Ambisonics for Loudspeakers

Eetu Backman
Aalto Universtiy
Bachelor's Programme EST / AAT

eetu.backman@aalto.fi

Abstract

Today exists mainly three surround sound formats, that can be implemented for loudspeaker layouts: channel based, scene based and object based. Channel based surround sound being the oldest format still in use, scene and object based audio formats represent the state of the art methods for reproducing three dimensional sound accurately for human perception.

In this literature review a scene based format called the ambisonics is represented. First, general theory of ambisonics is regarded consisting of brief introduction to ambisonic's mathematical concept, ambisonic format, sound capture and synthesis methods. Finally, current knowledge of ambisonic decoding methods and their effects on performance is revised.

1 Introduction

After the invention of surround sound, methods for recording, processing, saving, and playing back spatial audio needed to be developed. Creating a robust data format was essential for generalizing surround sound production and consumption. Currently, three different methods for capturing and encoding spatial audio information exist: channel-based audio, object-based audio, and ambisonics. In this article the current knowledge of ambisonics and it's decoding methods for loudspeakers will be reviewed

Michael Gerzon discovered and developed the first order ambisonics in the early 1970s. The ultimate idea behind ambisonics is to have a format for three dimensional audio, that holds the audio data in predetermined spherical harmonics and thus being independent of sound reproduction setup. [11] Although Gerzon's first order ambisonics did not lift off for commercial utilization due to it's lack of directional accuracy, the ambisonics in general has been under frenzied research and improvement the past 50 years.

The study on Ambisonics for Loudspeakers will commence with a recap to ambisonics theory, which is important to understand before studying it's adaptation on loudspeaker setups. After theory section the practical methods of realizing an ambisonic system will be summarized. Next different decoding methods for loudspeaker setups will be reviewed. Finally, the study will investigate the localization accuracy of ambisonics with different decoding and loudspeaker setups. Also the understanding of coloration effects caused by the decoding and loudspeaker setup will be examined.

2 Introduction to Ambisonics

In this section the basic theory and workflow of an ambisonics system is summarized. First the background and theory will be initialized based on mathematical grounds. Next a brief introduction on the workflow, B-format and encoding methods will be rehearsed and last but not least, the theory behind main decoding methods will be introduced.

2.1 Ambisonics workflow

The creation of ambisonic audio starts either from recording a scene using first- or higher order ambisonic microphones, or by synthesizing mono audio tracks to be included in the audio data. Then measured or created raw data will be encoded to B-format using an encoding method depending on the chosen order and type of raw data obtained. The B-format, discussed in more detail later, is the format, that will store and carry the data to be reconstructed. Last, the B-format will be decoded to loudspeaker signals and reproduced with a suitable loudspeaker setup. Figure 1 visualizes the data flow in ambisonics.

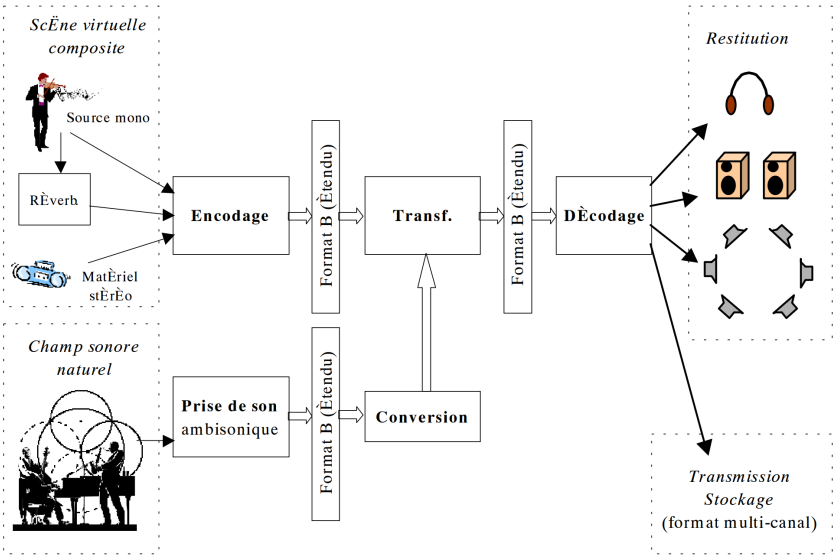

Figure 1: Visualization of ambisonic data flow. Acquired from[2].

