Proceedings of the Institute of Acoustics

FULL REALITY SURROUND SOUND - THE CHALLENGE OF THE FUTURE

D.G. Malham Music Technology Group, Department of Music, University of York,

1. INTRODUCTION

Since shortly after the remote transmission of sound (either spatially or temporally) became a possibility in the nineteenth century, ways have been sought to extend the transmission capabilities so that the systems could deal not only with the frequency and amplitude based temporal elements which make up a sound event but also the spatio-temporal ones. Although much progress appears to have been made with the development of HRTF based binaural systems, 5.1 (and other cinema style systems), Ambisonics, holophonics, Wave-Field Synthesis and Hyper-Dense Transducer Array technology, we are still a long way from having achieved truly *reality mimicking* performance. This paper examines some of the issues involved as well as the history of our efforts to move towards this and uses that background to suggest possible ways forward.

2. FULL REALITY SOUND

2.1 Definition

For those people lucky enough not to suffer from any significant impairment of either visual or aural senses, the visual sense tends to be regarded as more important than the sense of hearing for the perception of what we call, for want of another word, reality. When we think of a beautiful moorland scene, it is the shape of the hills and the purple of the heather, with the white dots of sheep scattered within it that we pay most attention to. The sound of the wind moving through that heather, the calling of the sheep and the plaintive cry of the curlew tend to be in the background. Yet imagine that scene without the sounds. Without them, it ceases to be a living, breathing part of nature, losing much of its depth and becoming just another picture.

This loss of "reality" which happens when the sound is lost, or is otherwise inadequately presented, is even more pronounced when the scene is an artificial rendition of reality. It is only necessary to consider what happens when the sound is lost on a film or a television set to see the truth of this assertion. The corollary to this is that we need to ensure that the degree of *reality equivalence* which we achieve is as high as possible. For a system to produce *Full Reality Sound*, it is obviously necessary for all departures from reality equivalence to be below the relevant thresholds of perception.

Our hearing is the only one of our five senses which is truly capable of providing us with fully three dimensional information about remote, ie non-contact, events. We are able to perceive the positions of acoustic sources wherever they are in the space around us and whether they are moving or stationary. We can even estimate the distance of the sound producing object as well as getting some idea of its nature and size. It is also worth noting that for life forms that have directional hearing, the 3-D sonic environment is one from which they cannot escape since, unlike sight, hearing cannot be cut off by anything as simple as night or a blindfold.

Given the importance of spatial hearing to our perception of the world, it is evident that it is vital to render correctly the spatial aspects of any audio system which is part of a reality mimicking system[1]. It is, however, essential that we understand what we mean by "render correctly" because this is dependent upon the purpose of the system. The primary variable governing this is the number of people to which the sound needs to be presented simultaneously. For a single individual, rendering the sound scene correctly simply means presenting them with the same set of audio precepts that they would have gained had the situation been real (a *one-one* mapping). Presenting full reality sound to a multiplicity of people simultaneously poses the question of what we mean by reality. Do we treat each person as a separate observer of the scene, with their own location and independent perceptions (*many-many* mapping), or is it more appropriate to provide them all with the same perceptions simultaneously (*one-many* mapping)? It should be noted that one-one and many-many mappings can be regarded as being, in some sense, equivalent. Either is a legitimate option and the choice of which to use should be made on aesthetical grounds although economics, and to a lesser extent technological considerations, usually dictate which route will be attempted. The word "attempted" is used advisedly, since it is the author's contention that we are still a significant way from achieving anything like full reality equivalence for either one-many or many-many mappings, although the one-one situation is somewhat better.

2.2 Spatial Hearing.

In order to assess what has to be done to achieve Full Reality Sound, it is worth reviewing briefly the known mechanisms that humans use for the spatial perception of sounds. It should be pointed out here that this list should not be regarded as exhaustive as it is by no means clear that our current level of knowledge can be regarded as complete.

- Arrival times. The time of arrival at the ears of the wavefront emitted by a sound object, 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. The minimum difference in arrival times between the two ears which can be perceived is dependent on the nature of the sound, varying between 5 microseconds and 1.5 milliseconds (Begault, 1994:44)

- Level differences. Sound level differences between the two ears. Sound from a source to the left of the head, for example, will arrive directly at the left ear, but will be diffracted around 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 obstructing 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*).

- Frequency response differences. The shape of the head and the external part of the ears imparts a frequency dependant 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 HRTF's also include ILD and ITD's. For this reason, it will

[1] I use the term "reality mimicking" here, rather than Virtual Reality, since VR has specific connotations which do not, I believe, cover things like the concert situation, where we might still wish to achieve full reality sound.

be referred to as the *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) 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.

