Parametric analysis of ambisonic audio

Contributions to methods, applications and data generation

Author: Andrés Pérez López

TESI DOCTORAL UPF / year 2020

THESIS SUPERVISORS
Dr. Adan Garriga Torres
Dr. Emilia Gómez Gutiérrez
Department of Information and Communication Technologies

To my cat

Contents

List of Figures

List of Tables

Chapter 1

Scientific Background

1.1 Conventions

1.1.1 Reference system

In what follows, we will make use of a right-handed coordinate system, where the positive x-axis points towards the *front*, the positive y-axis points towards the *left*, and the positive z-axis points towards the *zenith* (North Pole).

Any position in the unit sphere may be described in spherical coordinates by two angles: the *inclination* angle ϑ, which accounts for the aperture with respect to the z-axis, and the *azimuth* angle φ, which represents the counter-clockwise angle with respect to the x-axis from the top-view. The value ranges are $0 \leq \vartheta \leq \pi$ for the inclination, and $0 \leq \varphi \leq 2\pi$ for the azimuth. The spherical coordinate system used in this thesis is depicted in Figure 1.1.

Table 1.1 shows the spherical coordinate values for some reference points on the unit sphere. Notice that the poles ($\vartheta = 0, \pi$) are a special case for the spherical coordinate system – in that case, the azimuth angle is not defined.

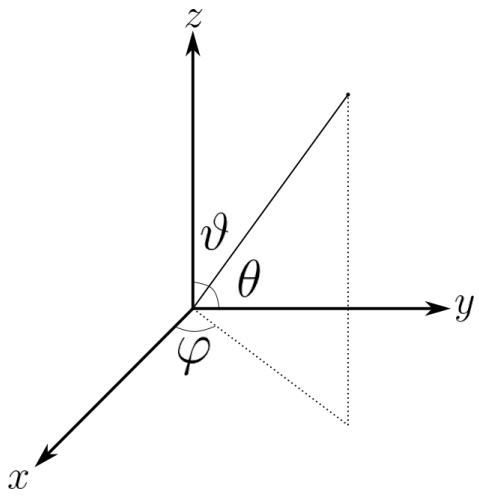

Figure 1.1: Spherical coordinate system used.

The transformation between spherical and cartesian coordinate systems is given by the following relationship:

$$\begin{aligned} x &= \cos\varphi \sin\vartheta \\ y &= \sin\varphi \sin\vartheta \\ z &= \cos\vartheta \end{aligned} \tag{1.1}$$

Table 1.1: Cartesian and spherical representation of characteristic points along the unit sphere.

Position	Cartesian	ϑ	φ
front	$[1, 0, 0]$	$\pi/2$	0
back	$[-1, 0, 0]$	$\pi/2$	π
left	$[0, 1, 0]$	$\pi/2$	$\pi/2$
right	$[0, -1, 0]$	$\pi/2$	$-\pi/2$
zenith	$[0, 0, 1]$	0	*
nadir	$[0, 0, -1]$	π	*

The *elevation* angle θ provides an alternative way of describing the relationship with respect to the z-axis. θ is defined as the aperture with respect to the xy-plane, with positive values towards the positive z-axis. The relationship between elevation and inclination angles is:

$$\theta = \pi/2 - \vartheta \tag{1.2}$$

For the sake of compactness, a point in the unit sphere will be often represented by $\Omega = (\vartheta, \varphi)$.

Given the periodic nature of the azimuth angle, the descriptive statistic operations applied to φ will refer to the 2π*-periodic* version or the operator; this situation does not affect the inclination/elevation coordinate.

1.1.2 Nomenclature

Throughout the Thesis, we refer to time-domain signals with lowercase, e.g. $x(t)$, with t as the time index.

Time-domain signals transformed by the Short-Time Fourier Transform (STFT) are represented with uppercase, e.g. $X(k, n)$, where $k \in [0, K-1]$ is the frequency bin index, and $n \in [0, N-1]$ the time frame index.

Multichannel signals are in general denoted by a subscript variable index, usually with the letter m; for example, $x_m(t)$ or $X_m(k, n)$. Signals with an integer subscript index, such as $x_0(t)$, represent a specific channel of the corresponding multichannel signal.

In the context of ambisonic, subscripts and superscripts are used in signal names with a specific meaning; check Section 1.3 for a detailed explanation.

Vector notation is represented with boldface characters, e.g. $\boldsymbol{X}(k, n)$. When used, the way to construct the vectors will be specified.

1.2 Spherical Harmonics

1.2.1 Definition

Spherical harmonics are continuous functions defined on the sphere surface. Due to their mathematical properties, any continuously differentiable spherical function can be decomposed as a combination of spherical harmonics, in what is known as the *Spherical Harmonics Expansion* [Jarrett et al., 2017].

Many different spherical harmonic definitions exist in the literature, with minor variations among them. In the following, we will use the real-valued, fully normalized spherical harmonics as defined by [Zotter and Frank, 2019]:

$$Y_n^m(\varphi, \vartheta) = N_n^{|m|} P_n^{|m|} \cos(\vartheta) \Phi_m(\varphi), \tag{1.3}$$

where the *normalization factor* N_n^m is:

$$N_n^m = (-1)^m \sqrt{\frac{2n+1}{2} \frac{(n-m)!}{(n+m)!}} \tag{1.4}$$

the *Legendre polynomials* P_n^m are defined as:

$$P_{n+1}^m = \begin{cases} \frac{2n+1}{n-m+1} x P_n^m, & \text{for } n = m, \\ \frac{2n+1}{n-m+1} x P_n^m - \frac{n+m}{n-m+1} P_{n-1}^m & \text{else}, \end{cases} \tag{1.5}$$

with $P_n^n = \frac{(-1)^n (2n)!}{2^n n!} \sqrt{1 - x^{2^n}}$ and the initial term $P_0^0 = 1$, and Φ_m is the azimuthal part of the spherical harmonics:

$$\Phi_m(\varphi) = \frac{1}{\sqrt{2\pi}} \begin{cases} \sqrt{2} \sin(|m|\varphi), & \text{for } m < 0, \\ 1, & \text{for } m = 0, \\ \sqrt{2} \cos(m\varphi), & \text{for } m > 0. \end{cases} \tag{1.6}$$

One of the properties of the spherical harmonics is orthonormality on the sphere surface:

$$\int_{\mathbb{S}^2} Y_n^m(\varphi,\vartheta) Y_{n'}^{m'}(\varphi,\vartheta)\,\mathrm{d}cos\vartheta\,\mathrm{d}\varphi = \delta_{nn'}\delta_{mm'}, \tag{1.7}$$

where δ_{xy} represents the Kronecker delta operator:

$$\delta_{xy} = \begin{cases} 1, & \text{if } x = y, \\ 0, & \text{else.} \end{cases} \tag{1.8}$$

The spherical harmonics depend on the *order* $n \geq 0$ and the *degree* m, $|m| \leq n$ for each value of n. In practice, the maximum order N, $n \leq N$ determines the spatial resolution of the sound field expansion.

