The generation of virtual acoustic environments for blind people

D A Keating

Department of Cybernetics, University of Reading,
Whiteknights, PO Box 225, Reading, ENGLAND

cybdak@cyber.reading.ac.uk

ABSTRACT

VR systems, like the cinema, tend to concentrate on the visual image, leaving the audible image as garnishing. If blind people are to make any use of virtual environments then the audible image needs to be greatly improved. An ideal system would encode the amplitude and direction of sounds in a way that was independent of the means used to portray the image to the user. The Ambisonic B-format describes four signals: a reference, and three vectors. These are simple to generate and manipulate and may easily be converted into loudspeaker feeds for one, two or three dimensional arrays of speakers.

The generation of binaural signals for headset based systems is made difficult by the complexity of the head related transfer function(HRTF). Steering the virtual image to take account of head movements is computationally expensive and requires detailed knowledge of the users HRTF if the image is to be at all realistic. An alternative is to make use of virtual loudspeakers around which the image is moved but which have a fixed location relative to the users head. If the speaker feeds are derived from a B-format signal this may easily be panned and tilted by manipulating the relative amplitudes of the three vectors according to simple mathematical rules. The HRTF of the user can be measured at the positions of the virtual speakers and the resultant relationship between the combination of all these signals and the two headphone signals need only be computed once.

This paper gives an introduction to the Ambisonic system and shows how B-format signals can be used to encode a virtual acoustic environment. It concludes by describing how the B-format signals may be manipulated and used to generate almost any number speaker feeds or headphone feeds.

Keywords: surround sound, ambisonics, acoustic modelling

1. INTRODUCTION

1.1 What is a virtual acoustic environment?

An acoustic environment (virtual or real) consists of a number of primary sound sources in 3-D space and a number of reflecting surfaces. Each of these surfaces will have characteristics depending on its sonic properties. For example hard solid surfaces will reflect all frequencies well but the addition of a soft covering will increase the absorption at high frequencies and hence lower the reflectance. Suspended panels absorb particular bands of frequencies; glass for example can be a good absorber at low frequencies but a good reflector at high frequencies. The sense of space perceived by an observer is primarily governed by the timing and direction of reflections from the surfaces. A sense of depth in stereo recordings of real sounds is usually due to the timing of early reflections from near surfaces such as the walls or floor of the room in which the recording was made.

1.2 Interacting with an acoustic environment

In traditional audio entertainment, or indeed cinema, the experience is totally passive. In a virtual acoustic environment the user is able to move around the environment thus increasing their involvement and allowing them to build up a better mental image of the acoustic space. To further the ability of the user to build up this image active involvement is required. If the user makes a sound then they should hear the response of the environment to this sound. This means not only that objects in the environment should react to sound (for example birds could stop chirping when disturbed by sound) but that the sound should be treated as any other primary source and appear to be reflected by the many surfaces in the environment. This would allow the 'acoustic' of otherwise quiet places to be determined by the user.

©1996 ECDVRAT and University of Reading, UK ; ISBN 0 7049 1140 X

2. REPRESENTING DIRECTION

As far as the user of a virtual environment is concerned, the environment consists of a number of sounds each of which comes from a particular direction. The method by which these sounds are finally to be portrayed to the user should not concern us at this moment. The biggest failing in surround sound systems for audio entertainment is that the designers failed to realise that the need was to encode the directions of the sounds, not the method by which the image was finally to be created. The ill fated Quadraphonic system of the 1970s set domestic surround sound back by about 15 years and Dolby Surround threatens to do the same again at the present time. The ambisonic system of representing direction (in 3-D) in audio signals was conceived at the same time as Quadraphonic and thus despite being based on sound psycho-acoustic and engineering principles (Gerzon, 1974) was tarred with the Quadraphonic brush.