2.2 Theory of Ambisonics

The theory of ambisonics begins from the problem of representing an arbitrary, three dimensional sound field mathematically so that it can be divided to it's spatial components, manipulated and again assembled with the use of loudspeakers or headphones. Solving the problem starts from incorporating the general wave equation 1, which is represented in spherical coordinates as acoustic waves are isotropic and involve spherical forms naturally. By deriving the Laplacian operator for spherical coordinates as in equation 2, also the wave equation will be solved for spherical components. [18]

$$\frac{\partial^2 p(t,r,\theta,\phi)}{\partial t^2} = c^2 \nabla^2 p(t,r,\theta,\phi) \tag{1}$$

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 \frac{\partial}{\partial r}) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}(\sin\theta \frac{\partial}{\partial \theta}) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \phi^2} \tag{2}$$

The general function of the acoustic pressure p is a function of time and space. The temporal and spacial behaviors of an acoustic wave are independent of each other and thus can the temporal and spacial components be solved separately. Because the problem to solve was explicitly to represent the acoustic wave in spherical coordinates, the temporal component can be omitted here.

$$\nabla^2 p(r,\theta,\phi) = k p(r,\theta,\phi) \tag{3}$$

After omitting the temporal component of the wave equation, we are left with solving the Helmholtz equation (equation 3) in spherical coordinates, which describes

the sound wave's behavior only in spatial domain. This can be divided further to it's radial, horizontal and vertical components through again separation of variables. Solutions for the radial component are spherical Bessel functions $j_n(kr)$ and solutions for the angular part make up the before mentioned spherical harmonics $Y_{nm}(\theta, \phi)$. Spherical harmonic functions' (equation 5) first component is the azimuthal oscillatory component (the spherical harmonics are periodical in spherical coordinates along the azimuth or horizontal angle). The second part is the polar component, which is described by the normalized associated Legendre polynomials [12]. B_{nm} represents the sound pressure at the boundary of the sphere without sources within and can be considered as the B-format signals of degree n and order m. Here degree n corresponds to ambisonics order [3].

In a real life scenario, the physical sound waves follow the solution of the wave equation with an infinite degree (ambisonic order). How ever for the ambisonic system, the solution must be truncated to a finite degree determined by the amount of microphones, channels or loudspeakers available. The truncation of the ambisonic order causes localization and coloration issues, which can be minimized to some degree in decoding.

$$p(r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} B_{nm}(k) j_n(kr) Y_{nm}(\theta, \phi) \tag{4}$$

$$Y_{nm}(\theta, \phi) = e^{in\theta} N_{nm} P_{nm} cos(\phi) \tag{5}$$

2.3 Format of Ambisonics

Ambisonics is a surround sound format, in which a recorded or synthesized sound field is divided (encoded) to it's orthogonal sound velocity components and again assembled (decoded) to form the original sound field via headphones or loudspeakers as accurately as possible. The data format of ambisonics is called the B-format. Each channel in B-format holds the audio captured by the (real) spherical harmonic determined for the channel in question. For example if a single sound source is rotated around an ambisonic recording setup, the B-format's X-channel will hold a signal of the original sound source, scaled by the x-directional figure-8 spherical harmonic. Thus a complete sound field can be constructed from the spherical harmonics with a resolution determined by the order of the system as mentioned above.

For each order exists a set of spherical harmonics, which are the solutions for the wave equation in spherical coordinates (equation 1). Each of the solutions is orthogonal over all orders and degrees, which means that for capturing and producing an original sound field with infinite spatial resolution requires the order of the system to be infinite [15]. Thus can be said, that any realistic ambisonic system is a truncation in spatial, spherical harmonics, just as a Fourier series can be a truncation of an arbitrary wave form in sinusoidal harmonic components. In figure 2 can be seen the three dimensional plots of the real spherical harmonics $Y_{nm}(\theta, \phi)$ up to third order.

