February 07, 2019

Symmetric Ear-Related Ambisonics Rendering

Alper Güngörmüsler

Andrew Allen

Symmetric Ear-Related Ambisonics Rendering

Abstract:

Immersive virtual reality (VR) and augmented reality (AR) environments rely on high-quality audio. To this end, VR and AR platform developers have adopted Ambisonics, which is a technique used to record, modify, and recreate a full-sphere surround sound. To render (or decode) the sound field as faithful as possible, developers often use a set of head-related transfer functions (HRTFs). However, the use of HRTFs, which are processed with truncated spherical harmonics, increases errors with increase in frequency scale. A new method for rendering spatial audio ambisonically that handles each ear independently shifts the paradigm. The conventional HRTF ambisonic encoding is modified to compute the spherical harmonic coefficients for a set of ear-related transfer functions (ERTFs), which enable an increase in sound quality encoding and rendering by using fewer ambisonic orders.

Keywords: Ambisonics, head-related transfer function (HRTF), ear-related transfer function (ERTF), B-format, high-order Ambisonics, virtual reality (VR), augmented reality (AR)

Background:

Immersive VR and AR environments rely on high-quality audio. To this end, many developers have adopted a technique called Ambisonics. While derivatives of this technique have been around since the 1970s, the interest to create immersive VR or AR environments has refueled interest and research in Ambisonics.

Ambisonics is a technique used to record, modify, and recreate a full-sphere surround sound. Unlike other multichannel audio formats, Ambisonics does not rely on encoding the speaker information, but rather the sound field created by multiple sound sources using spherical harmonics. This sound field, referred to as B-format, is then rendered (or decoded) into the listener’s speaker system or headphones.

Fig. 1 illustrates a first-order Ambisonics approximation of the B-format.

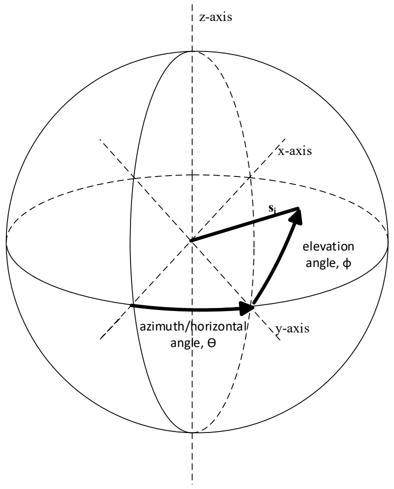


Fig. 1

A simple ambisonic encoder takes the sound signal S, azimuth angle Θ, and elevation angle φ to create a 3-dimensional sound field with four components. These components are often labeled W for the sound pressure, X for the front-minus-back sound pressure gradient, Y for the left-minus-right sound pressure gradient, and Z for the up-minus-down sound pressure gradient. The W component is often referred to as the zeroth-order function and corresponds to the omnidirectional microphone, while X, Y, and Z are the first-order components that represent figure-of-eight capsules that are oriented along the x, y, and z axis, respectively (see Fig. 1).

The components W, X, Y, and Z are computed as follows:

$$W = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_i \left\lfloor \frac{1}{\sqrt{2}} \right\rfloor, \qquad X = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_i cos\varphi_i \, cos\Theta_i, \qquad Y = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_i sin\varphi_i \, cos\Theta_i, \qquad Z = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_i sin\,\Theta_i$$

where i = 1, 2, 3, …(n-1), n.

However, the resolution of first-order Ambisonics is low. In practice, higher-level Ambisonics are needed to increase the quality of the spherical surround sound, where the resulting signal is referred to as high-order Ambisonics. Fig. 2 illustrates, zeroth-, first-, second-, and third-order Ambisonics.