Through the spherical harmonic expansion, any sound field may be represented with a limited spatial resolution by the finite combination of all spherical harmonics up to order N. For a given order n, the number of spherical harmonic functions is $2n+1$. With the accumulation of all orders up to N, the total number of spherical harmonics is given by $M = (N+1)^2$. Figure 1.2 depicts all spherical harmonics from orders 0 to 3.

1.2.2 Spherical array processing

Let us consider a sound field captured with a spherical microphone array, which contains Q capsules distributed around a spherical surface of radius R at the positions $\Omega_q, 1 \leq q \leq Q$. The captured frequency-domain signals $X_q(k)$ can be represented as the spherical harmonic domain signals $X_n^m(k)$ through the spherical harmonic transform of order n and degree m [Moreau et al., 2006]:

$$X_n^m(k) = \sum_{q=1}^{Q} X_q(k) Y_n^m(\Omega_q) \Gamma_n(kR), \tag{1.9}$$

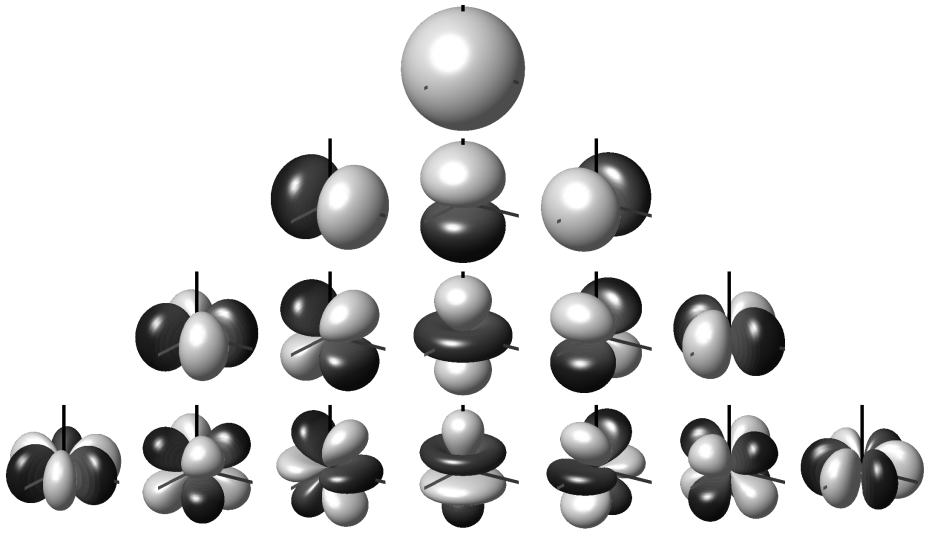

Figure 1.2: Spherical harmonics up to order $N = 3$. The rows correspond to the spherical harmonics of a given order n, and the columns span all possible degree values.

where the term $\Gamma_n(kR)$ models the radial transfer function, and depends on a number of factors, being the *sphere configuration* one of the most prominent considerations. Sphere configuration, in its basic form, refers to the physical properties of the baffle where the capsules are mounted, and it can be either *open* or *rigid*. While open configuration is the simplest solution, it might present numerical problems in the form of zeros in its frequency response. Conversely, a rigid baffle interferes with the sound field and might create undesired interferences, but it improves the numerical condition from the open case. Fig. 1.3 shows the simulated magnitude response of $\Gamma_n(kR)$ for a spherical array considering both configurations. The reader is referred to [Moreau et al., 2006] and [Rafaely, 2004] for a deeper insight into the topic of spherical microphone array design.

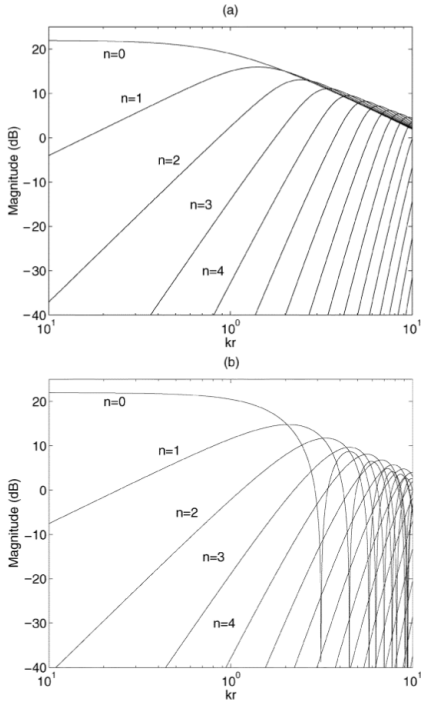

Figure 1.3: Magnitude of $\Gamma_n(kR)$ for different ambisonic orders, in the case of (a) rigid sphere, and (b) open sphere configurations. Adapted from [Rafaely, 2004].

By using the model from Eq. 1.9, the maximum spherical harmonic order N that can be retrieved with negligible spatial aliasing depends on the number of microphone capsules [Moreau et al., 2006]:

$$N \geq (Q+1)^2. \tag{1.10}$$

Furthermore, the sphere radius R has also an effect on the operational bandwidth of the microphone. According to [Moreau et al., 2006], the maximum aliasing-free operational frequency of a spherical microphone array is given by:

$$f_{max} = \frac{c}{2R\gamma}, \tag{1.11}$$

with c being the sound speed, and γ the maximum aperture angle between two capsules. It is important to notice the existence of a practical minimum frequency of the spherical microphone array, given by the low magnitude in low frequencies of high ambisonic order components, as shown in Fig. 1.3.

1.3 Ambisonics

1.3.1 Ambisonics Theory

Ambisonics is a spatial sound recording and playback technology initially developed during the 1970s [Gerzon, 1973], and further expanded into its modern formulation around the 2000s [Daniel, 2000]. Ambisonics is based on the idea of decomposing a sound field into its spherical harmonic representation.

Originally, the decomposition was limited to first-order spherical harmonics, as the so-called *First Order Ambisonics* (FOA); mainly because of practical limitations. The technique was later formalized for arbitrary spherical harmonic orders, known as *Higher Order Ambisonics* (HOA). In general, with the term *ambisonics* we will be referring to the latter definition.