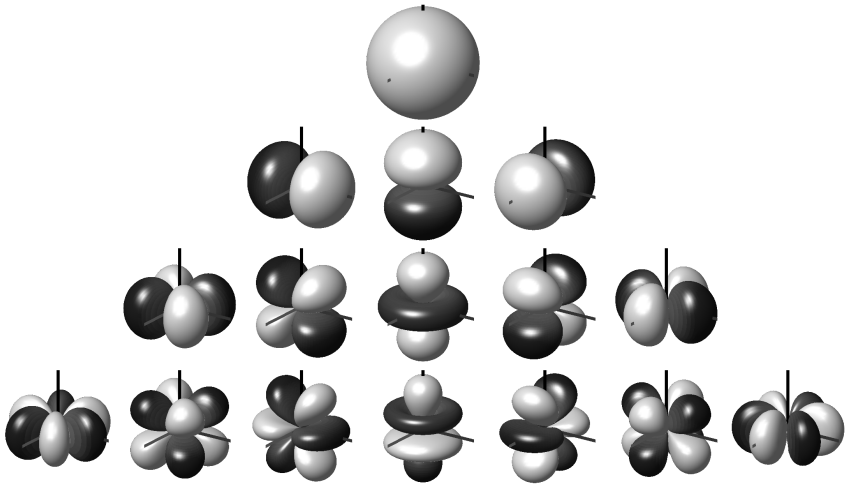

Figure 2: Real spherical harmonics and their modes up to third order. Ambisonics order corresponds to the order of the spherical harmonics. Imported from Ambisonics Wikipedia page. [1]

First order B-format is the simplest version and contains four channels; W, X, Y and Z. This is the lowest amount of channels with which a three dimensional sound field can be captured in ambisonics. [8]

While the first order B-format includes enough spherical harmonics to represent the entire spherical sound field, higher orders add velocity components in between the main lobes to allow for a single lobe to be more directional. The higher the order of the system, the more there are spherical harmonics and thus the better is the directional accuracy of the dataset.

2.4 Sound Capture for Ambisonics

To create a B-format file of a sound scene, the sound needs to be first captured either by recording or by synthesis. Here the basic knowledge of ambisonic recording is revised.

The B-format being composed of spherical harmonics of a sound scene, it is beneficial to record with microphone layouts, that allow for an easy conversion to those spherical harmonics. For a first order ambisonic system, two preferable microphone layouts exists: triple-MS and tetrahedral layouts [19]. The triple MS layout consists of three perpendicular figure-8 microphones and one omnidirectional microphone. This layout directly maps the physical sound scene to the channels of a first order B-format, making it a compelling option as no encoding is needed. The tetrahedral layout consists of four cardioid microphones in a regular pattern called the tetrahedron, which is the first of five platonic solids. Each B-format channel is a linear combination of the four microphone signals as can be seen in figure 3

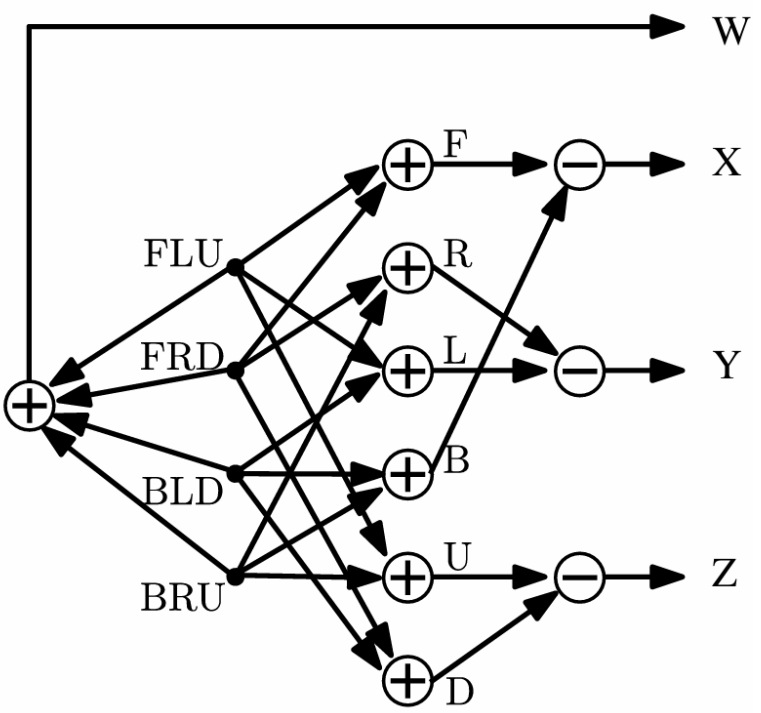

Figure 3: Encoding process of tetrahedral microphone signals to first order B-format. Imported from [19]