2.1 Ambisonic B-Format

The ambisonic B-format representation of sounds including direction consists of four signals. These signals in turn represent the zero and first order spherical harmonics in three dimensions (Gaskell, 1979; Gibson et al, 1972). The zero order harmonic (denoted W) is simply a sphere which has an amplitude which is independent of direction. The three first order harmonics consist of dipoles in three mutually orthogonal directions. These are (by convention) Front-Back (denoted X), Left-Right (denoted Y) and Up-Down (denoted Z). The B-format representation may be expanded to contain the second order harmonics which are quadrapoles. This gives the ability to improve the discrimination of direction (Gerzon, 1976; Bamford and Vanderkooy, 1995) but requires a total of nine signals to represent all three dimensions or five to represent the horizontal plane only. The advantage of increased discrimination is at present outweighed by the disadvantage of having to produce five or nine channels of audio output. The basic B-format representation which consists of a reference and three vectors is mathematically simple to produce and (with the exception of binaural) it is also simple to transcode into other formats.

2.2 Ambisonic UHJ

The ambisonic UHJ format is a hierarchical format which allows surround sound to be encoded in one to four channels. One channel gives a mono (directionless) signal, two give horizontal surround of low directivity but which is stereo compatible, three give horizontal surround of higher directivity and four give full periphonics (with height). Two channel recordings (strictly BHJ but usually referred to simply as UHJ) are available in the domestic audio formats. The two channel format uses the relative amplitudes of the two signals to encode left-right information and the relative phase to encode front-back information. UHJ recordings are usually produced from B-format masters, a soundfield microphone or similar pseudo-coincident microphone technique being used to provide the B-format signals directly for recording of the master.

2.3 Binaural

The binaural representation of direction attempts to mimic that which is available to humans; that is the two signals which are received by the ears. A major problem is that these signals are unique to each of us and so there is no standard definition of them. There are, however, a number of commonly used pseudo-standards; the BBC use a perspex disc of a certain diameter and define the signals as those which would be detected by microphones at a certain distance either side of the disc. The outputs from proprietary dummy heads such as those produced by Bruel and Kjaer or Senheiser might also be considered as standards. The mathematical relationship between the direction of the sound source and the two binaural signals is extremely complex even when generalisations and simplifications have been made. Thus not only is it difficult to produce these representations but if they are not to be used directly to provide the feed signals of headphones it is also difficult to derive any other signals from them.

2.4 Dolby Surround

The Dolby Surround (or Dolby Stereo) formats for cinema or home video surround sound are perhaps the most common surround sound formats in use today. Unfortunately there is no full definition of direction associated with these formats. Four signals are defined: Centre (for the dialogue channel), Right and Left (primarily for the stereo music soundtrack) and Surround (for all sound effects). Two channels are produced from these four by means of a simple phase-amplitude matrix. The primary use of the Dolby formats is to provide a surround effect to augment a dominant visual image and not to accurately reproduce an acoustic environment. Despite the lack of full definition there are proprietary processors available which will produce Dolby signals from mono sources and directional information. In order to reproduce the direction intended, the loudspeaker layout must be identical to that used by the manufacturer who produced the processor. This is not usually defined in detail, and even if it were the basic Dolby format is incapable of positioning signals from the rear with any kind of real accuracy (all the rear speakers carry the same signal). The more recent 5.1 channel (digital) format does allow positioning at the rear but is still not fully defined and suffers from the same problem as the earlier formats in that the speaker positions are fixed by the format.

2.5 Dolby Pro-Logic

Dolby Pro-Logic does not refer to a format in its own right but refers to the type of decoder used. A pro-logic decoder takes the normal two Dolby Surround signals and steers the four output signals depending on the dominant direction of the source. This gives better directivity than the simple decoder when there is a dominant direction but can offer no advantage if many directions are present at once. An additional limitation of the Dolby formats is that they have no provision for height information.