Ambisonic encoding

Let us consider a sound field composed of a point sound source S located in far-field at the angular position Ω_s. The sound pressure at the coordinate origin P can be expressed in terms of the spherical harmonic expansion of order N as:

$$P = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\Omega_s) S \tag{1.12}$$

The ordered set of values of all spherical harmonics up to order N, evaluated at the source position, is known as the *ambisonic coefficients*:

$$Y_n^m(\Omega_s) = [Y_0^0(\Omega_s), Y_1^{-1}(\Omega_s), \ldots, Y_N^N(\Omega_s)] \tag{1.13}$$

Furthermore, the process of multiplying the signal S by the ambisonic coefficients is known in the literature as the *ambisonic encoding*. The resulting signal vector is usually referred to as the *ambisonic* (or *B-Format*) signal S_n^m:

$$S_n^m = Y_n^m(\Omega_s) S \tag{1.14}$$

Note that, because of the superposition principle, a sound field composed of several different point sources can be broken down to the addition of the individual contributions.

Although the term *B-Format* was initially introduced as an alternative name for first-order ambisonic signals [Daniel, 2000], it is nowadays common to use it as a synonim of ambisonic signals, without any order restriction. We will use the latter acception in what follows.

Historically, the name *B-Format* was used as an opposite of *A-Format*, which describes the signals recorded by a tetrahedral microphone array [Gerzon, 1975]. The tetrahedron is the simplest and most common form of spherical microphone arrays (indistinctly referred to as ambisonic microphones) with uniform

capsule distribution. Again, the term *A-Format* is also currently employed for referring to the signals recorded by any spherical microphone array, regardless of the number or arrangement of capsules.

Likewise, the process of signal conversion from the spatial domain (microphone capsules) to the spherical harmonic domain (ambisonic signals), as in Eq. 1.9, is known as *A-B conversion*. A number of different approaches have been developed for this process, and the interested reader is referred to [Moreau et al., 2006] for more information.

In practice, there are two alternative ways to generate ambisonic signals. The first one is the *synthesis*, based on the direct application of ambisonics encoding (Eq. 1.12) to a monophonic signal. The second one is the *recording* with a spherical microphone array, followed by the aforementioned domain conversion.

Ambisonic Decoding

Conversely, the sound field reconstruction is performed by the *ambisonic decoding* operation. This process is equivalent to weight-and-sum beamforming in the spherical harmonic domain, and it is sometimes also referred to as the *virtual microphone* technique [Zotter and Frank, 2019].

Let us consider a loudspeaker located at the angular position Ω_p. In accordance with Eq. 1.12, the signal feed P is *decoded* from the ambisonic signal as:

$$P = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\Omega_s) S Y_n^m(\Omega_\ell) \alpha_n \tag{1.15}$$

where α_n is a weighting factor which accounts for the beam directivity. There are several standard weightings used for different purposes; their values are shown in Table 1.2, and the first-order directive patterns are plotted in Figure 1.4.

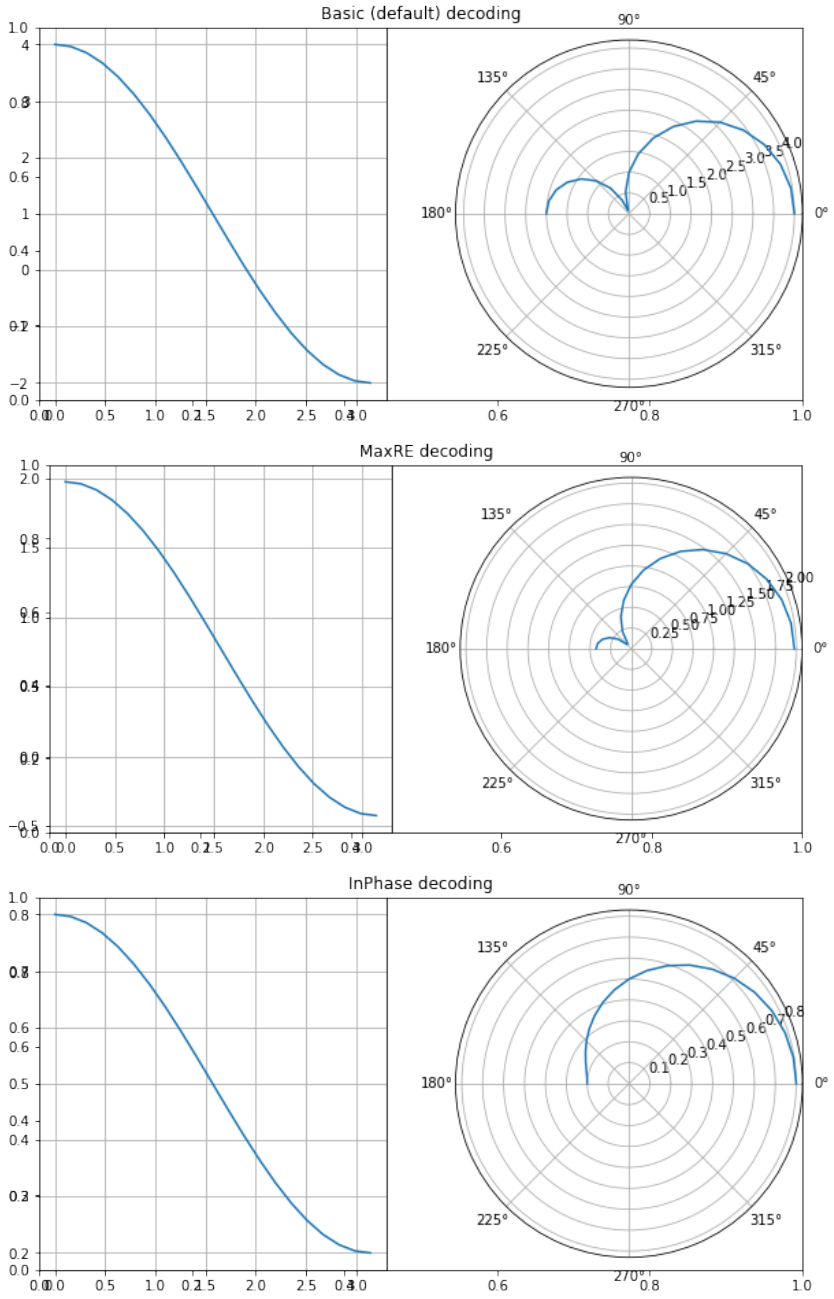

Figure 1.4: Directive patterns of first-order ambisonic decoding.

Table 1.2: Ambisonic decoding: standard values of $alpha_n$ weightings. Adapted from [Daniel, 2000].

