44).

(2) 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 *interaural level difference* (ILD).

(3) The shape of the head and the external part of the ears imparts a frequency-dependent response which varies with sound position and which is, in general, different for each ear. Although this is often referred to as the *head-related transfer function* (HRTF), strictly speaking the HRTF also includes the ILD and the ITD. For this reason, it will be referred to as the *head-related frequency response* (HRFR). For positions where ILDs or ITDs give ambiguous or nonexistent differences between ear signals (such as median plane signals) or where the listener has little or no hearing in one ear, this is the main positional sensing mechanism where head movement is not involved. It is also one of the two main mechanisms for distinguishing frontal sound sources from rear ones.

(4) We have the ability to change the position of our head in such a way that we can minimise the ITD, ILD and the difference between the HRFRs 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 interaural differences are increasing or decreasing for a particular direction of head movement.

The main cues for determining the distance of a sound source are:

(1) *The ratio of direct to reverberant sound.* In a reasonably reverberant environment, the energy in the reverberant field stays more or less constant for all combinations of listener/source positioning, which means that for a given source level the reverberation loudness remains the same, whereas the source loudness drops off with increasing distance. (It is this factor in particular which makes it difficult to place a 'sound object' closer than the nearest loudspeaker in a diffusion system.)

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

(3) *Progressive attenuation of higher frequencies with distance.* This is due to absorption by moisture in the atmosphere.

(4) *The reduction of loudness with distance.* This is due to the increase in the area of the wavefront as it moves away from the source.

The interpretation of the last two cues is heavily dependent upon acquired knowledge of both the spectra and loudness of the sound source, something which should be considered when using heavily manipulated or wholly artificial sound objects. Loudness as a distance cue is, in particular, 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.

We should note here that these are not the only ways that the body perceives sound and indeed other perceptual mechanisms can also provide directional cues. Unfortunately, because of the difficulty of working experimentally on, say, chest cavity pickup or bone conduction mechanisms, little work has been done on these means of perception and their directional discrimination capabilities. Instead, because of the relative ease with which headphone-based measurements can be made, almost all the major studies of directional hearing have concentrated on

headphones. Informal experimentation has, however, shown that such nonaural sound perception mechanisms should be taken seriously. In particular, I believe that the chest cavity may play a role 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 be true for headphone presentation. This may have serious implications for diffusion systems where the bass is presented over a limited number of subwoofers or for replay of electroacoustic works over headphones. Additionally, it is worth considering that the mechanisms of directional hearing described above may well only be components of a holistic, integrated directional perception facility.

3. SOUND SPATIALISATION TECHNIQUES

Sound can be spatialised in essentially three different ways.

(1) The system can attempt to provide signals *directly at the ears* similar to those which would have occurred had there been real sound sources in the intended positions. This is usually but not always done via headphones.
(2) A loudspeaker system can be designed to produce, in an extended space, a precomposed soundfield which, upon correct interpretation by the listener, will produce the spatial results desired by the composer.
(3) The performance space itself can be used to spatialise sound using a *loudspeaker orchestra* placed within it and controlled by a diffusion mixing desk operated by a suitably trained performer or by the composer in person.

3.1. Headphone-based systems

In this section we will consider systems that are intended for headphone listening and those which use the same approach but are modified so that loudspeakers can be used.[1] These are generally referred to as *binaural* systems.

This is perhaps the most obviously 'correct' way of approaching the problem of full three-dimensional (3D) spatialisation of sound. Exact duplication of what the ear would hear in a natural situation should produce the best reproduction. In fact, under a certain limited set of circumstances there can theoretically be no better or closer approach to real-world performance. There are, however, some very real problems.

For recording natural soundfields, binaural systems use *dummy head microphones*. These systems are constructed in the form of a model head with microphones inserted in the ears, although this may be simplified to a sphere, as in the Scheops device, or even a circular disk of material with microphones mounted at either side, as used by the BBC. This approach was, as far as can be ascertained, first adopted in the 1920s by Dr Harvey Fletcher and his team at Bell Labs (Sanal 1976: 832) and has been used in various forms ever since. When a synthetically constructed soundscape is produced using this method, each sound source must be treated using the appropriate HRTFs for the *source to left ear* and the *source to right ear* paths. The required HRTFs, which naturally have to be different for every 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' HRTFs (*generalised*),
(3) measured on a dummy head, which will itself usually have been generalised from the measurement of many individuals, or
(4) calculated from a mathematical model (*synthesised*).