3. MANIPULATING THE DIRECTION

Helmet based VR and telepresence systems (Griffin and Keating, 1992) require acoustic and visual images which track head movements. Such helmet based systems traditionally use headphones for the audio reproduction. VR head tracking systems are available commercially which give information on head position and orientation. If the head orientation signals are used to steer the surround audio then the imagery obtained from the headphones can be good. Any front-back ambiguity may be resolved by small head rotations although the effect of the headphones on the conch resonance (Moore, 1989) needs to be compensated for as this is a potential source of lack of realism. Even if the system is not helmet based then the surround signal may need to be steered as the user is likely to be able to rotate in the virtual environment albeit by joystick or similar control.

3.1 Steering Two-channel Sources

If we are considering a generalised surround signal steering block then the major problem is the large number of different formats which must be catered for. In addition to this, each of the formats has its own set of problems. We cannot, for example, rotate a stereo signal and produce a stereo format output as stereo is not a surround format. In Dolby surround, attempts to mix between speaker feeds (L, R, C & S) to achieve rotation yields strange results as the front and surround signals are simple combinations of left and right. Rotating the sound image by 180 degrees would in any case be impossible as there is no side information at the rear due to the speaker layout used.

Steering UHJ in 2-channel format is difficult as the direction information is encoded using amplitude and phase difference, thus a rotation is not a simple function of the two signals. The same is true to an even greater extent with binaural sources. Not only must the amplitudes and phases be manipulated in a complex fashion but the generalised head related transfer function (HRTF) must be known (Wenzel 1992). This makes steering these formats directly very difficult.

3.2 Steering Multi-channel Sources

Multi-channel formats fair little better if they are defined in terms of speaker feeds. The one format which is a notable exception to this is ambisonic B-format. The components of the B-format can be defined as follows: W (the reference signal) has a gain of 0.7071 regardless of direction, the vector Y (left-right) has a gain of sin(A) * sin(B), X (front-back) has a gain of cos(A) * cos(B) and Z (up-down) has a gain of sin(B). Where A is the angle of the source measured counter clockwise from the front and B is the angle of elevation.

Steering these signals is simplicity itself (Gerzon, 1975; Malham and Clarke, 1992). To rotate a signal C degrees counter clockwise the following equations are used:

$$newW := W \tag{1}$$

$$newX := X.\cos(C) - Y.\sin(C) \tag{2}$$

$$newY := Y.\cos(C) + X.\sin(C) \tag{3}$$

$$newZ := Z \tag{4}$$

In other words W and Z are unchanged and the new X and Y terms are combinations of the old terms modified by gains in the range -1 to +1.

Tumble (rotation about the Y-axis) by an angle D may be achieved as follows:

$$newW := W \tag{5}$$

$$newX := X.\cos(D) - Z.\sin(D) \tag{6}$$

$$newY := Y \tag{7}$$

$$newZ := Z.\cos(D) + X.\sin(D) \tag{8}$$

Tilt (rotation about the X-axis) by an angle E may also be added if required as follows:

$$\text{newW} := W \tag{9}$$

$$\text{newX} := X \tag{10}$$

$$\text{newY} := Y.\cos(E) - Z.\sin(E) \tag{11}$$

$$\text{newZ} := Z.\cos(E) + Y.\sin(E) \tag{12}$$

Each of the new vectors is thus a sum of each of the old vectors multiplied by a constant, where the constants depend on the manipulation required. The reference remains unchanged throughout. It makes a great deal of sense to produce our virtual environment using the B-format definition of direction as manipulating the signals to allow for movements of the user is so easy. If, however, we wish to incorporate recorded sounds which are in another format into our virtual environment they must first be converted into B-format.

4. CONVERTING SURROUND SOURCES TO B-FORMAT

Fortunately there already exist methods of producing pseudo B-format signals from stereo and UHJ sources. These are not ‘pure’ B-format as the directional information is contained in both amplitude and phase. The steering algorithms quoted earlier, however, are unaffected by this and give the same results as they would with ‘pure’ B-format.

