Spatial Hearing Mechanisms and Sound Reproduction

by D. G. Malham

There are a number of different cues the ear-brain combination uses to determine the position of a sound source. Although there are may be other, more subtle mechanisms, those we will be most concerned with as recording engineers are;

1. The time of arrival of the wave front of a sound event at the ears, or more specifically, the difference in arrival times at the two different ears. A sound source anywhere on a line from due front, through due above to due back (the median plane) will have its wave front arrive at the two ears simultaneously. Move the source away from this line and one ear will begin to receive the wave front before the other. This is known as the Interaural Time Delay or ITD. This effect is only usable up to a frequency where the wavelength of the sound approaches twice the distance between the ears. Over that, it provides only ambiguous cues.

2. Sound from a source to the left of the head, for example, will arrive directly at the left ear, but will have to travel “through” (!) the head – actually it is diffracted round – to get to the right ear and will also have to travel further. It will thus be quieter 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.

3. The shape of the head and the external part of the ears results in a frequency dependent response which varies with sound position. This is known as the Head Related Transfer Function or HRTF. For positions where ILD’s or ITD’s give ambiguous or non existent differences between ear signals (such as median plane signals) this is the main positional sensing mechanism. For a sound source not placed symmetrically with respect to the two ears will further result in a different response at each ear.

4. Our ability to change the position of our head in such a way that we minimise the ITD, ILD and the difference between the HRTF’s at the two ears. This is, or should be, the point at which we are directly facing the sound source.

Of these mechanisms, analog mixers can only use one easily for producing CONTROLLABLE sound positioning over loudspeaker systems. That one is the level difference. Fortunately, the ear-brain combination is very democratic, and whichever directional mechanisms is producing the most plausible results, that is the one whose opinion will be taken.

This is what enables us (within certain limitations) to use two speakers to produce a sound image with controlled perceived direction by simply feeding more signal to one speaker than the other. This is, of course, what we do when we pan sound across a stereo image. This system is known intensity panning and the sound images thus created between the speakers are known as phantom images. Around a 15dB difference between the speakers in a stereo rig will move the apparent sound position to the loudest speaker.

There are limitations, largely on the positioning of the listener, who must be positioned so that the apparent angular separation between the speakers is around 60 degrees for the best effect. Any wider and there tends to be a ‘hole’ in the middle of the image where for one thing, the sound dips in level. Although this can be compensated for there is a much more serious problem. That is the development of instabilities in the perceived position of the sound image which become progressively more extreme as the angle increases. Above a 90 degrees separation, central images become virtually impossible to maintain.

Even at the optimum 60 degrees, unless the listener is centrally placed between the two speakers, the image will pull towards the nearer speaker. This is a result of the difference in time delays between the two sound paths which acts as if it were an ITD. Furthermore, even for listeners in the central “stereo

seat" the apparent positions of sounds varies with frequency to the extent that high frequency sound images (above say 3 kHz) near the centre tend to be 1.6 to 2 times as wide as low frequency ones.

Despite all these caveats, it is possible to achieve highly satisfactory results with simple systems, if adequate care is taken. It does, however, highlight the fact that even after sixty years of research – the first stereo patents date from the early thirties – there are still many unknowns and much work still remains to be done.

PRODUCTION OF STEREO IMAGES

It is apparent from the above that we can produce stereo images by taking the output of our microphones and feeding different amounts to the left and right channels. This is done over and over in multimiked recordings and in the work we do in electroacoustic music. In many, if not most, cases the advantages of having such a considerable degree of control over the image outweighs the disadvantage of the lack of image depth which usually results in multimiked recordings having a somewhat unnatural feel.

There are, however, a number of simple techniques which can ameliorate this unnaturalness, provided the event we are trying to record is acoustically satisfactory. These involve the use of pairs of microphones, one feeding the left, the other feeding the right channel. The Americans favour using spaced omnidirectional microphones. If they are spaced a few feet apart there are both time and amplitude differences between the two signals. This set up produces a very (perhaps too!) spacious and open result.

More common in this country is the so-called coincident pair technique. In this a crossed pair of directional – usually cardioid – microphones are used with the capsules placed as close as possible. With cardioid mics, the angle between them should be around 100 degrees, but this will vary depending on the size of the ensemble.

A slight variation on this is to space the capsules by 2-3 cms. This retains most of the imaging precision of this method, whilst gaining some of the spacious quality (but not too much!) of the spaced omni system.