The individualised approach is the most successful and is potentially capable of producing reality-equivalent results, but the difficulty of measuring every possible HRTF for each user of a system means that this is currently only used in research systems. For most situations, generalised HRTF sets are used, but unfortunately this approach has a serious deficiency. 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 HRFRs can be significant, especially at higher frequencies. Because of the importance of these cues for front–back discrimination, *front–back reversal errors* become much more common. Sometimes this results in complete failure to perceive any sounds as being at the front (or rear). The problem can be ameliorated if the position of the listener's head can be tracked and used to select the appropriate set of HRTFs. If this is done, head rotation-based cues can be used for front–back discrimination, greatly reducing the number of such errors. With head tracking, even seriously mismatched HRTFs can become usable, although the effect is only present during head movements. It can be quite disconcerting for the listener if the image is continually swapping between correct presentation

[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.

(during movement) and incorrect (when still). It should be noted that even when using personalised HRTFs, problems occur if the system does not provide head tracking, since this results in the soundfield being fixed with respect to the head, rather than the exterior world. These problems are worse for, say, recordings that might be listened to from walkman-type systems, where the listener is moving, but may not be so serious for situations where the listener's head is normally less likely to be mobile, such as when working with a computer.

So far, we have been discussing the use of such binaural systems in a fairly theoretical manner. In practice, there are further significant limitations to this approach. The computational burden of the binaural approach is high, even for a single sound source. HRTFs are usually stored and processed as *impulse responses* typically comprising, at the commonest sample rate of 44.1 kHz, some 512 samples for each of the HRTF's source–ear paths, although various data reduction techniques can be applied (Begault 1994: 158) to reduce these numbers. The application of these HRTFs to the sound from each source is done with *finite impulse response* filters, so each sample of any one sound source will require some 1,024 multiple-accumulate cycles in order to produce the two ear signals, although again there are techniques for reducing this burden. As long as the sound imagery remains simple, this does not present a significant problem to 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 for relatively simple synthesised sound images, such as one finds in computer games. As soon as the imagery starts to get close to that of a real-world soundfield, the computational burden becomes excessive preventing their generation in real time, even using massively parallel supercomputers. The extra burdens of manipulating and interpolating between multiple sets of HRTFs result in this limit being reached much earlier when head tracking is in use. For the foreseeable future, soundfields of near real-world complexity, at least those synthesised using the direct HRTF approach, will only be realisable offline, 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. This results directly from the fact that there is no known way of changing the HRTFs applied to each sound source during the recording for new ones corresponding to the changed soundfield/head orientation, because there are simply too many unknown parameters. The same limitation applies to the output from offline full-complexity HRTF-based soundfield synthesis programs. This limitation can be circumvented by precomputing a high-complexity background soundfield against which a smaller number of active sources can be positioned. Using precomputation, a number of soundfields containing the same sonic sequences but with different orientations can 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, although it does impact significantly on the data storage requirements of the system. With the large hard disks used by modern computers and the appearance of large capacity, cheap storage media such as DVD, this may be less of a consideration.

Binaural material 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 separated. When binaural material is presented over loudspeakers, the right ear receives not only the signal emitted from the right-hand speaker, but also the one intended for the left ear emitted from the left-hand speaker. The same thing happens for the opposite ear path. It is possible to cancel a significant portion of this crosstalk by 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. Crosstalk cancellation systems require the orientation and location of the listener relative to the speakers to be precisely known for optimum operation. This is unlikely to be the case outside the laboratory, but careful design and a suitable set of compromises can result in very usable results, as evidenced by the number of two-speaker 3D surround-sound options now available on soundcards in PCs, in televisions and in other consumer audio devices. These techniques only work well over a very small area and so such systems cannot easily be applied in the concert hall, but binaural coding can nevertheless play a significant role in the composer's armoury of spatialisation methodologies. However, for a composer or performer wishing to present an electroacoustic work with well-defined spatial elements to a large audience, some form of loudspeaker-based diffusion is the only currently practicable approach. Loudspeakers also have the advantage that they stimulate non-ear-based sound perception mechanisms (such as body cavity resonances), as well as ear-based ones.