2.5 Sound synthesis for Ambisonics

In ambisonic sound synthesis, a mono source is located to a virtual sound field by sampling it with the spherical harmonics of the B-format. This can be regarded the same as recording a physical sound source with an ambisonic microphone, that directly maps the sound to the B-format. A first order B-format can be created from a mono source by multiplying it with the real spherical harmonics of the zeroth and first order as in the equations one through four. As can be seen, the multipliers are the same figures of eight as the pick up patterns of first order ambisonic microphones.

$$W(t) = \frac{s(t)}{\sqrt{2}} \tag{6}$$

$$X(t) = s(t) \cos\phi \cos\theta \tag{7}$$

$$Y(t) = s(t) \sin\phi \cos\theta \tag{8}$$

$$Z(t) = s(t) \sin\theta \tag{9}$$

2.6 Audio Decoding to Loudspeakers

Decoding for loudspeakers in ambisonics means the process, in which the signals transmitted for the loudspeakers are determined based on the B-format and the positions of the loudspeakers in the setup in question. The purpose of a decoding algorithm is to map the channels of the B-format to loudspeaker arrays, ensuring accurate spatial rendering. As the B-format is encoded the same, regardless of the loudspeaker setup, the decoder must have the knowledge of loudspeaker positioning.

In Ambisonics, the time differences between loudspeaker signals do not need to be explicitly created. This is because interaural time differences (ITDs) are naturally perceived from sound physically arriving from specific directions. In addition human brain has been discovered to perceive ITDs at low frequencies from signal level differences [5]. This allows ambisonics to create realistic localization cues without incorporating synthesized signal time differences, but only using amplitude panning methods.

Creation of the decoder starts from the principle that a sound field is reconstructed from a linear combination of all spherical harmonics available in the B-format of order N. Each of L loudspeakers denoted by l receives the linear combination with channel specific weights accordingly $s_l = \alpha_l W + \beta_l X + \gamma_l Y + \delta_l Z....$ The main challenge in decoding is to determine these weights for each loudspeaker to optimize the accuracy in reproduction, while minimizing defects caused by the truncation of the ambisonic order.

First ambisonic decoding method for loudspeakers is called the mode matching method or MMAD. The MMAD is based on analytically determining the loudspeaker signals, that would recreate the sound field that was originally recorded. In other words: If the B-format can be acquired by encoding the sound field with the spherical harmonics ($B = C * p$), the signals to be emitted by the loudspeakers can be acquired by decoding the B-format by the inverse of the encoding matrix C. the original sound field and the sound field generated by the loudspeakers being interchangeable

$$B = C * p_{acquired} \tag{10}$$

$$C^{-1} * B = p_{reproduced} \tag{11}$$

$$\implies C^{-1} * C * p_{acquired} = p_{reproduced} \tag{12}$$

$$\implies p_{acquired} = p_{reproduced} \tag{13}$$

If the loudspeaker setup contains more loudspeaker channels than the B-format contains directional channels, the decoding matrix is no more rectangular, rather under

determined, making its inversion impossible. Here the least-squares approximation is used as the decoder matrix instead and is called the Moore-Penrose inverse.