- Head movements. Our ability to change the position of our head in such a way that we minimise the ITD, ILD and the difference between the HRFR's at the two ears. This difference minima is, or should be, at the point at which we are directly facing towards (or away from) the sound source. Note that this is also the basis of the other (and possibly main) mechanism for *front-back discrimination* This is dependent upon observing whether inter-aural differences increase or decrease for a particular direction of head movement

The preceding mechanisms deal largely with the angular positions of sound objects Our ability to determine the distance of a sound source relies upon the following cues;

- Reverberation. 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)

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

- Loudness. Loudness reduces with distance due to the spreading of the wavefront.

- Loudness variation rate. Close sources show larger variations in loudness with listener movement and/or head turning, since the distance change is a larger fraction of the overall distance to the source than for more distance sources.

- High frequency losses. Higher frequencies are progressively attenuated with distance, largely due to absorption by water molecules in the atmosphere.

- Air related distortion. For high sound pressure levels, the increase in distortion with distance from the source, which results from the differing speeds of propagation of the positive and negative going peaks of the pressure wave (Czerwinski,2000), is a possible extra cue to source distance.

The interpretation of the last four 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 (Nielsen, 1993) have shown errors of more than two to one when subjects were asked to estimate the distance of a sound source.

It should be noted here that these are not the only ways that the body perceives sound and some of these other perceptual mechanisms 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 published 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 headphone presented information. Informal experimentation by the author has, however, shown that such non-aural sound perception mechanisms should be investigated more thoroughly and should probably be taken seriously. In particular, there is reason to 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. If proven, this would have serious implications for diffusion systems where the bass is presented over a limited number of subwoofers or where replay is over headphones. Additionally, it is worth remembering that the mechanisms of directional hearing described above are only components of the holistic, integrated directional perception facility that humans possess.

2.3 History and Assessment of Existing Systems.

Reality-mimicking systems can be roughly divided into those which use headphone presentation and those which use loudspeakers. There are, of course, hybrids based on the *transaural* system developed by Cooper and Bauck (Cooper and Bauck 1989). These use *interaural crosstalk cancellation* to allow material intended for headphone presentation to be presented over loudspeakers. This would not be possible without cancelling out the crosstalk caused by sounds from the left speaker reaching the right ear and vice versa There are other hybrid forms, such as the *Personal Sound Environment* (WWW [1]) combined headphones and loudspeaker system which is used for some IMax films.

2.3.1 Headphone based systems.

Headphone based systems, more commonly known as *binaural* systems are ones in which the sound is introduced directly into the ear canal via transducers mounted close to or in the ear, without directly interacting with any of the spatial hearing mechanisms mentioned above. As such, in one sense, the very first excursion into artificial spatial sound, Clément Ader's use of multiple sets of telephone transmitters and receivers in the first multichannel audio transmission system (Askew, 1981) might be regarded as being binaural. Although it is unclear whether the original intention was, in fact, to give some sense of sound source position via the use of spatially separated transmitters, the effect was certainly noted and remarked upon by many visitors to his exhibit at the 1881 Paris Exhibition of Electricity. However, there was no attempt in this system to implement anything which would mimic the cues generated by a human head, for instance by mounting the microphones in the "ears" of an artificial (*Dummy*) head so this was not full binaural.

The first real work of any significance in this field was conducted in the 1920's when a binaural/headphone based system was developed by Dr. Harvey Fletcher and his team at Bell Labs (Sanal 1976) but the work was not applied at the time, interest shifting instead to loudspeaker based systems which could be used in cinemas (Fox, 1982). However, binaural systems implemented either directly over headphones or indirectly over two speakers using crosstalk cancellation have become very popular in recent years. They have achieved widespread acceptance in Virtual Reality (VR) systems (mostly direct presentation), computer gaming (mostly indirect) and Home Theatre applications (indirect). This is largely due to the fact that they have relatively low resource demand in relation to their performance, at least when compared to any system using loudspeakers. This is certainly true for simple recordings made using *dummy head* techniques although initially this was not the case for synthetic soundscapes. Here, the computing power necessary originally limited the technique's application to areas where it use could be justified by the fact that it was obviously the

'correct' way of approaching the problem of full 3-d spatialisation of sound This perception was, and still is, supported by the concept that exact duplication of what the ear would hear in a natural situation will produce the best reproduction. Unfortunately, even ignoring the possibility that binaural systems do not act upon all significant sound perception mechanisms, there are major problems with binaural systems;

- No satisfactory methodology exists for recording real soundfields which will allow the use of the head movement cues, reducing reality equivalence.
- The wide variance in individuals makes the use of dummy heads and computational models of the head and body which do not closely match that of a particular individual unreliable at best and completely unusable at worst.
- The need to wear a piece of technology, the headphones, can provide a strong counter-cue, diminishing the degree of reality equivalence.
- Crosstalk cancelled loudspeaker presented binaural may generally be rejected for either one-many or many-many mappings because the very small "*sweet spot*" where it works correctly disrupts the sense of reality for most members of the audience, though it does have applications in specific situations which are essentially one-one mapped, such as computer gaming or televisions.