4.1 Converting Dolby Surround to B-format

It remains necessary to develop methods of converting Dolby surround, binaural and the multi-channel formats into B-format. Although it would be possible to decode the 2-channel Dolby surround signals directly the resultant surround channel (L-R) requires decoding by a modified B-type noise reduction decoder. It is therefore better to take the centre (C) and surround (S) channels from an existing Dolby-surround decoder with zero delay on the surround channel..

A transcoder developed by the author (Keating and Griffin, 1993) which was found to give acceptable results used the following equations:

$$W = 0.5\,(C - jS) \tag{13}$$

$$X = 0.7\,(C + jS) \tag{14}$$

$$Y = S \tag{15}$$

These are similar to the ‘super stereo’ or ‘enhanced’ equations usually used in ambisonic UHJ decoders but with increased difference signal. It is not suggested that these are optimal but they give reasonable results all the same.

4.2 Converting Dolby Pro-logic to B-format

Using a Dolby Pro-Logic decoder presents some problems as the outputs may or may not be ‘steered’ electronically by the decoder depending on whether there is any dominant direction present. A converter must therefore give acceptable results when presented with simple surround sound signals or (in the extreme case) only one active output. If for example there is only dialogue present the pro-logic decoder will reduce the left, right and surround output signal amplitudes considerably, leaving only the centre channel active.

A set of transcoder equations which give the same results as those given earlier for surround signals but also copes with Pro-Logic signals is as follows :

$$W = 0.25C + 0.177(L + R) - 0.5jS \tag{16}$$

$$X = 0.467C + 0.165(L + R) + 0.5jS \tag{17}$$

$$Y = 0.7(L - R) \tag{18}$$

4.3 Converting UHJ to B-format

Standard ambisonic decoder equations may be used to produce B-format from UHJ. These are as follows:

$$W = 0.667\Sigma + 0.11j\Delta \tag{19}$$

$$X = 0.556\Sigma - 1.099j\Delta \tag{20}$$

$$Y = \Delta + 0.513j\Sigma \tag{21}$$

where:

$$\Sigma = (L + R)/2 \tag{22}$$

$$\Delta = (L - R)/2 \tag{23}$$

4.4 Converting Stereo to B-format

Stereo signals may be considered as two mono signals and placed anywhere but the following equations allow the stereo image to be positioned to the front with a width (α) of up to 180 degrees:

$$W = 0.65\Sigma - 0.27j\Delta(\alpha/180) \tag{24}$$

$$X = 0.98\Sigma + 0.4j\Delta(\alpha/180) \tag{25}$$

$$Y = 0.75\Delta(\alpha/180) \tag{26}$$

4.5 Converting Binaural to B-format

The problem format is binaural because there is no standard definition of the Left and Right signals. These will in general be similar but will depend on the exact recording/synthesising technique employed. Although side information is at high frequencies carried in the relative amplitudes of the two signals this reverts to time difference (or phase) at frequencies below about 700 Hz. In addition front-back information is carried in subtle clues caused by the head and ear shapes. It is thus very difficult to decode or transcode the directional information present in binaural signals and for this reason this has not yet been considered by the author.

5. CONVERTING B-FORMAT SIGNALS TO SURROUND

In order to give versatility to a virtual acoustic environment generating system it is useful to be able to provide outputs in any of the common surround-sound formats

