Spatial Audio Toolbox v.1.0 Tutorial

Dimitar Kostadinov, Joshua D. Reiss
Email: josh.reiss@elec.qmul.ac.uk

1. Spherical Harmonics and Ambisonics

Spherical harmonics are functions in spherical polar coordinates (point in space is defined by r-distance, θ – azimuth angle and φ – elevation angle). They are used for investigating physical problems with spherical symmetry in three dimensions. They appear in quantum mechanics. Further they are used in representation of gravitational fields and magnetic fields of planetary bodies and stars. They are used in 3D computer graphics in almost every aspect of lighting. Spherical harmonic function is usually written as $Y_l^m(\theta,\phi)$, where l is the degree and m is the order of the function. In figure 1 are shown several visualizations of spherical harmonics functions.

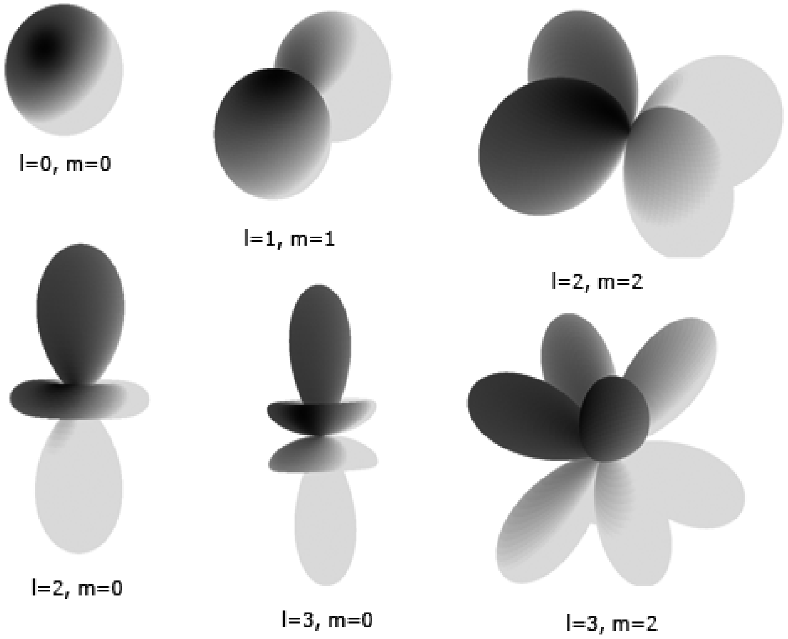

Fig. 1. Visualization of spherical harmonics.

Spherical harmonics form an orthogonal set of functions of order m and degree l.
The idea of using spherical harmonics in reproducing sound field is introduced by Micheal Gerzon. This technique is called Ambisonics. The idea is following: if we have a set of speakers which form a sphere, then using spherical harmonics for both encoding and decoding, we can recreate the sound field in the center of that sphere. This center is so-called sweet spot or listener's position. Each speaker should be equally distant from the center. In theory any reasonable set of speakers can be used by adding the proper gain and delay to each speaker.

Spherical harmonics can be described by following equations:

$$Y_{\ln}^k(\theta,\phi) = P_{\ln}(\sin\phi)\cos(n\theta) \text{ if } k=1 \qquad \textbf{(1)}$$

$$Y_{\ln}^k(\theta,\phi) = P_{\ln}(\sin\phi)\sin(n\theta) \text{ if } k=-1 \qquad \textbf{(2)}$$

P_{mn} is a semi-normalised associated Legendre function of degree **l** and order **n**. Associated Legendre functions are canonical solutions of Associated Legendre equation. It is a differential equation of second order. It is used in many areas of physics, chemistry, and applied mathematics with situations with spherical symmetry. If we seek for solution of Leplace's equation which is not necessarily axisymmetric we will come to the Associated Legendre equation and associated Legendre functions are such solutions. Each function is orthogonal to all of the others. This means that if we have infinite number of channels (infinite number of associated Legendre functions) we can define the sound field as it is in reality. Although this is not possible, using Spherical harmonics with higher order and degree we can describe the sound field more precisely.