Decoding	N	n			
		0	1	2	3
basic	0	1			
	1	1	1		
	2	1	1	1	
	3	1	1	1	1
max-rE	0	0.577			
	1	0.775	0.4		
	2	0.861	0.612	0.305	
	3	0.906	0.732	0.501	0.246
in-phase	0	0.333			
	1	0.5	0.1		
	2	0.6	0.2	0.029	
	3	0.667	0.286	0.071	0.008

The decoding equation 1.15 can be written in matrix form as:

$$P = S_n^m Y_n^m(\Omega_p)^T \alpha_n \tag{1.16}$$

where the superscript T represents the matrix transposition. This equation can be extended to the usual case of decoding to a loudspeaker array, comprised of L loudspeakers located at the positions $\Omega_L = [\Omega_{p_1}, \ldots, \Omega_{p_L}]$. In such case, the loudspeaker feed vector P_L can be written as:

$$P_L = S_n^m D, \tag{1.17}$$

where

$$D = \text{diag}(\alpha_n)[Y_n^m(\Omega_{p_1})^T, \ldots, Y_n^m(\Omega_{p_L})^T] \tag{1.18}$$

is a $M \times L$ matrix known as the *decoding matrix*, and diag(α_n) is a diagonal matrix of size M containing the values of α_n along the main diagonal. Although the matrix D is frequency-independent and depends solely on the loudspeaker array geometry, in practical scenarios it is usual to include frequency-dependent weightings, $\alpha_n(k)$, to improve the broadband sound field reconstruction [Daniel, 2000].

Furthermore, sound field reconstruction with Eq. 1.17 is only possible when the loudspeakers are evenly located on the 3D space; in other words, the speaker layout must take the form of one of the five *Platonic solids*: tetrahedron, cube, octahedron, dodecahedron or icosahedron. Provided that this condition is usually difficult to fulfil in real scenarios, there are several methods which allow ambisonic decoding for such *irregular* layouts. One of the most commonly used is the AllRAD method [Zotter and Frank, 2012]. AllRAD proposes a two step decoding: first, the ambisonic signal is decoded to a nearly-uniform layout of virtual speakers. Then, the signals of the virtual speakers are further distributed into the real speakers by the *Vector-Based Amplitude Panning (VBAP)* method [Pulkki, 1997].

1.3.2 Practical considerations

Due to historical and practical reasons, there are two aspects that must be taking into account when working with ambisonic signals: *channel normalization* and *channel ordering*. In the following, the term *channels* will be used as a synonym for spherical harmonics, as they are usually referred to in sound engineering contexts[1].

[1]In fact, ambisonic signals are inherently multichannel, even though each channel corresponds to a spherical harmonic, and not to a loudspeaker feed as in traditional *channel-based* audio.

Channel normalization

Let us consider the spherical harmonics $Y_n^m(\Omega)$ as defined in Eq. 1.3. Due to the orthonormal property showed in Eq. 1.7, they follow the *fully 3d normalized* or *N3D* channel normalization convention.

Alternatively, the *Schmidt 3d semi-normalized* or *SN3D* [Daniel, 2000] convention is also of widespread usage. The conversion between *N3D* and *SN3D* is driven by the following expression:

$$Y_n^m(\Omega)^{(\text{N3D})} = \sqrt{2n+1} Y_n^m(\Omega)^{(\text{SN3D})} \tag{1.19}$$

MaxN is another existing convention. It defines all spherical harmonics as having a maximum absolute value of 1:

$$\max_{\Omega} |Y_n^m(\Omega)^{(\text{MaxN})}| = 1, \forall (n,m) \tag{1.20}$$

Finally, the *Furse-Malham* (or *FuMa*) normalization only differs from *Max-N* in the scaling of the zero-th order component:

$$Y_n^m(\Omega)^{(\text{FuMa})} = \begin{cases} 1/\sqrt{2}, & \text{if } n = 0, \\ Y_n^m(\Omega)^{(\text{MaxN})}, & \text{else.} \end{cases} \tag{1.21}$$

Each of the normalization schemes has its own particularities. For instance, *N3D* is the most mathematically straightforward, and spherical harmonics defined in that way can be directly used for both encoding and decoding (as in Eqs 1.12 and Eq. 1.15) – however, from a sound engineer point of view, other normalization schemes with maximum values below the unity might be preferred, such as *SN3D*. Besides this, *FuMa* has been historically the default normalization [Gerzon, 1985], while the more modern *N3D* and *SN3D* were popularized after J. Daniel's work [Daniel, 2000].

As a summary, Figure 1.5 displays the different normalization schemes. The reader is referred to [Carpentier, 2017] for an extensive review on the topic.

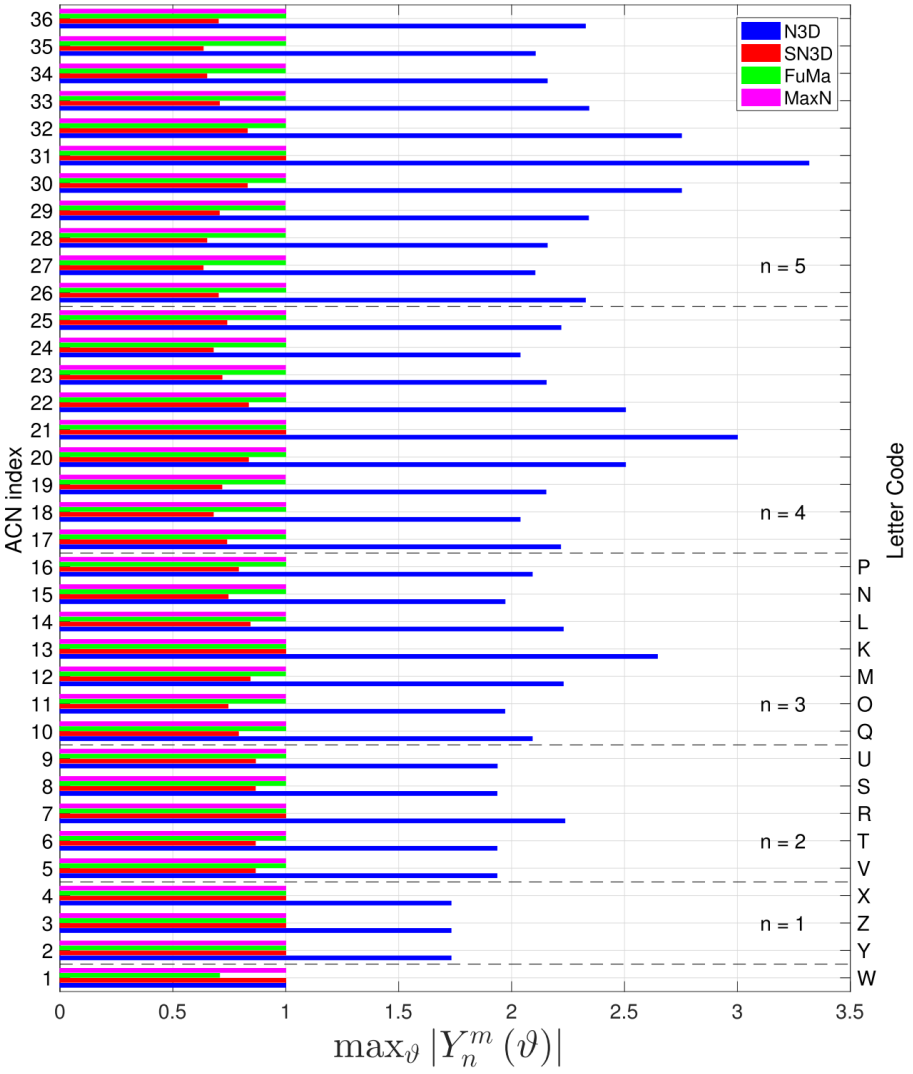

Figure 1.5: Maximum value of each ambisonic channel up to order 5, for all different normalization schemes. Image from [Carpentier, 2017].