5.1 Converting B-format to Dolby Surround

The author has used a standard UHJ encoder driving a two-channel Dolby surround decoder with some success. The following transcoder equations for four channel discrete or two channel Dolby surround should be a good starting place for readers wishing to experiment:

$$C = W + 0.71X \tag{27}$$

$$L = W + 0.71Y \tag{28}$$

$$R = W - 0.71Y \tag{29}$$

$$S = jW - 0.71jX \tag{30}$$

or:

$$L = (0.86 + 0.35j)W + (0.25 + 0.25j)X + 0.35Y \tag{31}$$

$$R = (0.86 - 0.35j)W + (0.25 - 0.25j)X - 0.35Y \tag{32}$$

5.2 Converting B-format to UHJ

The standard UHJ encoder equations are:

$$L = (0.4699 - 0.17j)W + (0.0928 + 0.255j)X + 0.3278Y \qquad (33)$$

$$R = (0.4699 + 0.17j)W + (0.0928 - 0.255j)X - 0.3278Y \qquad (34)$$

5.3 Converting B-format to Stereo

The UHJ encoder equations above may be used to provide a stereo compatible output but the following simpler equations may be used:

$$L = W + 0.35X + 0.61Y \qquad (35)$$

$$R = W + 0.35X - 0.61Y \qquad (36)$$

These have the advantage that ninety degree phase shifters are not required but suffer from the disadvantage that sounds from directly behind the user are quieter than if the UHJ encoder is used. As stereo is not a surround format the rearward sounds will appear to come from the front

5.4 Converting B-format to Binaural

The translation of B-format signals to binaural could potentially be as difficult as binaural to B-format but fortunately we can cheat. It would be very difficult to derive a generalised transfer function which translated B-format signals directly to left and right ear signals. The complete head related transfer function (HRTF) (Wenzel, 1992) would have to be known and filters derived from this to modify the left and right ear signal amplitudes and phases. Determining the HRTF over a full sphere experimentally would require a great deal of time in an anechoic chamber; whilst implementing the translation functions thus derived would require a powerful DSP system or massive amounts of analog electronics.

The HRTF may of course be computed (Rasmussen and Juhl, 1993) rather than measured but once again the implementation of such a complex transfer function is difficult. The solution is to only measure or compute the function at a few fixed angles. An ambisonic decoder driving four speakers placed around the listener is capable of giving a good illusion of direction in the horizontal plane. It can be argued therefore that if the HRTF is known for these four positions then that is all the information that is required. (In fact symmetry reduces this to two). Calculating the cumulative effect of these speakers at the ears results in right and left ear signals in terms of the four speaker feeds. Each speaker feed is defined in terms of the B-format signals W, X and Y (for horizontal only) and so an overall definition of the right and left ear signals in terms of the B-format signals may be derived very simply. This then gives us the translation of B-format to binaural. The number of virtual speakers is of course arbitrary and a greater number will give better results but will require the HRTF to be known at a greater number of source directions. This method would of course also work for periphonics if a three dimensional array of at least eight loudspeakers was considered. Increasing the order of the spherical harmonics from first to second would give improved directivity but would require a minimum of six virtual speakers in the horizontal plane and twelve if all three dimensions were required.

6. CHOICE OF ACOUSTIC OUTPUT METHOD

The first choice that must be made is whether to use headphones or to use an array of loudspeakers.

6.1 The advantages and disadvantages of using headphones

Headphones have many advantages over loudspeaker arrays. They may already be part of the VR system available to the system developer, and if not they are inexpensive to add. In use they are compact and most importantly they limit the sound they produce to the ears of the user. This becomes of great significance if the user is to interact with their environment by means of a microphone. They do, however, require a head tracking system if head movements are to give the normal additional directional information used to resolve ambiguities. They also require drive signals which are complex to calculate if realism is to be achieved.

6.2 The advantages and disadvantages of using loudspeakers

Arrays of loudspeakers may be driven from B-format signals via a simple decoder (Gerzon, 1980). Many different layouts may be used with different numbers of loudspeakers. At least four loudspeakers are required for horizontal surround sound and at least six for full periphonics (although eight are preferable as the birectangular layout allows compatibility with horizontal surround and stereo). The loudspeakers should be identical but need not be full-range units as a sub-woofer may be driven from the W output of the B-format signals. The cost of such a set-up is still high as four or more 2-channel (stereo) amplifiers are also required to drive these speakers. The set-up is not easily portable and so is best set-up in its own room. This also avoids the sound from annoying other people in the vicinity of the VR set-up. The main advantage of loudspeaker arrays is that errors in decoding degrade the image gradually, whereas those in the binaural (headphone) system can cause massive perceived errors in direction or highly ambiguous results.