Special case for MMAD occurs, when the amount of B-format channels equals the amount of loudspeakers and the loudspeaker configuration is regular, thus making the encoding matrix rectangular and orthogonal. When these requirements are met, the inversion of the encoding matrix becomes it's transpose and finding the decoding matrix becomes extremely easy. This decoding methods is also known as the sampling method, as each loudspeaker receives a sample of every spherical harmonic according to the loudspeakers angular position.

$$D = C^{-1} = C^T \tag{14}$$

Mode Matching decoding method can be adapted to create the energy preserving ambisonic decoding method. The energy preserving method retains direction independent energy, when using irregular loudspeaker setups. This is achieved by forcing the pseudo-inverse decoding matrix to be orthonormal, when calculating the singular value decomposition. Orthogonality is forced by removing the singular value matrix from the singular value decomposition of the pseudo-inverse decoding matrix, thus removing any energy scaling factors.[7]

One of the most recent ambisonic decoding methods is called the all-round ambisonic decoding or ALLRAD [9]. In ALLRAD original ambisonic channels are first mapped to a virtual loudspeaker setup, in which the loudspeaker positioning obey a regular design. This is done to preserve energy- and mode-matching, which is only achiavable using the t-design. Lastly, the virtual loudspeakers are mapped to physical ones with Pulkki's vector base amplitude panning, thus allowing for irregular loudspeaker setups[16].

3 Perceived localization accuracy

A virtual source played back through loudspeakers has two perceivable characteristics related to source location: source location and source width. These characteristics are angle dependent due to non-ideal qualities of an ambisonic system such as truncation of the system order and irregularity of the loudspeaker layout. In this chapter existing research of comparison between different decoding methods is revised regarding localization related defects.

In his dissertation [13] Mr. Matthias studied the effect of decoder order weighting on localization accuracy. The study was conducted through listening tests in a regular loudspeaker layout with two seating positions, in the center and off-center. Also Mr. Pulkki's vector based amplitude panning and multiple direction amplitude panning methods were in the comparison pool.

While the listening position was in the center, all of the four panning methods had great results for localization accuracy. Only when off-center position was used, great deviation between perceived and actual source location was measured. Between the two weighting methods, $maxr_E$ weighting had greater results, while $maxr_v$ weighting resulted in high deviation, split perceived direction and low confidence across tests. This coincides with the theory of side lobe minimization resulting in a wider acceptable sweet spot. However the use of $maxr_E$ weighting cannot replace use of center listening position. While in theory in-phase weighting would result in the widest sweet spot due to complete removal of the side lobes, the main lobe becoming excessively wide results in the worst localization accuracy [6].

The importance of loudspeaker layout regularity has been shown in [4], [7] and [10]. In [10] error in perceived direction caused by a semicircular layout was compared between sampling, energy preserving (EPAD) and ALLRAD decoding methods. The results were only simulated with an energy vector model, which in [13] was considered to only loosely correlate with actual human perception of direction. How ever the simulated results suggest that mapping error can be corrected for by EPAD and ALLRAD, but a missing loudspeaker or irregular aperture angle between loudspeakers results in direction dependent source width with all decoding methods. Also in Mr. Matthias's dissertation was found that aperture angle limits the narrowest possible perceived source width.

The localization accuracy of ambisonic decoding and loudspeaker systems was mainly estimated with directionality vectors for acoustic velocity $\boldsymbol{r_v}$ and energy $\boldsymbol{r_E}$, which are determined as follows: [2]

$$\boldsymbol{r_v} = \frac{\sum_{l=1}^{L} g_l \boldsymbol{\theta_l}}{\sum_{l=1}^{L} g_l} \tag{15}$$

$$\boldsymbol{r_E} = \frac{\sum_{l=1}^{L} g_l^2 \boldsymbol{\theta_l}}{\sum_{l=1}^{L} g_l^2} \tag{16}$$

In [13] was noticed, that both direction models coincide only loosely with listening tests regarding perceived panning angle. How ever the accuracy of the energy vector model could be improved by accounting for direction dependency of loudness perception caused by the shape of human head and ears. This could be implemented by weighting the vector components with weight values mapped from a head related transfer function.