Channel ordering

Channel ordering refers to the manner in which spherical harmonics, inherently organized in the 2D space by dimensions n and m, are sorted into a one-dimensional vector.

The *ACN* (from *Ambisonic Channel Number*) scheme follows from the mathematical description given in Eq. 1.13. The spherical harmonics are first ordered by ascending order n and, inside each order, by ascending degree m. The index of a given channel $i \in [0 \dots M-1]$ can be thus obtained by the following relationship:

$$i = n^2 + n + m \tag{1.22}$$

Historically, first-order ambisonic audio has followed what it might be called *traditional B-Format* channel ordering [Gerzon, 1985]. By this scheme, the four channels of a FOA signal S_n^m are referred to by the axis where the corresponding spherical harmonic steers, plus the name W for the zeroth order component:

$$S_n^m(\Omega)^{\text{(FuMa CO)}} = [W, X, Y, Z] \tag{1.23}$$

where:

$$\begin{aligned} W &= S_0^0(\Omega) \\ X &= S_1^1(\Omega) \\ Y &= S_1^{-1}(\Omega) \\ Z &= S_1^0(\Omega) \end{aligned} \tag{1.24}$$

This nomenclature was extended to second and third order, and is currently known as the *Furse-Malham* or *FuMa* channel ordering. The channel names use all english alphabet letters from K to Z in third order and, although there would be enough letters to go up to fourth order, the unconvenience of the system was clear [Malham, 2003]. Figure 1.5 shows the equivalence between *FuMa* (*"letter code"*) and *ACN* channel names.

In practice, there exist two main combinations of channel normalization and ordering schemes:

- The *classical* approach, usually limited to first-order ambisonics, which uses *FuMa* normalization and channel ordering[2].
- The *modern* approach, inspired by the *ambix* file format [cite ambix], with *SN3D* normalization and *ACN* channel ordering.

Anyhow, the *classical B-Format* channel naming and ordering is still widely used when referring to first-order ambisonics.

1.4 Parametric Spatial Audio Analysis

Trough parametric analysis, sound fields may be described in terms of a small amount of sound sources and associate parameters. Such representation might reduce to a great extent the complexity of processing methods [Jarrett et al., 2017].

One of the most successful sound field parametric models is DirAC [Pulkki, 2007], which was originally conceived as a method for impulse response processing and spatial sound reproduction [Merimaa and Pulkki, 2005].

DirAC (acronym for *Directional Audio Coding*) is a perceptually motivated time-frequency (TF) domain method, based on the assumption that any sound field may be reproduced with high perceptual quality by considering two parameters: the sound field diffuseness and the most prominent sound *Direction-of-Arrival* (DOA) [Pulkki et al., 2018].

[2]In general, it may be expected that *early* ambisonic material follow these conventions without any explicit mention to them.

Let us consider a *SN3D*-normalized first-order ambisonic signal in time-frequency domain, $S_n^m(k,n)$. For the sake of clarity, we will use in this section *FuMa* channel notation and ordering (Eq. 1.23):

$$S_n^m(k,n) = [W(k,n), X(k,n), Y(k,n), Z(k,n)] \tag{1.25}$$

Given this representation, we can express the *pressure* $P(k,n)$ of the sound field as:

$$P(k,n) = W(k,n) \tag{1.26}$$

as well as the sound *pressure-gradient* (or *velocity*) $\boldsymbol{U}(k,n)$ as:

$$\boldsymbol{U}(k,n) = -\frac{1}{\rho_0 c}[X(k,n), Y(k,n), Z(k,n)], \tag{1.27}$$

where ρ_0 is the mean medium density, and c is the sound speed.

The *active intensity* $\boldsymbol{I}(k,n)$, defined as the amount of transmitted acoustic energy, can be expressed in terms of sound pressure and velocity [Fahy and Salmon, 1990]:

$$\begin{aligned} \boldsymbol{I}(k,n) &= \Re\{P^*(k,n)\boldsymbol{U}(k,n)\} \\ &= -\frac{1}{\rho_0 c}\Re\{W^*(k,n)[X(k,n), Y(k,n), Z(k,n)]\}, \end{aligned} \tag{1.28}$$

where * represents the complex conjugate operator.

An estimate of the instantaneous DOA $\Omega(k,n)$ can be extracted from the intensity vector, interpreting each of its time-frequency bins as a point in the cartesian space. Effectively, the sound propagation direction is the opposite to the observed arrival direction.

$$\Omega(k,n) = \angle(-\boldsymbol{I}(k,n)), \tag{1.29}$$

with $\angle$ representing the spherical angle operator of a cartesian vector. The result of this computation must be understood as the direction of the net energy flow, which in the case of a single plane-wave will correspond to the source position.

Another useful parameter is the *energy density* $E(k,n)$ [Stanzial et al., 1996]:

$$\begin{aligned} E(k,n) &= \frac{1}{2\rho_0 c^2}|P(k,n)|^2 + \frac{1}{2}\|\boldsymbol{U}(k,n)\|^2 \\ &= \frac{1}{2\rho_0 c^2}\Big(|W(k,n)|^2 + \|[X(k,n),Y(k,n),Z(k,n)]\|^2\Big). \end{aligned} \tag{1.30}$$

Finally, the *diffuseness* $\Psi(k,n)$ can be computed from the sound intensity and energy density [Merimaa and Pulkki, 2005]:

$$\begin{aligned} \Psi(k,n) &= 1 - \frac{\|\langle \boldsymbol{I}(k,n)\rangle\|}{c\langle E(k,n)\rangle} \\ &= 1 - 2\frac{\|\langle \Re\{W^*(k,n)[X(k,n),Y(k,n),Z(k,n)]\}\rangle\|}{\langle |W(k,n)|^2 + \|[X(k,n),Y(k,n),Z(k,n)]\|^2\rangle}, \end{aligned} \tag{1.31}$$

where the symbols $\langle\cdot\rangle$ represent the expectation operator, which is usually implemented as time-domain averaging.