6.3 Experimental Results

Initial experiments by the author used an ambisonic decoder driving four speakers placed around a dummy head, which in turn was driving a pair of headphones. The speakers and head were placed in a dead room to prevent room acoustics from having an effect. In this simple experiment the B-format signals were derived from a two channel UHJ source and were not steered.

The image from the loudspeakers was realistic and open. A 'walk-around' test showed that the directions of the sounds were being reproduced correctly. There was some lack of stability of the image if the listeners head was rotated but this should be less of a problem with a true B-format source rather than one decoded from 2-channel UHJ.

The headphone signal was also perceived to be open, with little of the 'in the head' localisation common with headphones. The 'walk-around' test however showed the classic front-back ambiguity to be present, with different subjects perceiving different directions. Is it thought that the addition of a head tracker steering the signals will help considerably. The decoder is now being implemented using DSP to replace the 'acoustic' decoder used in the initial experiment. This will eliminate any unwanted effects due to low frequency room resonances or imperfections in the loudspeakers or dummy head, and should allow the measured HRTF from the user to be used rather than that of the dummy head.

7. CONCLUSIONS

In this paper the problems of providing VR users, particularly those without the benefit of sight, with a satisfactory image of a virtual acoustic environment have been considered. The ambisonic B-format representation of direction of acoustic sources has been suggested as a sensible one to use in a VR system. The methods by which other surround formats may be converted to B-format have been given, as have methods of converting B-format into other formats. A solution to the particular problem of producing binaural signals has been suggested and the early experimental results show this method to hold promise.

8. REFERENCES

J S Bamford and J Vanderkooy (1995), Ambisonic Sound for Us, *Proc. 99th AES Convention*, Pre-print No 4138, October 1995.

P S Gaskell (1979), Spherical Harmonic Analysis and Some Applications to Surround Sound*, BBC Research Dept. Report*, BBC RD 1979/25.

M A Gerzon (1974), Surround-Sound Psychoacoustics, *Wireless World*, **80**, April, pp. 36-39.

M A Gerzon (1975), Panpot and Soundfield Controls, *NRDC Ambisonic Technology Report No 3*, August.

M A Gerzon (1976), Maximum Directivity Factor of n-th order Transducers, *J. Acoust. Soc. Am*, **60**, July, pp. 278-280.

M A Gerzon (1980), Practical Periphony: The Reproduction of Full-Sphere Sound, *Proc. 65th AES Convention*, London, Feb. 1980.

J J Gibson, R M Christensen and A L R Limber (1972), Compatible FM Broadcasting of Panoramic Sound*, J. Audio Eng. Soc.*, **20**, Dec., pp. 816-822.

M Griffin and D Keating (1992), Factors Affecting Sound Synthesis and Presentation Within Virtual Reality Systems, *Proc. I.O.A.*, **14**, Part 5, pp. 273-280.

D A Keating and M P Griffin (1993), The Production of Steerable Binaural Information From Two-Channel Surround Sources, *Proc. I.O.A*, **15**, Part 7, pp. 173-185.

D G Malham and J Clarke (1992), Control Software for a Programmable Soundfield Controller, *Proc. I.O.A.*, **14**, Part 5, pp. 265-272.

B C J Moore (1989), *An Introduction to the Psychology of Hearing*, 3rd Edition, Academic Press, London, 1989.

K B Rasmussen and P M Juhl(1993), The Effect of Head Shape on Spectral Stereo Theory, *J. Audio Eng. Soc.*, **41**, No 3, March.

E M Wenzel (1992), Localisation in Virtual Acoustic Displays, *Presence*, **1**, No 1.