The fact that there are these general differences in approach between the British recordings and American ones is a result of historical differences in the aims of the research teams in the two countries who developed the original stereo techniques during the early 1930's.

In Britain, Alan Blumlein's team was most interested in providing good stereo images in a domestic environment, with as much a sense of "being there" as possible. So, they were dealing mostly with a situation in which there would only be a small number of listeners who would be able to cluster themselves in or around the "stereo seat". In this case, one pair of crossed coincident mics could be used to create amplitude panned material in two channels to be fed to two speakers. The microphones he used were figure of eight types, crossed at 90o so that as sounds crossed the stage in front of the mic pair, the level coming from one decreases whilst the level from the other increases. This arrangement gives a very natural sound, although the front sound stage is not as wide on the speakers as it is in real life (angular distortion) and the sound is more reverberant since the sounds from the rear of the mics is picked up equally loudly but is mapped onto the frontal image produced by the pair of speakers. This is why most modern usage of this coincident pair technique uses mics with cardioid polar patterns to reduce the rear pickup.

The team at Bell Labs in the States were much more concerned with providing stereo to large audiences, e.g.. for film sound (although Blumlein's original patent mentions film sound frequently). As such, many listeners would not be in the ideal stereo seat and there would be a considerable problem with the "hole in the middle", especially since this is where much of the dialogue would be expected to be coming from. Accordingly, they worked with three channels (the centre channel on film is still called the dialogue channel) feeding three speakers. The channels were derived from either widely spaced microphones (often referred to as a curtain of mics) or via a rather complex panpot. While this approach does not give as good results in the domestic situation as Blumlein stereo, it works very well in the area for which it was designed.

AMBISONICS

Ambisonics is a method of recording information about a soundfield and reproducing it over some form of loudspeaker array so as to produce the impression of hearing a true three dimensional sound image. I deliberately say “impression” to stress the fact that if you truly wished to reproduce the soundfield present in a two metre sphere up to say 20 kHz then you can argue from information theory that you would need many, many channels and loudspeakers. Estimates of the number have varied from 400,000 upwards! In practise, all you can actually do is to determine how much information we can capture with some sensible combination of microphones and then to find some way of using that information to fool the ear into hearing a full soundfield.

Attempts to provide directional information in artificially reproduced sound images started in the late nineteenth century when a “broadcast” of a concert was made in France using multiple telephones spaced along the front of the stage, transmitting over wires to a similar number of telephone receivers. Quality was, of course, rather poor but an impression of direction was undoubtedly gained.

In the late 1920’s and early 1930’s a more formal basis for directional reproduction was laid down by Alan Blumlein in Britain and the RCA company in the States. The techniques they developed were for systems using only a small number of channels of information for reproduction over a pair of loud-speakers.

The technique developed by Alan Blumlein consisted of a pair of microphones with figure of eight characteristics, mounted as close together as possible and with the front lobe of one mic pointing 45 degrees to the left of the front-back line and the front lobe of the other pointing 45 degrees to the right. Although this does provide excellent stereo imagining it does have a problem. Because of the figure of eight characteristics sounds coming from the rear are also picked up and when reproduced over a pair of loudspeakers these sounds are folded over and mapped onto the front sound stage. This results in a sound which is too reveberant for many ears.

“Purist” recording engineers who like the simplicity and accuracy of the Blumlein technique have modified it in order to remove this perceived problem. By replacing the figure of eight microphones with ones with cardioid characteristics and changing the angle between them so that it just includes the desired sound stage, it is possible to use the cardioid mic’s lack of response to rearward sounds to reduce the mapping of rear reveberant sounds onto the front reproduced sound stage. This results in a much more acceptable if less accurate sound image. (As a matter of practice the angle between the mics should not be more than about 120 degrees or less than 90).

It does, however, seem a pity to throw away this information when we already have insufficient. The dummy head technique can be employed to utilise this lost information although only for headphone listening. (Work has been done and is still being done to get better results over loudspeakers for dummy head recordings but problems still remain to be solved). By using some form of analogue of the human head with microphones picking up sound where the ears should effectively be and then reproducing these signals over headphones very good results can be obtained with sounds appearing to come from all directions, not just the front. Unfortunately, the best results with the most stable images come from a dummy head which closely matches the listeners. However, the more closely the head matches that of any one listener, the worse the results may get with other listeners. Even if you try to generate some kind of average head you can come unstuck. One set of recordings I heard a few years ago, which were made using a head based on several years of painstaking measurements of all the colleagues and students of a Continental European researcher, gave absolutely stable and very precise results except for the fact that to me and all the other British people who listened to it, the front and back directions were transposed. The BBC approach where the head is just a disk of perspex with microphones placed a few centimetres either side of it gives a more universally acceptable result at the expense of true precision.