Even though Eq. 1.31 (known as *DirAC's diffuseness*) is one of the most common ambisonic diffuseness estimators, several alternative formulations exist. Other diffuseness estimation procedures include the *coefficient of variation method* [Ahonen and Pulkki, 2009] and the more recent *COMEDIE* estimator [Epain and Jin, 2016]. In any case, in what follows, the term *diffuseness* and the symbol Ψ will refer by default to Eq. 1.31.

As a mathematical convenience, we will define the *B-Format coherence* as the complement of the diffuseness:

$$\Delta(k,n) = 1 - \Psi(k,n) \tag{1.32}$$

In conclusion, Figure 1.6 plots the spectrograms of the DOA $\Omega(k,n)$ and diffuseness $\Psi(k,n)$ of a FOA recording, which consists of a sound source located at the front, plus a moderate amount of reverberation and background noise.

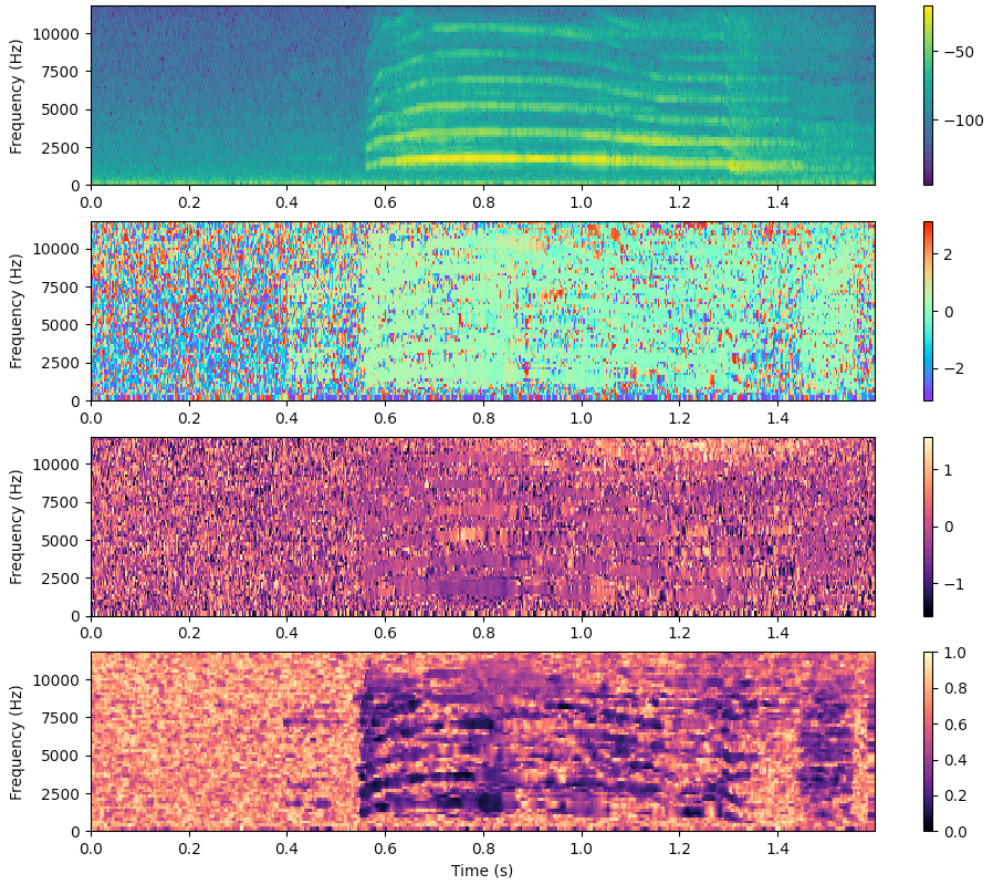

Figure 1.6: Parametric time-frequency spatial audio analysis of a first order ambisonic recording. From top to bottom: 1.) Magnitude spectrogram of the omnidirectional channel. 2.) and 3.) Azimuth and elevation of the estimated instantaneous narrowband DOAs $\Omega(k, n)$. 4.) Instantaneous narrowband diffuseness $\Psi(k, n)$.

1.5 Spatial Coherence Analysis

In the context of microphone array signal processing, diffuseness is commonly estimated through the *Magnitude Squared Coherence* (MSC) [Elko, 2001] between two frequency-domain signals S_1 and S_2, as a function of the *wavenumber* k and the capsule distance r:

$$\text{MSC}_{12}(kr) = \frac{|\,\langle S_1(kr) S_2(kr)^* \rangle\,|^2}{\langle |S_1(kr)|^2 \rangle \, \langle |S_2(kr)|^2 \rangle}, \tag{1.33}$$

where the $\langle \cdot \rangle$ operator represents the temporal expected value, and * defines the complex conjugate operator. In the case of spherical isotropic noise fields, Eq. (1.33) can be expressed in terms of microphone directivity patterns $T(\phi, \theta, kr)$ as [Elko, 2001]:

$$\begin{aligned}\text{MSC}_{12}(kr) &= \frac{|N_{12}(kr)|^2}{|D_{12}(kr)|^2} \\ &= \frac{|\int_0^{\pi}\int_0^{2\pi} T_1(\phi,\theta,kr) T_2^*(\phi,\theta,kr) e^{-jkr cos\theta} sin\theta d\theta d\phi|^2}{|\sqrt{\int_0^{\pi}\int_0^{2\pi} |T_1(\phi,\theta,kr)|^2 sin\theta d\theta d\phi}\sqrt{\int_0^{\pi}\int_0^{2\pi} |T_2(\phi,\theta,kr)|^2 sin\theta d\theta d\phi}|^2}.\end{aligned} \tag{1.34}$$

Moreover, the general expression of the directivity of a first-order differential microphone is given by the following relationship:

$$T_i(\Omega_i) = \alpha_i + (1 - \alpha_i)\cos\Omega_i, \tag{1.35}$$

where $i \in [1, 2]$ is the microphone index, Ω_i is the angle between wave incidence and microphone orientation axis, and $\alpha_i \in [0, 1]$ is the directivity parameter of the microphone i, which ranges from bidirectional ($\alpha_i = 0$) to omnidirectional ($\alpha_i = 1$).