3.2. Loudspeaker-based systems

There are a number of possible loudspeaker techniques that can be used for spatial reproduction of electroacoustic works in the concert hall. Currently,

the most popular is the use of an *orchestra* of loudspeakers which are placed around the performance venue so as to allow sound diffusion artists to explore the relationships between the acoustics of the space and the sound materials of the performed composition. Loudspeaker orchestras such as *Beast*[2] or the *Gmebaphone*[3] feature large numbers of loudspeakers, usually with a wide variety of characteristics. A skilled diffusion artist will place the speakers so as to excite many of the different acoustic properties of the performance space, yielding near or distant sound images by employing the variety of loudspeaker distances available, differing colourations through the use of arrays of tweeters, bass bins or mid-range-only drivers, and the ability to vary the *reality* of the sound images from real, where they come from a single loudspeaker, to totally unreal, when there is a large proportion of the orchestra in use. This approach is entirely appropriate for many electroacoustic works, but it does represent a continuation of the aesthetic of the separated composer and performer. This aesthetic may not be appropriate for all electroacoustic compositions, especially since one of the factors which separates composers of electroacoustic works from those of purely acoustic ones is the far greater degree of control which they can, if they wish, exercise over the final sound of their piece in performance. In order for this control to be available not just over the timbral and temporal aspects of a piece but also over the spatial ones, other approaches need to be considered.

In order to make available this level of preperformance determination of the spatial elements in a piece, systems which in some way attempt to create the illusion of a real soundfield directly within the listening space need to be used. The term *illusion* is used advisedly since, despite claims to the contrary, it is at least impractical with current technology to reproduce fully a predetermined 3D soundfield of reality-equivalent complexity over any significant area, owing to the large number of information channels that would be necessary.[4] Nevertheless, there are several ways in which a limited number of channels can be used to create a subset of the soundfield that contains a set of cues of a sufficiently unambiguous nature for the illusion presented to the listener to be acceptable. In the simplest 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,[5] or a partial, usually horizontal plane only, surround image. Note that here we are dealing with *transmission* channels, not with *reproduction*, i.e. loudspeaker drive, signals. In some systems, the loudspeaker drive signals may be significantly larger in number than transmission channel signals. For the purpose of this paper, we will limit discussion to three main types of system, namely *stereo*, *Cinema Style* surround and full 3D surround based on *Ambisonic* technology.

3.2.1. 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 in normal usage stereo is taken to mean 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. In order to simplify matters, we will only discuss two-channel, two-speaker systems here, although occasionally stereo systems are extended to two or three channels driving three or more loudspeakers. 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 at earlier, it is intended to cover only a limited *sound stage*, usually in front of the listener.

There are two main ways of producing a stereo image. They rely on the use of either amplitude differences or time differences between the two speakers. The first is by far the most common approach, being embodied in the ordinary *pan* function, as well as the many recordings made with *coincident pairs* of directional microphones as their main or even only stereo source. 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 700 Hz), an amplitude difference of between 15 and 19 dB is sufficient to move the sound fully into the loudest speaker, assuming a subtended angle of 60 deg between the speakers 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*,

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

[2] Birmingham Electroacoustic Sound Theatre: http://www.bham.ac.uk/music/ea-studios/BEAST/

[3] Groupe de Musique Electroacoustique de Bourges: http://www.gmeb.fr/

[4] 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 even a small (2 m diameter) spherical volume 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, Verheijen and Van Tol (1995).

[5] Although it dates back to much earlier than this (Askew 1981, Fox 1982).

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, Barker and Edeko 1985: 315). Above 700 Hz the apparent angular source location produced by this rule increases, although it has been found (Clark, Dutton and Vanderlyn 1958: 108) that multiplying the $(L - R)$ component by 0.7 above this frequency can partially compensate for this. This is a rather simplified application of a more complex, frequency-dependent directional coding rule, for which a more comprehensive exposition is available in Bennett *et al*. (1985). Even though this requirement has been known since Blumlein's work in the 1930s (Blumlein 1931), this frequency-dependent rule is rarely used. Fortunately, sufficiently strong cues are produced for sounds within the lower band using the law of sines for most people to obtain good results from stereo even without this *stereo shuffling*. 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 *et al*. 1958: 109), and it is the difference between these two mechanisms which results in the difference in apparent source location. A comprehensive coverage of this is also given in Gerzon (1994).

Stereo has a number of limitations, the main ones being:

(1) 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 deg are generally unacceptable;
(2) the increasingly poor performance as the listener moves off-axis; and
(3) difficulties with image stability under head rotation such that in the limit, where the listener is parallel to the speakers rather than facing them, it is impossible to generate stable central phantom images (Thiele and Plenge 1977).