2.3.2 Loudspeaker Based Systems

In our search for a high level of reality equivalence, we can immediately dismiss two channel stereo, satisfying though this might be when the recording is good, since, unless the image is highly distorted, it only presents a small part of even the horizontal image and no vertical component, except inadvertently. We can also dismiss the existing *cinema style* systems, since despite appearing to cover the whole of the horizontal image, this fails to meet *homogeneity* (Malham, 1999a) criteria and, in any case, no serious attempt is made to deal with the vertical dimension.

There are currently only three systems which even attempt any significant degree of reality equivalence. They are:

- Wave field synthesis (Boone, 1996)
- Holophony (Nicol, 1998)
- Ambisonics (Gerzon, 1972,1975)(Fellgett, 1975)

The three systems are all related, as Nicol and Emerit have shown (Nicol, 1999), in that they all attempt to re-create (or create) the wavefronts in a soundfield, using multiple speakers as the 'secondary' sources required by the *Huygens*[2] principle. Of the three, only Ambisonics can achieve an exactly reality equivalent image and that only at the central point, the others sacrificing this in favour of being able to provide an acceptable, but not reality equivalent, image over a larger area. In essence, all three attempt to produce appropriate reality equivalent wavefronts using large numbers of speakers (although Ambisonics can successfully use as few as four for correct operation in the horizontal plane

[2]The *Huygens Principle* states that a propagating wavefront may be regarded as consisting of a large (ideally infinite) number of secondary sources.

only) disposed around the listening area. Holophonics and wave field synthesis are *volume solutions* which both use large numbers of microphones to sample the acoustic wavefronts over a significant area whereas Ambisonics uses a small array of microphones to sample the wavefronts crossing a single point in the recording space, thus making it a *point solution*. For the volume solutions mentioned above, spatial aliasing results in (non- zero) errors which are similar at all points in the listening area and which increase with frequency, due to the finite spacing of the transducers in both recording and reproduction arrays. As a point solution, Ambisonics is immune to spatial aliasing at the sampling point. It does, however, produce increasing, frequency dependent, errors as you move away from the central point. This can be ameliorated by going to a higher order system, such as *second order* (Malham, 1999b)(WWW 2). Even with second order, however, whilst the performance remains better in most respects than other systems, the errors outside the relatively small central area are sufficient to disqualify the Ambisonics approach from being fully reality equivalent, at least when used on its own. Nicol and Emerit proposed a hybrid of Ambisonics and Holophonics which combines the best features of both but still seems some way from reality equivalence.

Another approach, proposed recently by the current author, departs from the idea of attempting to create wavefronts based on Huygens and uses, instead, real sound sources. In the *Hyper Dense Transducer Array* (Malham 1999a) there is a transducer for every individually distinguishable sound source position which, depending on the nature of the sound source and whose research results you accept, means transducers placed between 1° to 6° degrees apart. There would therefore be from something over a thousand to many tens of thousands of transducers placed on the surface of a notional sphere[3] around the listeners. The driving fore behind this idea was the comments by many people that any system based on *phantom imaging* was bound to be distinguishable as artificial because the perceptual characteristics of the phantom images and real ones diverge, at least in part because their phase velocities are different (Daniel, 1998).

With good transducers, many-many mapped reality equivalence is potentially achievable for any source position outside the sphere, but for source positions inside the sphere, the system would have to transfer to one of the Huygens based wavefront reconstruction systems. As we have seen, this would result in a departure from reality equivalence except in the central area. Here the angular separation no longer translates to a linear separation large enough to cause spatial aliasing at the frequencies of interest. Unfortunately, this effectively rules this interesting approach out in terms of reality equivalence when used over a large (concert hall sized) area..