Using equations (1) and (2) we can encode Ambisonics of any order using:

$$S_c = Y_{mn}^{k}(\theta,\phi)S_i \qquad \textbf{(3)}$$

S_i is the input signal. S_c is the encoded channels. Usually Latin alphabet is used W,X,Y,Z,T… for the encoded channels, but if we have higher than 4th order Ambisonics it is unsuitable to use. In equation (3) θ is the azimuth angle and φ is the elevation angle of the sound source. The distance from the listener and the sound source is equal to the radius of the sphere. Thus, we can only place a sound source on the sphere. This is not big issue since the direction of the sound source is more important for practical use. In the following table 1 are shown encoding weights to 3rd order.

Order	l,n,k	Channel	Natural(SN3D)	FuMa weights
0	0,0,1	S_1 (W)	1	$1/\sqrt{2}$
1	1,1,1	S_2 (X)	$\cos(\theta)\cos(\phi)$	1
1	1,1,-1	S_3 (Y)	$\sin(\theta)\cos(\phi)$	1
1	1,0,1	S_4 (Z)	$\sin(\theta)$	1
2	2,0,1	S_5 (R)	$(3\sin^2(\theta)-1)/2$	1
2	2,1,1	S_5 (S)	$(\sqrt{3}/2)(\cos(\theta)\sin(2\phi)$	$2/\sqrt{3}$
2	2,1,-1	S_5 (T)	$(\sqrt{3}/2)(\sin(\theta)\sin(2\phi)$	$2/\sqrt{3}$
2	2,2,1	S_5 (U)	$(\sqrt{3}/2)(\cos(2\theta)\cos^2(\phi)$	$2/\sqrt{3}$
2	2,2,-1	S_5 (V)	$(\sqrt{3}/2)(\sin(2\theta)\cos^2(\phi)$	$2/\sqrt{3}$
3	3,0,1	S_5 (K)	$\sin(\phi)(5\sin^2-3)/2$	1
3	3,1,1	S_5 (K)	$(\sqrt{3/8})\cos(\theta)\cos(\phi)(5\sin^2(\phi)-1)$	$\sqrt{45}/32$
3	3,1,-1	S_5 (M)	$(\sqrt{3/8})\sin(\theta)\cos(\phi)(5\sin^2(\phi)-1)$	$\sqrt{45}/32$
3	3,2,1	S_5 (N)	$(\sqrt{15}/2)\cos(2\theta)\sin(\phi)\cos^2(\phi)$	$3/\sqrt{5}$
3	3,2,-1	S_5 (O)	$(\sqrt{15}/2)\sin(2\theta)\sin(\phi)\cos^2(\phi)$	$3/\sqrt{5}$
3	3,3,1	S_5 (P)	$(\sqrt{5/8})\cos(3\theta)\cos^3(\phi)$	$\sqrt{8}/5$
3	3,3,-1	S_5 (Q)	$(\sqrt{5/8})\sin(3\theta)\cos^3(\phi)$	$\sqrt{8}/5$

Table 1. 3rd order Ambisonics coefficients and FuMa weights.

To avoid confusion we mention that Ambisonics order **M** is different than Legendre function order **n**.

Furse-Malham weights are correcting coefficients used for engineering purpose. Although they are not mathematically correct they are often used for 1st, 2nd and 3rd order Ambisonics.

In theory with higher order Ambisonics we have spherical harmonics of higher order and degree, thus the reproduction of the sound field will be with higher quality. As we mentioned above spherical harmonics are used with problems with spherical symmetry. In our situation we have a virtual sphere. With spherical harmonics of higher order and degree we can describe the sound field more accurately.

Using Ambisonics of higher than 3rd order would be difficult and very unpractical, because of the number of speakers needed. For example for 4th order we need 25 speakers, for 5th order – 36 speakers. The amount of speakers required can be given the formula:

$$K_m = (m+1)^2 \quad \textbf{(4)}$$

K_m is the minimum amount of speakers needed for order m. We can use as many speakers as we have, but they can't be less than required. For encoding higher order Ambisonics equation **(3)** can be used.