3.3.2. Cinema Style surround

In Cinema Style surround systems, as typified by *Dolby 5.1*, additional channels are added to those in the standard 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.[6] These are rarely used directly in conjunction with the front speakers because of problems caused by the wide spread of the typical film audience. The signals going to the surround speakers are usually subject to a delay by the replay system. This is intended to ensure that the attention of those seated near the rear of the cinema is not drawn away from the screen by hearing sound from the surround channels prior to that arriving from the front. The 0.1 (in 5.1) refers to the presence of a *low-frequency effects*[7] (LFE) channel which may be used to drive a separate subwoofer. Although, for commercial reasons, Cinema Style systems are increasingly being pressed into use for music recording and composition, they are not really designed for the purpose. It can be argued that the ideal system for recorded music would be one in which the image of the reproduced soundfield, whether recorded or synthesised, was both homogeneous and coherent.[8] By deliberate design, Cinema Style surround does not meet these criteria, although it is possible to circumvent this to a greater or lesser extent in the studio or by using computer processing.[9] For the composer of electroacoustic music, the somewhat simplistic approach exemplified by these Cinema Style systems may be extended, by careful tailoring of the speaker feeds, to one in which an approach to the homogeneous/coherent criteria is made within the context of a particular system's actual layout. However, it has long been recognised (Weiland 1975) that for spatialisation based on this approach to work well in different systems, for instance that of a concert hall instead of the composer's studio, similarity of layout is essential. This would require standardisation of loudspeaker locations in composition and performance spaces or, at the very least, the careful description by a composer of the loudspeaker array which is to be used for any particular piece. This implies in turn that performance venues should be both willing and able to comply with the composer's wishes.

From this it can be seen that for a composer working in a studio to have good control over the spatial elements in performance, one possibility would be to have matching arrays of speakers in the

[6] In the recently announced *Dolby EX 6.1* channel system there are both side and rear surround channels.
[7] Also known as *low-frequency enhancement*.
[8] In a *homogeneous* system, no direction is preferentially treated. In a *coherent* system, the image remains stable for different listener positions (though the image may change as, indeed, a natural soundfield does).
[9] See, for instance, 'Surround Sound Special', *EQ*, Issue 10, October 1997, pp. 70–107 or Rumsey (1998).

composition and performance locations. Alternatively, if it is required that differences between the two in either number or position of loudspeakers be accommodated, a *transformation matrix* between the layouts needs to be defined. The wide variation between performance spaces makes it unlikely that the standardisation approach would be viable in most cases, so it makes sense to go for a transform-based system unless the work is only to be performed in a specific location. A good exemplar of this approach is the Ambisonic system devised in the 1970s by Michael Gerzon, Peter Fellgett, Peter Craven and Geoffrey Barton (Gerzon 1973, 1975, Fellgett 1975) and independently developed by Cooper and Shiga (Cooper and Shiga 1972). In the Ambisonic system, the sounds and their directional components are encoded vectorially in a set of spherical harmonics of which, in the simplest fully 3D case, there are four. These signals are known collectively as the *B Format* signals. By applying a suitable *transformation matrix* (or *decoder*) to these four signals, almost any regular, 3D array of speakers can be used. The results over the whole of the sphere around the listener can be nearly as good as stereo is capable of in front of the listener. The nature of B format is such that, whether it contains a single sound source or a multiplicity of them in a multiplicity of different positions, it can be treated for computational purposes as a single entity. It can be subject to transformations, such as rotation, tilting, tumbling or mirroring, using similar mathematical operations to those used to manipulate a graphical object. Many different transforms can be applied simultaneously to an arbitrarily complex B format-coded soundfield using just one multiplication of the 4×1 input signal matrix with a 4×4 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 contemporary PCs or workstations. The approach can even be used to form the basis of a *spatial computing engine* within a system 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 to reduce the computational loading problems which are associated with pure binaural systems employing realistic or near-realistic soundscapes. By placing all the sound sources in a B-format soundfield including, if required, complex natural soundfields recorded with a *Soundfield* microphone (Gerzon 1975, Farrah 1979), the processing involved in manipulating the soundfield is much simplified compared to that required 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 HRTFs. As this method only employs a single fixed set of HRTFs, it is possible to do all necessary operations on standard hardware, even when full head tracking is in use.