Ambisonics, on the other hand, goes back to the original ideas of Alan Blumlein and builds on them. By just adding an omnidirectional microphone to the pair of figure eight units it can be shown that you can capture ALL the information that it is possible, with such simple low order microphones, to capture about the horizontal soundfield at that point. It is, of course, assumed that you have arranged to have the capsules TRULY coincident, that is all three capsules are acoustically at exactly the same place in the soundfield. This impossibility becomes even more difficult when you add an up-down oriented figure eight capsule in order to record height information as well. This problem has been overcome in

the Soundfield microphone which uses four small capsules situated on the surface of a notional sphere to sample the incoming sounds. By some clever mathematics it is possible to generate the signals which would have been given by our four truly coincident capsules-at least up to some reasonably high frequency. (It should be noted that in Ambisonics the horizontal figure eight units are mounted front-back and side-to-side rather than at 45 degrees).

Having got the information recorded in this form, the task of producing the illusion has to be accomplished. This is completely separate from the task of capturing the information in the first place and is based on an amalgam of various theories of hearing covering both low (below 700 Hz) and high frequency mechanisms. The decoder must be adjustable for different speaker layouts.

The question must be posed 'How does this approach differ from the Quadraphonic systems?. Quadraphonics, or more properly Quadrifontal, systems were based on a very simple theory. If mono sound systems can be regarded as a hole in a concert hall wall and stereo systems as two holes AND are better then four holes MUST be better still. Unfortunately this is simply untrue since the extra information carried is partially redundant and causes considerable confusion and instability in the perceived images, particularly along the sides.

Extensive listening tests over many years show Ambisonic recordings to be at least as good as any other form of recording at capturing sound images and far better than most, but what is its applications in electroacoustic music? To understand these we need to look at some basic theory on Ambisonics.

BASIC AMBISONIC TECHNOLOGY

The Ambisonic surround sound system is essentially a two part technological solution to the problems of encoding sound directions (and amplitudes) and reproducing them over practical loudspeaker systems in such a way as to fool the ears of listeners into thinking that they are hearing the original sounds correctly located. This can take place over a 360 degree horizontal only sound stage (pantophonic systems) or over the full sphere (periphonic systems). Systems using the so-called 'B' format signals to carry the recorded information require three and four channels respectively for full encoding of sounds to the kind of accuracy achievable with first order microphones (cardioid, figure eight etc.). Reproduction requires four or more loudspeakers depending on whether it is pantophonic or periphonic, size of area etc. Practical minima are four for horizontal only, eight if you require height as well. The important thing to note is that there is no need to consider the actual details of the reproduction system when doing the original recording or synthesis, since if the B format specifications are followed and suitable loudspeaker/decoder setups are used then all will be well. In all other respects the two parts of the system, encoding and decoding, are completely separate.

ENCODING EQUATIONS

The position of a sound within a three dimensional soundfield is encoded in the four signals which make up the B format thus;

X = cosA.cosB (front-back)

Y = sinA.cosB (left-right)

Z = sinB (up-down)

W = 0.707 (pressure signal)

where A is the anti-clockwise angle from centre front and B is the elevation. If you limit the positions of sounds to within the unit sphere by ensuring that

(x? + y? + z?)

is always less than or equal to one then the equations can be more simply written as;

X = x

Y = y

Z = z

W = 0.707

where x,y,z are the coordinates of the sound source. The value of W is given as 0.707 rather than 1, since this allows for a more even distribution of levels within the four channels. This convention should be adhered to as the decoder designs are predicated on this. There is a catch in this simplicity, however, since if you attempt to move off the surface of the notional unit sphere and in towards the centre, the dropping levels in the X,Y,Z channels will reduce the overall sound level, rather than there being the expected increase as the apparent position of the sound source moves nearer the centre.

One fix that will keep the overall level pretty well constant is to make W vary thus;

W = 1 – 0.293(x? + y? + z?)

Further modifications can be made to allow for an overall increase as sounds move to the centre position, which corresponds more closely to reality.

ENCODING A MONOPHONIC SOUND INTO AMBISONIC B-FORMAT.