For first-order differential microphones, there is a closed-form expression for the numerator and denominator of Eq. (1.34):

$$
\begin{aligned}
N_{12}(kr) &= \frac{\alpha_1\alpha_2 sin(kr)}{kr} \\
&+ \frac{(1-\alpha_2)(1-\alpha_2)(x_1x_2+y_1y_2)}{(kr)^3}(sin(kr)-krcos(kr)) \\
&+ \frac{z_1z_2}{kr^3}[((kr)^2sin(kr)+2krcos(kr))(1-\alpha_1)(1-\alpha_2)+2sin(kr)(1-\alpha_1)(1-\alpha_2)] \\
&+ \frac{z_1}{(kr)^3}[j(kr)^2\alpha_2 cos(kr)(\alpha_1-1)+jkr\alpha_2 sin(kr)(1+\alpha_1)] \\
&+ \frac{z_2}{(kr)^3}[j(kr)^2\alpha_1 cos(kr)(\alpha_2-1)+jkr\alpha_1 sin(kr)(1+\alpha_2)], \\
D_{12}(kr) &= \frac{\sqrt{3\alpha_1^2+(1-\alpha_1)^2}\sqrt{3\alpha_2^2+(1-\alpha_2)^2}}{3},
\end{aligned}
\tag{1.36}
$$

where x_i, y_i and z_i are the cartesian coordinates of the wave incidence angle $\Omega_i = (\varphi_i, \vartheta_i)$.

1.6 Reverberation

In the context of room acoustics, reverberation refers to *"the energy of a sound source that reaches a listener indirectly, by reflecting from surfaces within the surrounding space occupied by the sound source and the listener"* [Begault and Trejo, 2000]. Conversely, in anechoic or free-field conditions, where reverberation is not present, only the direct path of the sound source exists. Assuming linearity and time-invariance, room reverberation can be fully characterised by its impulse response (IR).

Reverberation models often consider two differentiated parts of the reverberant tail, based on both physical and perceptual characteristics: the *early reflections* and the *late reverberation*. Early reflections, as the name suggests, refers to the individual sound

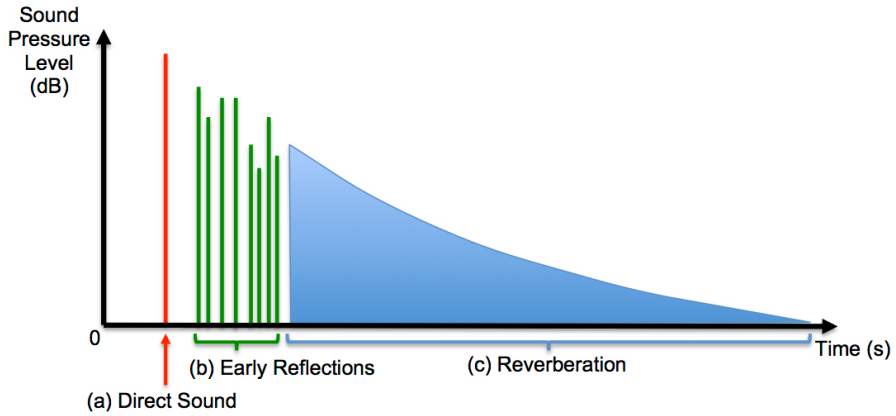

Figure 1.7: Room impulse response model, from [Murphy et al., 2017].

paths arriving to the listener after a few reflections on the room surfaces, which cause some degree of attenuation. Early reflections typically arrive with a time difference between 1 and 80 ms after the direct path [Begault and Trejo, 2000]. The term late reverberation encompasses all sound paths arriving to the listener after many reflections. Since the temporal density of such reflections increases with time, late reverberation is often modelled in statistical terms. An schematic representation of a room impulse response (RIR) is shown in Figure 1.7.

By following this model, a RIR $h(t)$ can be described as a sequential combination of responses:

$$h(t) = h_D(t) + h_R(t), \tag{1.37}$$

where $h_D(t)$ and $h_R(t)$ represent the *direct* (direct path plus early reflections) and *reverberant* (late reverberation) components of the RIR. respectively.

The room impulse response is a function of both the source and the receiver locations. Different levels, delays and directions of direct path and early reflections can obtained from measurements in the same room. However, it is generally assumed that the late reverberation is fixed for a given room, regardless of source/receiver positions.

Room reverberation plays an important role in psychoacoustics. While early reflections are usually perceived together with the direct path as a single auditory event, due to the *precedence effect* [Haas, 1972], late reverberation has often an influence on the received signal. In the specific case of speech, late reverberation is associated with a loss of intelligibility [Braun, 2018]. In the context of spatial perception, it has been shown that early reflections help the localization and externalization of sources [Rudrich and Frank, 2019], while the late reverberation is associated with a spaciousness perception of the room [Begault and Trejo, 2000].

There are a number of measurable parameters which help to characterise room acoustics. Perhaps one of the most widespread is the *reverberation time* T_{60} [Kuttruff, 2016]. It represents the time required for the reverberant sound field power to decay by 60 dB. Reverberation time can be accurately computed from the room geometry [Sabine, 1927] or from the IR [Schroeder, 1965].

In the latter case, the T_{60} value is usually estimated from the *Energy Decay Curve* (EDC), which is defined as:

$$\mathrm{EDC}(t) = 10 \log_{10} \sum_{t'=t}^{\infty} h^2(t), \tag{1.38}$$

where $h(t)$ represents the room impulse response. The values are normalized such that the maximum peak of the curve corresponds to 0 dB.

The EDC is usually modelled as a straight line in logarithmic scale. Therefore, the T_{60} estimation is performed by estimating the slope of a straight line between two reference levels on the EDC

Table 1.3: Reverberation time computation: usual reference levels

	EDT	T_{10}	T_{20}	T_{30}
$L_{max}(dB)$	0	-5	-5	-5
$L_{min}(dB)$	-10	-15	-25	-35

time series. Some of the most used reference levels receive specific names: *Early Decay Time* (EDT), T_{60}, and reverberation times T_{10}, T_{20} and T_{30}. Table1.3 shows their correspondent reference levels, where the maximum energy peak is normalized to 0 dB. An schematic representation of the reference levels is depicted in Figure 1.8.

An alternative parameter is the *decay rate* α_{60}, which is related to reverberation time T_{60} as:

$$\alpha_{60} = \frac{3\ln(10)}{T_{60}}(\text{dB/s}). \tag{1.39}$$

The decay rate is thus the slope of the EDC curve, in logarithmic scale, expressed in dB per second.

To conclude, it is important to notice that reverberation time is frequency-dependent. Accordingly, it is usual to report it for octave or third-octave bands, or alternatively to provide its value at a specific frequency.