3.2.3. Ambisonic surround sound

A single sound source can be *Ambisonically* encoded into B format by forming the four output signals from the single input signal thus:

$$W = \text{input signal} * 0.707,$$

$$X = \text{input signal} * \cos A * \cos B,$$

$$Y = \text{input signal} * \sin A * \cos B,$$

$$Z = \text{input signal} * \sin B,$$

where A is the anticlockwise angle of rotation from the centre front and B is the angle of elevation from the horizontal plane. The 0.707 multiplier on W is a result of engineering considerations related to achieving 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 distancing algorithm (Gerzon 1992).

By changing all four signals equally, a complete soundfield can easily be processed (say filtered, or controlled in volume) without disrupting any of the directional coding. To change the directional elements, a transform must be applied to change the 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 B-format 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. This is all that is required and the total number of sound sources in the input soundfield is irrelevant.

Note also that these B-format signals make no reference to loudspeaker positions. In fact, no particular loudspeaker layouts need be considered when dealing with Ambisonically encoded sound. There are only two main criteria which need to be borne in mind. Firstly, there needs to be a certain minimum number of speakers for effective presentation. For 2D systems the requirements are four speakers in a rectangle, and

for 3D systems, eight speakers in a cuboid is the minimum. In general, the more speakers, the better a system will perform, so long as they are evenly distributed around the central listening position. This latter rule can be ignored to some extent, so long as the speakers can be made to appear as if they are acoustically in the correct place by judicious use of delays and gain adjustments. The drive signal requirements for any particular layout and number of speakers can be met by suitable adjustments of the decoding algorithm. The design of the decoding algorithm is possibly the most complex part of the whole system and as such will not be dealt with at any length here. For an essentially complete analysis of the latest decoding technology, known colloquially as the *Vienna* technology, see Gerzon (1992) and US Patent No. 5,757,927, 'Surround Sound Apparatus', also by Gerzon. The Vienna technology is perhaps more appropriate to domestic-scale listening and, as I have indicated in other papers, some compromises have to be made for systems which need to work over the larger areas involved in concerts (Malham 1993). One of the most comprehensive recent treatments of decoding technology was presented by Jerome Daniel at the September 1998 Audio Engineering Society Convention (Daniel, Rault and Polack 1998).

For simple, experimental evaluation, the following rules can be followed:

- Choose an even number of speakers and arrange them as pairs at opposite ends of a line passing through the centre point of the listening area.
- Spread the speakers as evenly as possible around the centre point. Uneven spread affects both the positional accuracy and the extent to which there is an unwanted variation in the loudness of a sound as it moves around the space.
- Feed the speakers with a signal combining W and the directional components, X, Y and Z, each multiplied by the factors given by the following rules:

  W signal multiplied by 1 for small areas and 1.414 for large ones,

  X signal multiplied by $\cos A * \cos B$,

  Y signal multiplied by $\sin A * \cos B$,

  Z signal multiplied by $\sin B$,

  where, as in the generation of B format, A is the anticlockwise angle of rotation from the centre front of the loudspeaker and B is its angle of elevation from the horizontal plane.
- Experiment with the level of the W signal, since reducing it will improve the 'focus' for a central listener at the expense of noncentral ones, and increasing it will have the opposite effect.

Ambisonics in this form, known as *first order*, is not able to provide signals which are limited to single speakers. Sound images are produced by the cooperation of many speakers and whilst this produces one of the great advantages of Ambisonics – the near-complete disappearance of the speakers as perceived sources of sound – it also means that if it is desired to provide loudspeaker orchestra diffusion simultaneously, this needs to be done via a separated diffusion mix (although the same speakers can often be used). As we move to higher orders of system, with more channels in the B format (nine in second order, fifteen in third order (Gerzon 1973)), this will be less and less of a problem.

4. CONCLUSIONS

In this paper some of the perceptual and technical issues involved in the spatialisation of audio have been considered. From the development in the nineteenth century of an ability to present sounds remotely (in either space or time), music has moved, at least in some respects, through more than a full circle. The path stretches from the millennia when it was always part of a three-dimensional acoustic environment, though spatial elements were then rarely a deliberately exploited part of the music, through the early remote presentations with their removal of most of the spatial elements in the music, and up to the present era when recording engineers are finding ways of more accurately presenting the spatial aspects of the musical experience to a listener at home and composers are finding new ways of using space within their music. We are still in the process of learning how the ear/brain perceives sound, especially sound in space, and there is a long way to go technically before we are able to produce fully reality-equivalent systems. We are therefore currently in no position to define or describe the optimum way of spatialising sound either for reproduction or composition purposes. Indeed, whilst it may be possible to do so for reproduced music if the optimum experience at home is defined as one which matches that in the concert hall, the optimum system for composition purposes must remain always a decision of the composer, to be made on musical, not technical, grounds.