We have already encoded the channels. Now we have to generate a signal for every speaker using the encoded channels. We can do that using the following matrix equation:

$$S = C^{-1}.B = \frac{1}{n}C^{T}.B \quad \textbf{(5)}$$

S is the output signal vector. C is the re-encoding matrix. B is the vector of encoded channels. n is the number of speakers. The re-encoding matrix should be square with the size $n \, \mathrm{x} \, n$. Thus, the number of speakers must be the same as the number of encoding channels. In B-format (first order Ambisonics) we can use the following equation to describe output signal of each speaker:

$$p_j = \frac{1}{L}(W + X(\cos\theta_j \cos\phi_j) + Y(\sin\theta_j \cos\phi_j) + Z(\sin\phi_j)) \quad \textbf{(6)}$$

In outher words equations (5) and (6) represent the contribution of each speaker to the sound field. In equation (6) $\mathrm{p_j}$ is the output signal, L is the number of speakers, θ_j is the azimuth angle of each speaker, ϕ_j is the elevation angle of each speaker, W, X, Y, Z are the encoded channels in B-format. We can use as many speakers as we want. Equation (6) can be extended to higher order Ambisonics. It can be written as follows:

$$p_j = \frac{1}{L}(K.B) \qquad \textbf{(6.1)}$$

K is the decoding vector of coefficients equal to the encoding vector if the speaker's position was the sound source's position.

Speaker arrangements are an important part of reproducing the sound field. Different types can be used. In general they must be on the virtual sphere surface. In figure 2 and 3 are shown two types of speaker arrangements. The order of Ambisonics is 14th for better visualization.

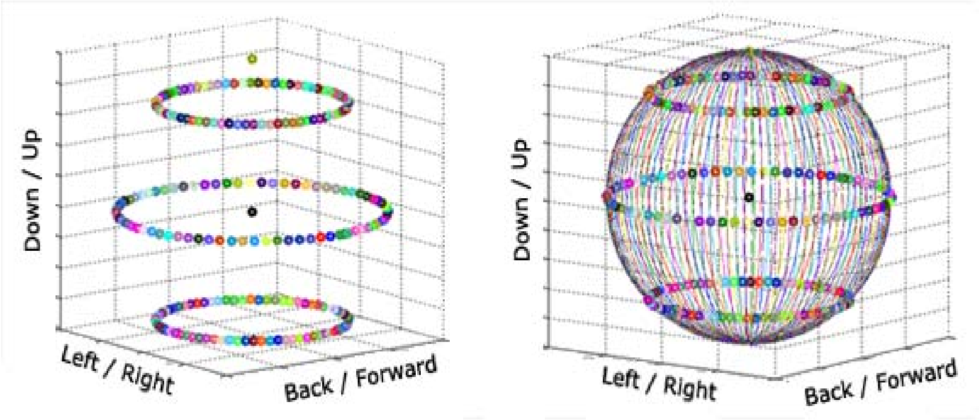

Fig. 2. Speaker arrangement in three rings with rotation of azimuth angle.

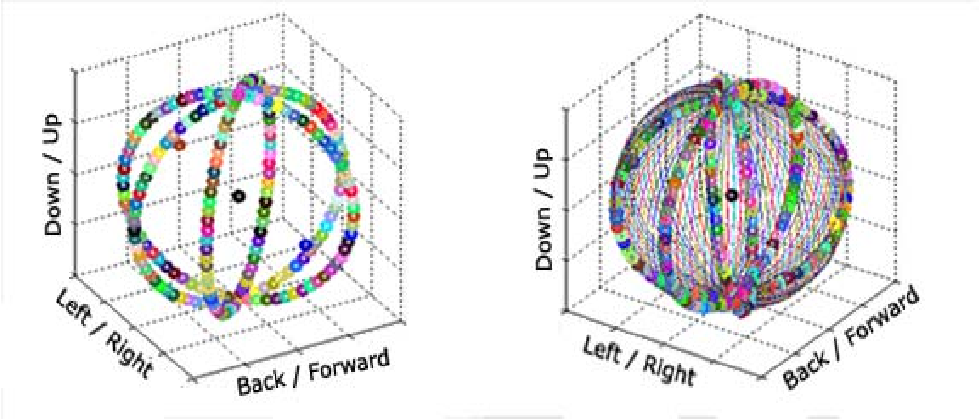

Fig. 3. Speaker arrangement in three rings with rotation of elevation angle.