Since the decoder designs are predicated on the basis that sounds being positioned in Ambisonic B-format are placed on the surface of or within a notional unit sphere the maximum radius a sound may be placed at can be thought of as 1 – this is frequently referred to as the 'Unit Sphere'. If the sound is moved outside this sphere the directional information will not be decoded correctly and sounds will tend to pull to the nearest speaker.

Initially all transformations will place sounds on the surface of the unit sphere.

If a monophonic signal is to be placed on the surface of a unit sphere, then its coordinates will be, referenced to centre front;

x = Cos A * Cos B

y = Sin A * Cos B

z = Sin B

These coordinates directly relate to the B-format signal levels thus;

X = input signal * Cos A * Cos B

Y = input signal * Sin A * Cos B

Z = input signal * Sin B

W = input signal * 0.707

The 0.707 multiplier on W is there as a result of engineering considerations related to getting a more even distribution of signal levels within the four channels when taking live sound from a Soundfield microphone. A is the anticlockwise angle of rotation from the centre format and B is the angle of elevation from the horizontal plane. These multiplying coefficients, (Cos A * Cos B etc.), will position the monophonic sound anywhere on the surface of the soundfield, producing the B-format encoded output signals. These signals are equivalent to three figure-of-eight microphones at right angles to each other, together with an omnidirectional unit, all of which have to be effectively coincident over the frequency range of interest

SOUNDFIELD MANIPULATIONS

Definition of the coordinate system for B-format soundfield manipulations.

If a B-format signal is to be transformed, for example rotated and tilted, then the four channels of the signal must be scaled by the correct coefficients. The following standard definitions are made about the way the sound will move to a new position. They are provided in order to keep the equations coherent and minimise the confusion that can all too easily occur. Please keep to these conventions whenever discussing or using Ambisonic technology.

* Definition (1) Positive angles of rotation are anti-clockwise or by convention rotation to the left.
* Definition (2) A rotation is defined as a circular movement about a pre- defined axis, normally taken as the Z-axis, this being the same as an anti-clockwise movement in the horizontal plane.
* Definition (3) A tilt is defined as a rotation about an axis lying on the horizontal plane, for example the x-axis which is the same as an anti-clockwise movement in the vertical left-right plane.
* Definition (4) A tumble is defined as a rotation about the Y-axis. This is the same as an anti-clockwise movement in the vertical front-back plane. Note that a tumble is the same as a tilt that has first been rotated by 90 degrees about the Z-axis.

Figure 1 shows the graphical representation of this where A = the angle of rotation , B = the angle of elevation.

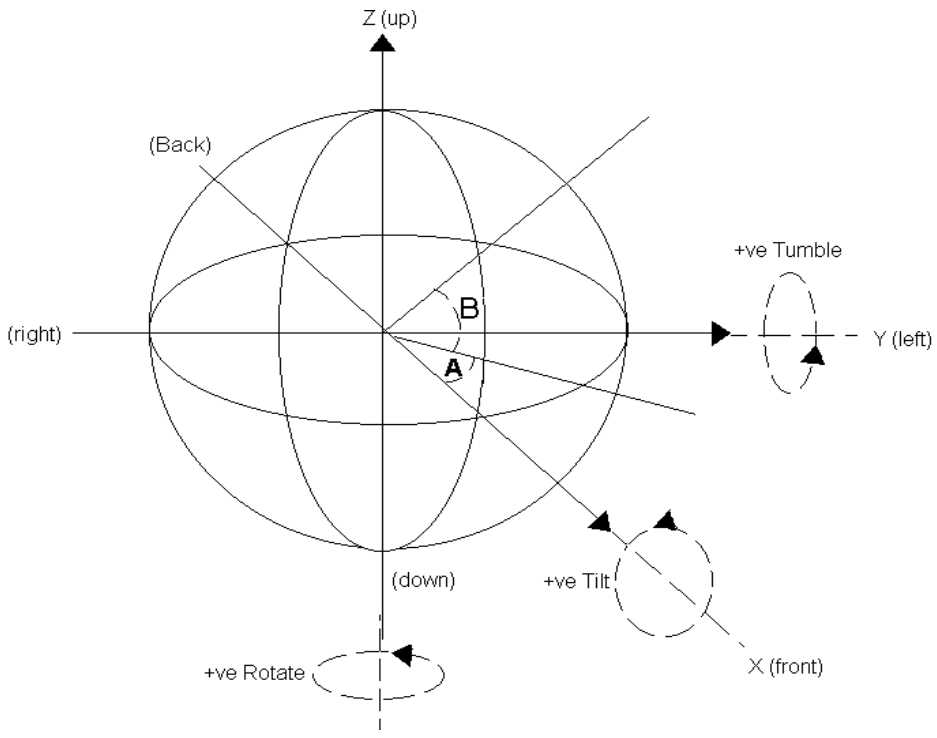

Figure 1. Basic transformations of the Ambisonic soundfield.