4 Coloration effects of Ambisonics

In addition to a limited angular resolution, another problem rises from an ambisonic system being of finite order; spatial comb-filtering [14]. In ambisonics, a spatial comb filter is created due to the destructive and constructive interference of pressure waves

arriving from adjacent loudspeakers, that reproduce together coherent signals. The comb filtering then can be detected as severe coloration of sound, especially when listener or sound image is moving [17], or when the listener is away from the physical sweet spot[13]. The amount of coloration is affected mainly by the amount of loudspeakers used [17] [19]. The more loudspeakers exists for given ambisonic system order, the more of them play back coherent signals in unison. This results in denser interference pattern and thus higher modulation frequency of changing coloration across panning angle [19]. Then the coloration can be reduced with the cost of reducing the localization accuracy by decreasing the number of loudspeakers. Reduction in the number of loudspeakers causes the remaining loudspeakers to be further apart, which decreases the frequency and intensity of the perceived coloration. Here a reverberant room can actually help in coloration reduction by reducing the depth of the comb filter[14].

Another way to reduce the coloration effect is to window the spherical harmonics as a function of order [13]. Reduction of side lobes reduce the number of loudspeakers playing back coherent signal, resulting in less coloration. Between $maxr_v$ and $maxr_E$ weighted decoders, $maxr_E$ resulted in less coloration for listening position in the center and off center.

5 Conclusions

Ambisonics has come a long way after invented by Gerzon 50 years ago. The invention of energy preserving and ALLRAD methods have made the use of ambisonics more feasible for irregular layouts, which has been one of the hardest problems to come by. Also the possibility to create higher order ambisonics helps with creating more accurate representation of the original sound field with a sweet spot remaining for higher frequencies. Still we are far from being able to identically recreate a sound field especially for a wider audience than one person, as ambisonics always remains a truncation of the ideal world. For a human head sized sweet spot to have 16 kHz bandwidth would need a regular layout of almost 1000 loudspeakers, thus making reaching for the ideal both expensive and troublesome, for now.

Use of a regular loudspeaker layout produces automatically orthogonal decoding matrix with mode matching and thus sampling decoders, which also creates the optimal localization accuracy. Moving out of the regular layout, sampling and MMAD are no more panning invariant and have poor localization accuracy, meanwhile energy preserving and ALLRAD methods achieve the smallest errors in panning angle. Use of regular loudspeaker layout and central listening position remains the most effective methods for creating good localization accuracy. Furthermore this permits the use of a sampling decoder, making the decoding process effortless as (pseudo-)inverse will not be needed.

What comes to coloration, use of $maxr_E$ weighting is recommended for the coloration to be less distracting. Also keeping the amount of loudspeakers minimal in the

limits of the system order helps by reducing the angular modulation frequency of the coloration effect. Lastly, a reverberant room reduces the depth of the comb filter effect, resulting also in less distracting coloration.

References

[1] Ambisonics - wikipedia, 2024.

[2] Arteaga, D. Introduction to ambisonics.

[3] Boehm, J. Decoding for 3d.

[4] Eric M. Benjamin, Richard Lee, A. J. Localization in horizontal-only ambisonic systems.

[5] F. Matthias, F. Zotter, A. S. Producing 3d audio in ambisonics.

[6] F. Matthias, F. Z. Localization experiments using different 2d ambisonics decoders.

[7] F. Zotter, H. Pomberger, M. N. Energy-preserving ambisonic decoding.

[8] Florian, H. An introduction to higher-order ambisonics.

[9] Franz Zotter, Matthias Frank, A. S. The virtual t-design ambisonics-rig using vbap.

[10] Franz Zotter, Frank Matthias, H. P. Comparison of energy-preserving and all-round ambisonic decoders.

[11] Gerzon, M. Periphony with-height sound reproduction.

[12] Jorge, T. A formulation of ambisonics in unconventional geometries.

[13] Matthias, F. Phantom sources using multiple loudspeakers in the horizontal plane.

[14] Matthias, F. How to make ambisonics sound good.

[15] Poletti, M. Three-dimensional surround sound systems based on spherical harmonics.

[16] Pulkki, V. Virtual sound source positioning using vector base amplitude panning.

[17] Solvang, A. Spectral impairment of two-dimensional higher order ambisonics. 267–279.

[18] Young, P. Helmholtz's and laplace's equations in spherical polar coordinates: Spherical harmonics and spehrical bessel functions.

[19] Zotter, F., and Frank, M. *Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality.* Springer-Verlag, 2019.