Both types of the arrangements are with equal elevation and azimuth angle step change. It is essential that in practical situation we can’t have these accurate arrangements. The black dot in the middle as shown in figure 2 and figure 3 is the listener’s position.

2. Sound panning in stereophony. Vector Based Amplitude Panning

First we will present the amplitude panning in stereophony. We must mention that usually audio engineers change the gain levels of the speakers until they perceive the sound from desired location. However sometimes automated sound source panning is needed. The relation between the gains and the desired location is given with equation **(11)**, where θ is the angle to the virtual sound source and θ_0 is the angle to the speaker. This is shown in figure 4.

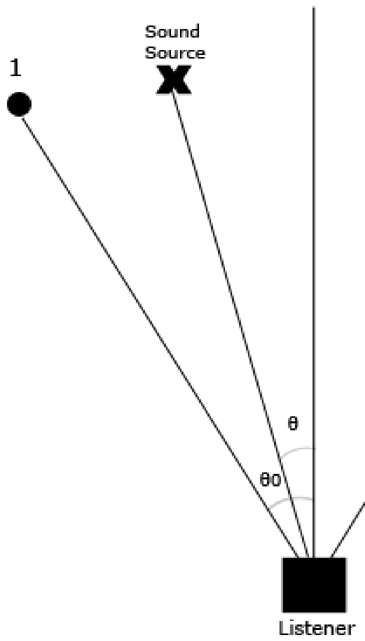

Fig. 4. Panning a sound source in stereophony. θ is the angle to the virtual sound source and θ_0 is the angle to the speaker.

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \qquad \textbf{(11)}$$

Equation **(11)** gives only the ratio between gain factors. They cannot be directly calculated from it. To find the gain factors we must normalize them using equation **(12).**

$$\sqrt{g_1 + g_2} = 1 \quad \textbf{(12)}$$

This sound panning is in stereophony with 2 speakers. If we want to pan a sound in 3D space, we might use VBAP.

Vector based amplitude panning is a technique for panning a sound source in 3D space using 3 speakers. In this technique speaker position and sound source position is defined by vectors. The “0” of the coordinate system is the listener position as shown in figure 6.

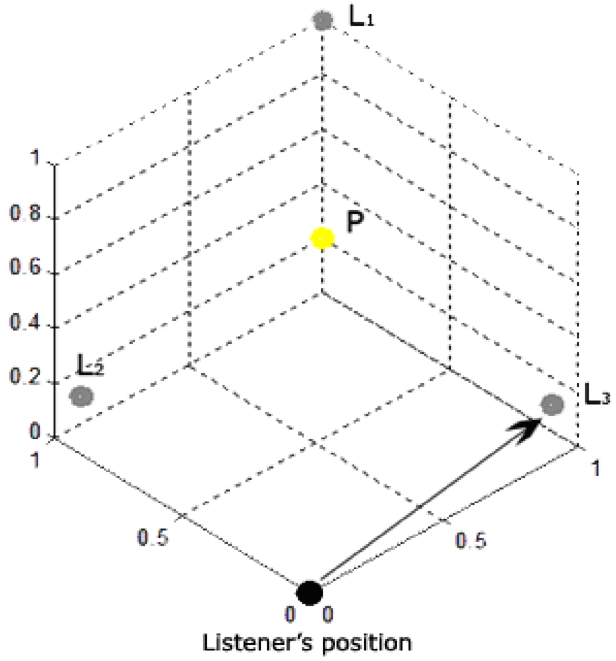

Fig. 5. Panning a sound source within speakers array in VBAP. P is the sound source. L_1, L_2 and L_3 are speakers.