ROTATING A POINT ABOUT THE Z-AXIS

If A is the positive angle of rotation and C is the angle between the X-axis and the untransformed position, (x,y), we have;

$x = r*\cos C$, $y = r*\sin C$

$x' = r*\cos (A+C)$, $y' = r*\sin (A+C)$

simplifying;

x’ = r*cos C * cos A – r*sin C * sin A

y’ = r*cos C * sin A + r*sin C * cos A

and substituting for x and y

x’ = x * cos A – y * sin A y’ = x * sin A + y * cos A

w and z remain unchanged since the rotation is about the Z-axis, for points on the surface of the unit sphere w = 0.707. If the same procedure is applied to the tilt and rotate equations this gives the following;

TILT

x’ = x

w’ = w

y’ = y * cos B – z * sin B

z’ = y * sin B + z * cos B

TUMBLE

x’ = x * cos B – z * sin B

w’ = w

y’ = y

z’ = x * sin B + z * cos B

These equations can now be combined to perform transformations such as rotate-tilt which give an angular rotation of the whole input soundfield to the left by an angle of A from the centre front. Then it tilts the B-format soundfield by an angle B from the horizontal.

ROTATE-TILT

x’ = x * cos A – y * sin A

w’ = w

y’ = x * sin A * cos B + y * cos A * cos B – z * sin B

z’ = x * sin A * sin B + y * cos A * sin B + z * cos B

Any combination of the many possible soundfield manipulations can be realised by using one matrix of scaling coefficients thus;

X’ = K1.X + K2.W + K3.Y + K4.Z

W’ = K5.X + K6.W + K7.Y + K8.Z

Y’ = K9.X + K10.W + K11.Y + K12.Z

Z’ = K13.X + K14.W + K15.Y + K16.Z

where K1 – K16 are the scaling coefficients formed by the soundfield manipulations applied to the incoming signals X, W, Y, and Z. X’, W’, Y’ and Z’ are the resultant B-format output signals.

So far, we have only discussed sounds coded ‘on the surface of the Unit Sphere’. This somewhat counter-intuitive convention was forced on us by the technology (analog) which was available for (practical) use when Ambisonics was first developed. Whilst the Soundfield microphone (which mimics the ideal combination of three truly coincident figure 8 microphones plus an omni) preserves the distancing cues in a natural acoustic, if the position of a sound source is artificially constructed using analog technology it is very difficult to do more than just have the signal gradually become more and more diffuse as it moves off the surface of the sphere and in towards the listener. This is because, with conventional analog Ambisonic panpots, the omnidirectional (W) signal increases as the sound source moves towards the centre to compensate for a corresponding drop in level of the directional (X, Y, Z) signals. As a result, the image becomes more and more diffuse as the sound is panned towards the centre

In contrast to this, the levels of all four components increase as a real sound source approaches a soundfield microphone until at closest approach, the relevant directional components undergo a rather rapid phase (or more accurately polarity) reversal, whereafter the levels of all components start reducing again as the source moves away on the opposite of the microphone (Fig. 2). The optimisation of the precise law the amplitude curves should follow to match the subjective effects is still under investigation. However, it is important to realise that the perceived distance of an acoustic source is only weakly dependent on its loudness. Experiments in anechoic chambers have shown errors of more than two to one in subjects asked to guess the distance of a sound source. In fact, the cues we use for judging distance are significantly more complex. They include:

* The ratio of direct to reverberant sound – in a reverberant environment, the energy in the reverberant field stays more or less constant for all combinations of listener/source positioning, (so for a given source level, the reverberation loudness remains the same) whereas the source loudness drops off with increasing distance.
* The pattern of directions and delays for the first few (i.e. the early) reflections off surfaces in the environment. These change as source and/or listener positions change.
* Higher frequencies drop progressively more with distance, due to absorption by moisture in the atmosphere.
* The loudness drops off with distance.

The last two are heavily dependent on acquired knowledge of both the spectra and loudness of the sound source. For an electroacoustic composition where sounds may well not bear any relation to those the listener is used to, this poses interesting problems, or opportunities, if these cues are used on their own.