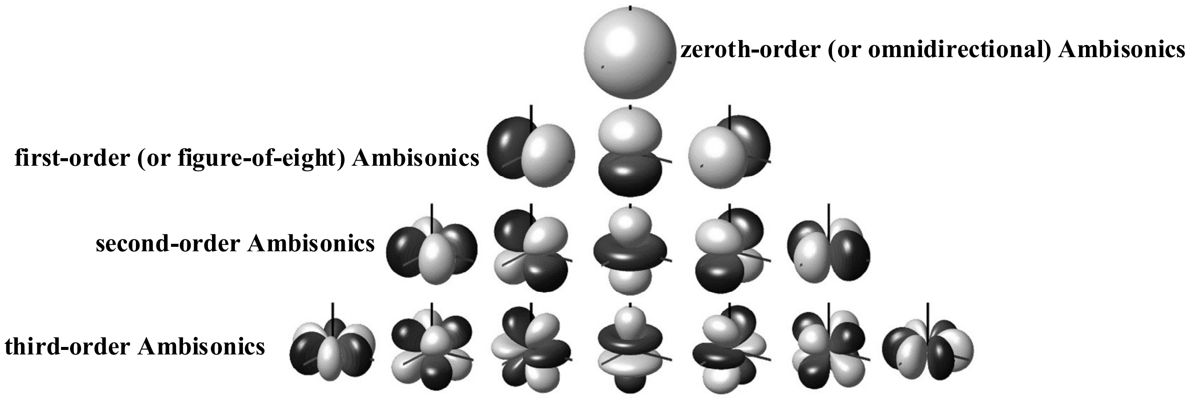

zeroth-order (or omnidirectional) Ambisonics
first-order (or figure-of-eight) Ambisonics
second-order Ambisonics
third-order Ambisonics

Fig. 2

Fig. 2 shows how the signal components increase in complexity with high-order Ambisonics. In addition, Fig. 2 demonstrates that to increase the order of the Ambisonics to an m-th order, there need to be $(m + 1)^2$ signal components, which makes encoding and rendering challenging.

The Ambisonic rendering process reconstructs the encoded sound field at the origin of the spherical space. The center of this spherical space is referred to as the “sweet spot.” The radius of the sweet spot increases with an increase in the number of ambisonic orders. In VR and AR environments, this often means that rendering is done at the left and the right headphone speakers; this is known as binaurally-rendered Ambisonics. To render the sound field as faithful as possible, developers often use a set of head-related transfer functions (HRTFs). Performing convolutions on signals from each loudspeaker with the set of HRFTs provides the listener with a faithful reproduction of the sound source. The use of HRTFs and an increase in the order of high-order Ambisonics broaden the size of the sweet spot, thus, rendering the encoded sound in a more-faithful manner.

Even though the use of HRTFs and high-order Ambisonics broadens the sweet spot, current binaurally-rendered Ambisonics files suffer from error produced by a small sweet spot and the ears being outside that sweet spot for many frequencies.

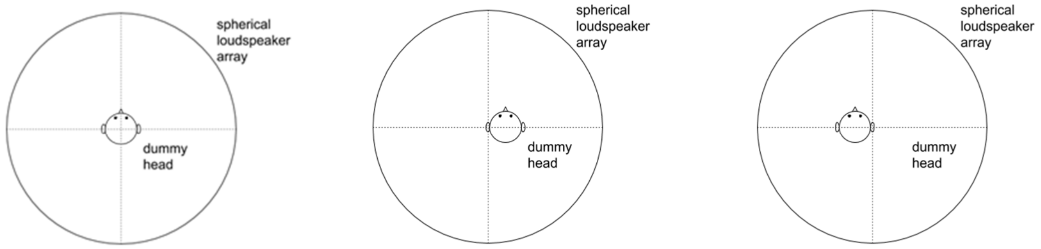
Description:

The use of a conventional head-related transfer function (HRTF), which is processed with truncated spherical harmonics (a common practice in Ambisonics), increases errors with increase in frequency scale. A new method for rendering spatial audio ambisonically that handles each ear independently shifts the paradigm.

In binaurally-rendered Ambisonics, it is advantageous to place the emphasis on the ears and not merely the head. The VR and AR experiences are enhanced by the higher quality rendered sound on each ear. Given that the location of the ears is symmetrical regarding the head, by employing symmetry, the implementation of this rendering method is achieved with marginally added computations than the conventional ambisonic rendering, while significantly increasing the quality and the precision of the rendered spherical sound.

A set of ear-related transfer functions (ERTFs) are generated. The ambisonic encoding of a spatial sound source for both ears is done independently producing two ambisonic signals. Unlike the conventional set of HRTFs, the set of ERTFs permits the binaurally-rendering of each ear's ambisonic signal. Furthermore, the exploitation of symmetry effectively and efficiently enables the encoding and rendering of the signals with marginally added cost, while producing great dividends.