$L_1(x_1,y_1,z_1)$,$L_2(x_2,y_2,z_2)$,$L_3(x_3,y_3,z_3)$ are vectors which point to each of the 3 speaker locations. $P(x_p,y_p,z_p)$ is a vector which points to the sound source. The idea is that the signal can appear coming from point P if each speaker has gain according to its position in Cartesian coordinate system. In other words each speaker has a contribution to the sound field, according to its position. For example, if $P \equiv L_1$, then the gain of speaker L_1 will be 1. All other gains must be 0. The gain for each speaker is calculated using equation **(13)**

$$\begin{aligned} &P = g_1 L_1 + g_2 L_2 + g_3 L_3 \\ &P^T = g L_{123} \\ &g = P^T L^{-1}{}_{123} \end{aligned} \qquad \textbf{(13)}$$

The gains must be normalized using equation **(12)**, in order to have values lower than 1.

In theory can use any 3 speakers to pan a sound source, but it is essential to use the closest 3 to the sound source. In some applications we might want to use more than 3 speakers to pan a source. In order to do that we can chose another triplet of speakers in such a way that none of the speakers should be used in both triplets. The gains can be calculated using equation **(13)** for both triplets separately. To normalize them we can use equation **(14)**.

$$\sqrt{\sqrt{\sum_{n=1}^{3} {g_n}^2}^{\,2} + \sqrt{\sum_{m=1}^{3} {g_m}^2}^{\,2}} = \sqrt{\sum_{n=1}^{6} {g_n}^2} = 1 \qquad \textbf{(14)}$$

With VBAP we have certain restrictions such as amount of used speakers. We can use either 3 speakers or an amount multiple of 3. Using more than 1 triplet for one sound source however can be difficult and undesirable sometimes. In the next section we will present method in which many speakers can be used for panning a sound source.

3. Vector Distance Panning

In this technique any amount of speakers can be used for panning a sound source. Each speaker is defined with a vector pointing to it. In The Cartesian coordinate system speaker S$_1$ has a vector $\overrightarrow{S_1}(x, y, z)$ which points to its location. The zero of the coordinate system is the Listener's position as shown in figure 6.

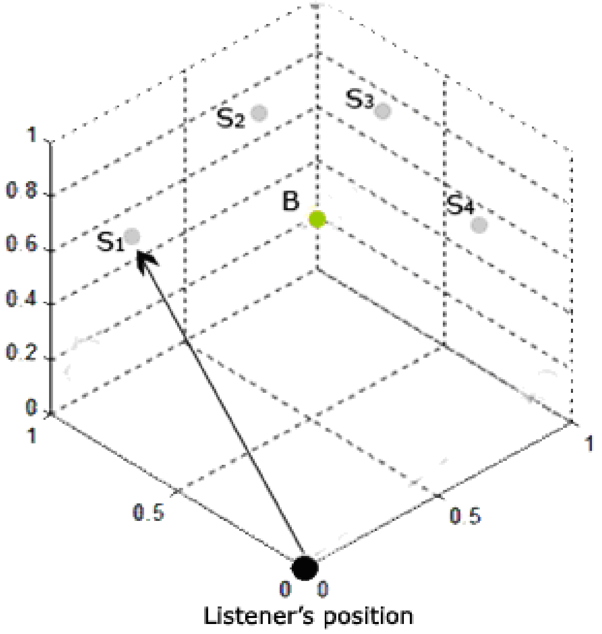

Fig. 6. Panning a sound source within speakers array with VDP. B is the sound source. S1, S2, S3 and S4 are speakers.