3 CONCLUSIONS

The conclusions to be drawn from the foregoing is that none of the approaches to spatialisation mentioned can meet the criteria for true reality equivalence on their own. Whilst it is possible that new approaches may be developed that will be able to meet this criteria, the best option at present appears to be to develop the hybrid speaker/headphone approach such. Close (to the head) sounds and any other appropriate sounds would therefore be reproduced binaurally using non-obtrusive head tracked headphones and individualised HRTF's and all other sounds would be handled by one of, or a combination of, the Huygens based volume solutions. In this way, both the direct air-ear and the non air-ear sound perception mechanisms would be properly dealt with for both one-one and many-many mappings. The hybrid approach would, however, be difficult to use with one-many mapping when the listeners are in the same acoustic space since the speaker based presentations are not isolated. It

[3]In practise, they would probably be placed on flat surfaces with delays and amplitude shading used to place them effectively on the spherical shell.

should, however, be usable when the 'many' are isolated, as in non-concert situations such as broadcast or recordings.

It has taken almost a century to make any significant progress beyond Ader's 1881 system. Much of what happened between then and the last quarter of the twentieth century being more in the way of relatively minor improvements or formalisations. In the last twenty years the pace of change, at least in the research arena, has considerably quickened but with the appearance of multichannel systems (Home Theatre, DVD, multichannel internet etc.) we are now also seeing the results filtering out to the consumer which in turn is stimulating further developments both technically and on the artistic side. There has never been a more exciting time to be involved in surround sound and there are at least some small signs that we may be moving towards the achievement of full reality sound systems.

4 REFERENCES.

Askew, Anthony "The Amazing Clément Ader" Studio Sound, Volume 23, no's 9, 10 and 11, September, October, November 1981

Begault, D. R. "3-D Sound for Virtual Reality and Multimedia" AP Professional, Boston, San Diego, New York, London, Sydney, Tokyo, Toronto

Boone, M.M., Verheijen, E.N.G. and Jansen, G. 1996 " Virtual Reality by Sound Reproduction Based on Wave Field Synthesis" 100th Convention of the Audio Engineering Society, preprint 4145, Copenhagen, May 1996

Cooper, D.H. and Bauck, J.L. for Transaural Recording' *Journal of the Audio Engineering Society*, Vol 37, No. 1/2, Jan/Feb 1989 pp. 3-19.

Czerwinski, Eugene, Voishvillp, Alexander, Alexandrov, Sergei and Terekhov, Alexander "Propagation Distortion in Sound Systems - Can We Avoid It?" *Journal of the Audio Engineering Society* Vol. 48, No.1/2 January/February, 2000 pp 30-48

Daniel, Jérôme, Rault, Jean-Bernard and Polack, Jean-Dominique 1998 'Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions', *preprint no. 4795, 105th Audio Engineering Society Convention*, Sept. 1998,. (Corrected version available by contacting the authors at Centre Commun d'Etudes de Télé-diffusion et Télécommunications, Cesson Sévigné, France)

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

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

Gerzon, Michael A.'Periphony: With-height Sound Reproduction' *Journal of the Audio Engineering Society*, Vol. 21 No. 1 Jan/Feb 1972 pp.2-10

Gerzon, Michael A. 'Ambisonics Part two: Studio Techniques' STUDIO SOUND, August 1975. pp24 - 30.

Malham, D.G. 1999a "Homogeneous and non-homogeneous surround sound systems" proceedings, AES UK 'Audio: the second century' Conference, pp25-34, London, 7-8 June 1999

Malham, D.G. 1999b "Higher order Ambisonic systems for the spatialisation of sound." Proceedings

of the International Computer Music Conference, Beijing, October 1999, pp 484-487 (see WWW 2 for corrected version)

Nicol, R. and Emerit, M. 1998 " Reproducing 3D-Sound for Video Conferencing: A Comparison Between Holophony and Ambisonic. Proceedings of the First COST-G6 Workshop on Digital Audio Effects (DAFX98) Barcelona November 1998, pp17-20 http://www.iua.upf.es/dafx98/

Nicol, R. and Emerit, M."3D-Sound Reproduction over an Extensive Listening Area: a Hybrid Method Derived from Holophony and Ambisonic" The Audio Engineering Society 16th International Conference on Spatial Sound Reproduction, Helsinki 1999, preprint no. s66819

Nielsen, Søren H. "Auditory Distance Perception in Different Rooms" *Journal of the Audio Engineering Society*, Vol 41, No.10, Oct. 1993, pp 755–770

Sanal, Arthur J. "Looking Backward" Journal of the Audio Engineering Society, Vol. 24, No. 10 December 1976 p. 832

WWW 1 http://www.sonics.com/products/pse.html viewed on 23/09/2000.

WWW 2 http://www.york.ac.uk/inst/mustech/3d_audio/secondor.html viewed on 27/09/2000.