REFERENCES

Askew, A. 1981. The amazing Clément Ader. *Studio Sound* **23**(9, 10, 11).

Begault, D. R. 1994. *3-D Sound for Virtual Reality and Multimedia.* London: Academic Press.

Bennett, J. C., Barker, K., and Edeko, F. O. 1985. A new approach to the assessment of stereophonic sound system performance. *Journal of the Audio Engineering Society* **33**(5): 314–21.

Blumlein, A. D. 1931. British Patent Specification 394,325. Reprinted in John Eargle (ed.) *Stereophonic Techniques*, pp. 32–40. New York: Audio Engineering Society, 1986.

Boone, M. M., Verheijen, E. N. G., and Van Tol, P. F. 1995. Spatial sound-field reproduction by wave-field synthesis. *Journal of the Audio Engineering Society* **43**(12): 1,003–12.

Clark, H. A. M., Dutton, G. F., and Vanderlyn, P. B. 1958. The "Stereosonic" recording and reproducing system. *Journal of the Audio Engineering Society* **6**(1): 102–17. As reprinted in John Eargle (ed.) *Stereophonic Techniques*, pp. 81–96. New York: Audio Engineering Society, 1986.

Cooper, D. H., and Bauck, J. L. 1989. Prospects for transaural recording. *Journal of the Audio Engineering Society* **37**(1/2): 3–19.

Cooper, D. H., and Shiga, T. 1972. Discrete matrix multichannel stereo. *Journal of the Audio Engineering Society* **20**(5): 346–60.

Daniel, J., Rault, J.-B., and Polack, J.-D. 1998. Ambisonics encoding of other audio formats for multiple listening conditions. *Preprint No. 4795*, *105th Audio Engineering Society Convention*, September 1998 (corrected version available by contacting the authors at Centre Commun d'Etudes de Tele-diffusion et Telecommunications, Cesson Sevigne, France).

Farrah, K. 1979. The Soundfield microphone. *Wireless World*, November issue, pp. 99–103.

Fellgett, P. 1975. Ambisonics. Part one: general system description. *Studio Sound*, August issue, pp. 20–2 and 40.

Fox, B. 1982. Early stereo recording. *Studio Sound* **24**(5): 36–42.

Gerzon, M. A. 1973. Periphony: with-height sound reproduction. *Journal of the Audio Engineering Society* **21**(1): 2–10.

Gerzon, M. A. 1975. The design of precisely coincident microphone arrays for stereo and surround sound. *Preprint No. 20*, *50th Convention of the Audio Engineering Society*, March 1975, London.

Gerzon, M. A. 1992. Psychoacoustic decoders for multispeaker stereo and surround sound. *Preprint No. 3406*, *92nd Convention of the Audio Engineering Society*.

Gerzon, M. A. 1994. Applications of Blumlein shuffling to stereo microphone techniques. *Journal of the Audio Engineering Society* **42**(6): 435–53.

Malham, D. G. 1993. 3-D sound for virtual reality systems using Ambisonic techniques. Invited paper presented at the *VR93 Conference* organised by Meckler Ltd, April 1993, London.

Rumsey, F. 1998. Microphone and mixing techniques for multichannel surround sound. *Journal of the Audio Engineering Society* **46**(4): 354–8.

Sanal, A. J. 1976. Looking backward. *Journal of the Audio Engineering Society* **24**(10): 832–3.

Thiele, G., and Plenge, G. 1977. Localisation of lateral phantom sources. *Journal of the Audio Engineering Society* **25**(4): 196–200.

Varese, E. 1959. *Poeme electronique Le Corbusier*. Paris: Editions de Minuit. Quoted in *Edgar Varese* by Fernand Oullette, p. 83. London: Calder and Boyars, 1973.

Weiland, F. C. 1975. *Electronic Music – Musical Aspects of the Electronic Medium*. Institute of Sonology, Utrecht State University (internal publication).