The gain for each speaker can be calculated in two steps. First we must find the distance between the speaker and the sound source using equation (15).

$$D_n = S_n B = \sqrt{(X_{Sn} - X_B)^2 + (Y_{Sn} - Y_B)^2 + (Z_{Sn} - Z_B)^2} \quad \textbf{(15)}$$

D$_n$ is the distance between speaker and the sound source. S$_n$ is the speaker position. B is the sound source. We need to calculate the gains for them. To do that, we need to find the closest speaker to the sound source:

$H=\min(D_n)$ **(16)**

H is the distance to the closest speaker. Essentially, the closest speaker must have the highest gain. The gain factor for each speaker can be calculated using equation (17).

$$g_n = \frac{H}{D_n} \qquad \textbf{(17)}$$

The gain of the closest speaker will be 1. All other gains will be lower than 1. The gain of every speaker represents its ratio to the closest one. This means that the speaker closest to the sound source will have highest gain. In order to have real gains instead of ratios, we need to normalize the gains by using equation (18). This is also done in order to have balanced system with constant energy.

$$\sqrt{\sum_{n=1}^{N} {g_n}^2} = 1 \quad \textbf{(18)}$$

N is the number of speakers that we are using.

The gains for each speaker are independent of the listener's position as it can be seen. Only the distances to the virtual sound source are important. The listener's position is important only for the time delay, which will be discussed in the next section.

It can be seen that even if we don't know the exact location of each speaker in the Cartesian coordinate system, the distance can be measured by any other suitable approach in stead of using equation (15). This means that we have the opportunity to arrange the speakers in particular room as we need to, without creating a virtual coordinate system. This technique is useful, as it can be used in almost every practical situation. The number of speakers is not restricted. Their positions are flexible. This technique is similar to VBAP, where gain factors are derived from speaker's position also. But rather than using directional components of the vectors, here we use the distance as a whole to calculate the gains.

4. Time Delay in VBAP and VDP

We need to assure that the sound from each speaker will arrive at the same time to the listener. The proper delay should be added. Equation (19) can be used to calculate the delay in samples.

$$G_n = \frac{M - D_{0n}}{V_s \cdot \frac{1}{fs}} \qquad \textbf{(19)}$$

D_{0n} is the distance from each speaker to the listener's position. It can be calculated using equation (20).

$$D_{0n} = S_n O = \sqrt{(X_{Sn} - X_O)^2 + (Y_{Sn} - Y_O)^2 + (Z_{Sn} - Z_O)^2} \quad \textbf{(20)}$$

In equation (19) M is the distance to the furthest speaker, V_s is the Speed of Sound, fs is the sample rate. G is the delay given in samples, n is the speaker number.

5. References

[1] Ville Pulkki and Matti Karjalainen, "*Multichannel Audio Rendering Using Amplitude Panning*", IEEE Signal Processing Magazine [118] May 2008

[2] Roger K. Furness, *"Ambisonics - An Overview" ,* AES 8th International Conference

[3] Vllle Pulkki, *"Virtual Sound Source Positioning Using Vector Base Amplitude Panning"*, J. Audio Eng. Soc., Vol. 45, No. 6, 1997 June

[4] U. Zölzer, *DAFX: Digital Audio Effects,* Wiley, 2002.

[5] D. Malham, "*Higher order Ambisonics systems"*, http://www.york.ac.uk/inst/mustech/3d_audio/higher_order_ambisonics.pdf, 2003.

[6] Dave Malham, "*Higher order Ambisonic systems*", abstracted from *"Space in Music - Music in Space",* an Mphil thesis by Dave Malham, submitted to the University of York in April 2003, revised and passed in December 2003

[7] *Martin Neukom, Jan C. Schacher, "AMBISONICS EQUIVALENT PANNING*"**,** Zurich University of the Arts Institute for Computer Music and Sound Technology

[8] Jérôme Daniel, Rozenn Nicol , and Sébastien Moreau, "*Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging*"**,** 114th Convention 2003 March 22–25 Amsterdam, The Netherlands

[9] Martin J. Morrell and Joshua D. Reiss, "*A Comparative Approach to Sound Localisation within a 3D Sound Field*"**,** 126th Convention 2009 May 7–10 Munich, Germany

[10] A. C. King, J. Billingham, Stephen Robert Otto, *"Differential equations"*

[11] Adomas Paltanavičius "*Spherical Harmonics Generator*", http://adomas.org/shg/