Figure 2 illustrates the relationship between the amplitudes of the directional signals x,y,z and the omnidirectional signal w as the sound source passes through the microphone.

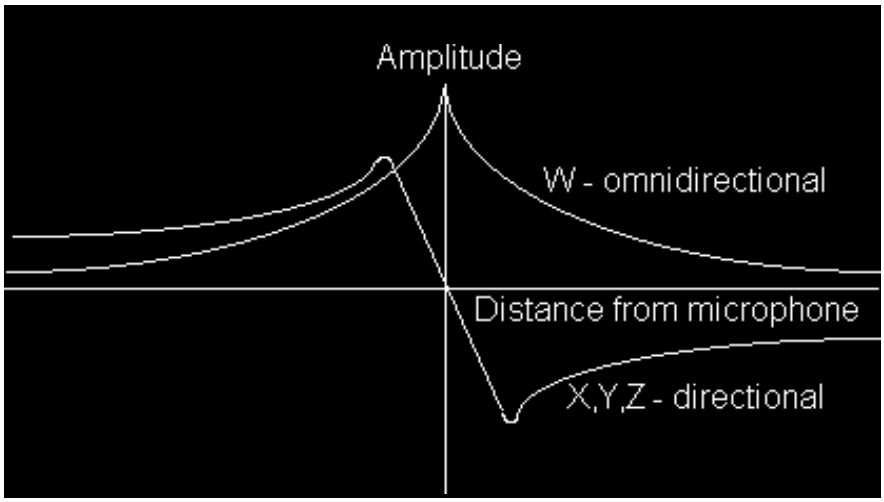

Figure 2. The relationship between the amplitudes of the directional signals x,y,z and the omnidirectional signal w as the sound source passes through the microphone.

AMBISONICS AND STEREO

The B format signals are not, of course, in any sense stereo compatible. It is, however, possible to combine the three (X,W,Y) components required for horizontal work in such a way that not only is a good stereo compatible two channel system produced but with a suitable decoder much of the original surround sound image can be recovered. The resulting (horizontal) soundfield is not perfect but by careful design of the encoding equations it is possible to place the defects in areas such as the rear image where the ear is less susceptible.

This encoding method, which is called UHJ coding, is used to produce stereo compatible Ambisonic records, tapes and broadcasts. The X,Y and W signals are matrixed into two channels using the following transform;

Left = (0.0928 + 0.255j)X + (0.4699 – 0.171j)W + (0.3277)Y

Right = (0.0928 – 0.255j)X + (0.4699 + 0.171j)W – (0.3277)Y

This would all seem relatively easy if it were not for the ‘j’ in the equation. What this indicates is that that particular signal is phase shifted by ninety degrees with respect to the ‘normal’ version of that signal, over the full audio band. In order to do that, each of these three signals must be passed through its own pair of wide-band phase shift (or all pass) networks. Within each pair, the output of one must be set up so that it has a phase shift that differs by ninety degrees from the output of the other member of the pair at all audio frequencies. This will give the required effect of a ninety degree phase shift. Existing encoders do this with analog circuitry but is entirely possible to write a computer program to do this or to implement the required filter equations in a digital signal processor.

This two channel member of the UHJ family of codings can be supplement with a third channel to remove the remaining anomalies for horizontal reproduction. This can be of reduced bandwidth without degrading things very far if it is necessary for operational reasons – for instance if transmitting it using subcarrier modulation on an FM transmitter. A fourth channel can be added to give height information. The decoding equations are such that a decoder for any of the levels will always extract the correct information from high level inputs – in other words the system is upward compatible.

The best reference for UHJ is Michael Gerzon’s article “Ambisonics in Multichannel Broadcasting and Video” in the Journal of the Audio Engineering Society, Vol. 33 No. 11, November 1985 pp. 859-871. In order to make the equations simpler, he gives them in terms of sum and difference signals, rather than left and right signals, but do not be put off by that.

Note that the original equations published in the patents had a factor of (0.3225 +- 0.00855j)Y rather than the (0.3277)Y quoted here, which are as published in Michael Gerzon’s 1985 article. According to an e-mail from Geoffrey Barton to the surrsound mailing list The formula (in the patents – DM) is

wrong. We changed the locus slightly sometime before 1980 to remove the ‘j’ term in y, thus saving a phase-shifter section and about 25% of the component cost of an encoder. All the commercially available encoders, including the Audio+Design units) used the modified version, all the Minim decoders and the Meridian are designed for this version.