The *Direct to Reverberant Ratio* (DRR) is another relevant acoustic parameter. DDR represents the ratio between direct and reverberant parts of the RIR, as defined in Eq. 1.37:

$$DRR = 10\log_{10}\frac{\sum_{t=1}^{L_D} h_D^2(t)}{\sum_{t=1}^{L_R} h_R^2(t)}, \tag{1.40}$$

with L_D and L_R as the length of the direct $h_D(t)$ and reverberant $h_R(t)$ filters, respectively. At a psychoacoustic level, the direct to

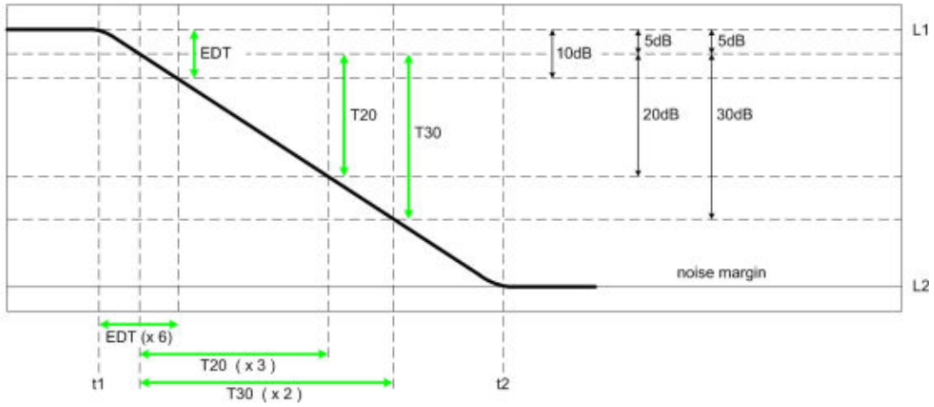

Figure 1.8: Room impulse response model, adapted from [AV_INFO, 1995].

reverberant ratio is one of the main cues for distance perception [Begault and Trejo, 2000].

Since the direct path and early reflections (but not the late reverberation) depend on the relative position between source and receiver , the filter $h_D(t)$ and therefore the DRR are as well location-dependent. For a given room, the source-receiver distance that produces a DRR of 0 dB is known as the *critical distance*.

1.7 Signal Models

Let us consider a sound source represented by the signal $s(t)$, located in a given acoustic enclosure characterised by its room impulse response $h(t)$. The resulting reverberant signal $x(t)$ can be therefore described as the *convolutive mixture* of the source and the RIR:

$$x(t) = s(t) * h(t). \tag{1.41}$$

When dealing with multichannel room impulse responses, as it is the case in ambisonics, the multichannel reverberant signal $x_m(t)$ is obtained by the convolutive mixture of each RIR channel independently:

$$x_m(t) = s(t) * h_m(t). \tag{1.42}$$

The time domain convolution operation, under certain assumptions, is equivalent to the multiplication in frequency domain. By doing so, Eq. 1.41 can be expressed as:

$$X(k,n) = S(k,n)H(k,n). \tag{1.43}$$

Eq. 1.43, also known as the *Multiplicative Transfer Function (MTF) model* is only valid when the length of the filter $h(t)$ is smaller than the length of the analysis window used in the STFT.

On the contrary, when the filter $h(t)$ spans across several analysis windows, the resulting model is referred to as the *Convolutive Transfer Function (CTF) model*:

$$X(k,n) = \sum_{l=0}^{L_h-1} H(k,l)S(k,n-l), \tag{1.44}$$

where L_h is the length of the filter $H(k,n)$ in time frames.

Bibliography

[Ahonen and Pulkki, 2009] Ahonen, J. and Pulkki, V. (2009). Diffuseness estimation using temporal variation of intensity vectors. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 285–288. IEEE.

[AV_INFO, 1995] AV_INFO (1995). Reverberation time. http://www.bnoack.com/. Accessed on June 26th, 2020.

[Begault and Trejo, 2000] Begault, D. R. and Trejo, L. J. (2000). 3-d sound for virtual reality and multimedia.

[Braun, 2018] Braun, S. (2018). *Speech dereverberation in noisy environments using time-frequency domain signal models*. PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

[Carpentier, 2017] Carpentier, T. (2017). Ambisonic spatial blur. In *142nd Audio Engineering Societiy Convention*. AES.

[Daniel, 2000] Daniel, J. (2000). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris VI.

[Elko, 2001] Elko, G. W. (2001). Spatial coherence functions for differential microphones in isotropic noise fields. In *Microphone Arrays*, pages 61–85. Springer, New York.

[Epain and Jin, 2016] Epain, N. and Jin, C. T. (2016). Spherical harmonic signal covariance and sound field diffuseness. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1796–1807.

[Fahy and Salmon, 1990] Fahy, F. J. and Salmon, V. (1990). *Sound intensity*. Acoustical Society of America.

[Gerzon, 1973] Gerzon, M. (1973). Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10.

[Gerzon, 1975] Gerzon, M. A. (1975). The design of precisely coincident microphone arrays for stereo and surround sound. In *Audio Engineering Society Convention 50*. Audio Engineering Society.

[Gerzon, 1985] Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871.

[Haas, 1972] Haas, H. (1972). The influence of a single echo on the audibility of speech. *Journal of the Audio Engineering Society*, 20(2):146–159.

[Jarrett et al., 2017] Jarrett, D. P., Habets, E. A., and Naylor, P. A. (2017). *Theory and applications of spherical microphone array processing*, volume 9. Springer.

[Kuttruff, 2016] Kuttruff, H. (2016). *Room acoustics*. Crc Press.

[Malham, 2003] Malham, D. (2003). 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*.

[Merimaa and Pulkki, 2005] Merimaa, J. and Pulkki, V. (2005). Spatial impulse response rendering i: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127.

[Moreau et al., 2006] Moreau, S., Daniel, J., and Bertet, S. (2006). 3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone. In *120th Convention of the AES*, pages 20–23.

[Murphy et al., 2017] Murphy, D., Shelley, S., Foteinou, A., Brereton, J., and Daffern, H. (2017). Acoustic heritage and audio creativity: the creative application of sound in the representation, understanding and experience of past environments.

[Pulkki, 1997] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, 45(6):456–466.

[Pulkki, 2007] Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516.

[Pulkki et al., 2018] Pulkki, V., Delikaris-Manias, S., and Politis, A. (2018). *Parametric time-frequency domain spatial audio*. Wiley Online Library.

[Rafaely, 2004] Rafaely, B. (2004). Analysis and design of spherical microphone arrays. *IEEE Transactions on speech and audio processing*, 13(1):135–143.

[Rudrich and Frank, 2019] Rudrich, D. and Frank, M. (2019). Improving externalization in ambisonic binaural decoding. In *DAGA 2019 Fortschritte der Akustik*.

[Sabine, 1927] Sabine, W. C. (1927). *Collected papers on acoustics*. Harvard University Press Cambridge, MA.

[Schroeder, 1965] Schroeder, M. R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(6):1187–1188.

[Stanzial et al., 1996] Stanzial, D., Prodi, N., and Schiffrer, G. (1996). Reactive acoustic intensity for general fields and energy polarization. *The Journal of the Acoustical Society of America*, 99(4):1868–1876.

[Zotter and Frank, 2012] Zotter, F. and Frank, M. (2012). All-round ambisonic panning and decoding. *Journal of the audio engineering society*, 60(10):807–820.

[Zotter and Frank, 2019] Zotter, F. and Frank, M. (2019). *Ambisonics*. Springer.