The generation of the set of ERTFs is achieved by simply modifying the well-established technique used to generate HRTFs. The conventional technique of generating HRTFs generally consists of placing a binaural dummy-head inside a spherical array of loudspeakers and centering the head so that the center of the sphere is at the center of the head between the ears. In the HRTF-centric setup, an impulse is generated from each loudspeaker and the response is recorded at microphones placed in both ears. Fig. 3 illustrates how the HRTF-centric setup is modified to generate the set of ERTFs.



a) Conventional HRTF Setup b) Novel Left ERTF Setup c) Novel Right ERTF Setup

Fig. 3

Fig. 3a demonstrates the conventional head ambisonic encoding, whereas Fig. 3b and 3c demonstrate the novel per-ear ambisonic encoding. To more effectively encode spatial sound sources, the ERTF technique keeps track of each ear's position and encodes sources relative to each ear position, instead of the center head position.

Fig. 4 helps demonstrate how conventional ambisonic encoding is modified to compute the spherical harmonic coefficients in the novel ERTFs.

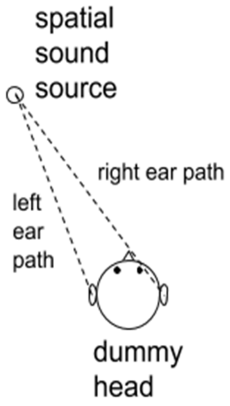


Fig. 4

Fig. 4 shows a spatial sound source with two independent paths to each ear. The method of conventional ambisonic rendering is modified to accomplish per-ear ambisonic rendering. Instead of convolving the signals using the left or right-ear from the HRTFs, one can convolve the per-ear signals with the ERTFs.

The ERTF technique is appealing because of the use of symmetry. Regarding the ears, there is symmetry around the forward (or y) axis of the head. Furthermore, the ears lie directly on the forward (or y) and up (or z) axis origins and only deviate symmetrically along the side-to-side (or x) axis (refer to Fig. 1). This allows for efficient encoding and rendering.

Mathematically, one can show that:

$$EL_W = ER_W, \qquad EL_X = -ER_X, \qquad EL_Y = ER_Y, \qquad EL_Z = ER_Z$$

where *EL* is the first-order spherical harmonic ERTFs of the left ear, and where *ER* is the first-order spherical harmonic ERTFs of the right ear.

For arbitrary orders, negative degrees are inverted between the ERTFs of the left and right ear. For encoding, the non-negative degrees are identical for both ears, therefore non-negative degrees only need to be encoded once. However, two sets of negative degrees are computed since they differ between the ears.

Therefore:

$$XL_W = XR_W, \qquad XL_X \neq XR_X, \qquad XL_Y = XR_Y, \qquad XL_Z = XR_Z$$

where *XL* is the B-format left-ear-encoded sources, and where *XR* is the B-format right-ear-encoded sources.

For rendering, one set of ERTFs for the non-negative degree convolutions is needed. However, the sum of two sets of negative degrees is convolved with one set of ERTFs. Mathematically, one can show that:

$$L = EL * XL$$

$$R = ER * XR$$

where * stands for the convolution operation.

The first-order spherical harmonic ERTFs can be expanded as:

$$L = EL_W * XL_W + EL_X * XL_X + EL_Y * XL_Y + EL_Z * XL_Z$$

$$R = ER_W * XR_W + ER_X * XR_X + ER_Y * XR_Y + ER_Z * XR_Z$$

Utilizing the symmetry, the first-order spherical harmonic ERTFs can be represented as:

$$L = EL_W * XL_W + EL_X * XL_X + EL_Y * XL_Y + EL_Z * XL_Z$$

$$R = ER_W * XL_W - EL_X * XR_X + EL_Y * XL_Y + EL_Z * XL_Z$$

By simplifying it is shown that:

$$J = EL_W * XL_W + EL_Y * XL_Y + EL_Z * XL_Z$$

$$K = EL_X * (XL_X - XR_X)$$

$$L = J + XR_X$$

$$R = J - XL_X$$

To appreciate an arbitrary order of degrees, negative degrees and non-negative degrees are grouped as:

$$J = \sum \left(EL_p * XL_p\right)$$

$$K = \sum [EL_n * (XL_n - XR_n)]$$

$$L = J + \sum (XR_n)$$

$$R = J - \sum (XL_n)$$

where the n-subscript denotes the negative degrees, and where the p-subscript denotes the non-negative degrees.

The mathematical proof shows that the additional cost in rendering symmetric ear-related Ambisonics is the addition (or subtraction) related to the non-negative degrees. The paradigm has shifted; instead of using HRTFs and an ever-increasing order in high-order Ambisonics to broaden the conventional sweet spot, ERTFs essentially double and re-align the sweet spot with marginally added cost.