SOUND SPATIALIZATION BY HIGHER ORDER AMBISONICS: ENCODING AND DECODING A SOUND SCENE IN PRACTICE FROM A THEORETICAL POINT OF VIEW...

Rozenn Nicol

Orange Labs
TECH/OPERA/TPS
2 Avenue Pierre Marzin, 22307 Lannion, France
rozenn.nicol@orange-ftgroup.com

ABSTRACT

An overview of HOA technology is presented. First, HOA defines a format of spatial audio which has many attractive properties, such as scalability and flexibility. Besides, this format is independent of the encoding (i.e. microphone signals) and decoding (i.e. loudspeaker signals) formats. Second, HOA provides tools to record, or create, and render a spatial sound scene. These tools, which rely on a specific encoding and decoding of spatial information, will be analysed and discussed from a both theoretical and practical point of view. Third, the final issue is the assessment of the virtual sound scene that is (re)created by HOA. The toolkit of available methodologies and criteria is examined.

1. INTRODUCTION

Ambisonics and its generalization Higher Order Ambisonics (H.O.A.) are one promising technology of sound spatialization, allowing one to record or create a spatial sound scene. Since the early work of Gerzon [1], a hughe amount of both hardware and software developments has provided microphone arrays and plugins, which today place HOA as a real and recognised tool of spatial audio engineering. The flexibility of HOA rendering, which is compliant with various equipements, including conventional multichannel setup (eg 5.1), loudspeaker array of headphones, is a strong advantage. The objective of this paper is not to present particularly new information, but rather to gather all what is known or still questioned about HOA audio tools, including all the steps (and the available choices involved within) from the recording of sound sources to their playback. However, because HOA is also a spatial audio format which is able to represent and model a sound scene, this issue will be first discussed.

2. A MODEL FOR REPRESENTING A SPATIAL SOUND SCENE

2.1. Spherical harmonics expansion

HOA technology is based on the expansion of an acoustic wave on the eigenfunctions of the acoustic wave equation within spherical coordinates (r: radius, φ: azimuth angle, θ: elevation angle) [2]. These eigenfunctions are defined by spherical Bessel functions $j_m(kr)$[1] et $n_m(kr)$[2] and/or spherical Hankel functions $h_m^+(kr)$[3] and $h_m^-(kr)$[4], in combination with spherical harmonics $Y_{mn}^{\sigma}(\varphi,\theta)$. The latter describe the angular variation of the acoustic wave, whereas the former account for the radius dependencies.

The whole space is divided into two subspaces: one subspace Ω_1 where all the sound sources are gathered and one subspace Ω_2 where no acoustic source is present and which defines therefore the listening area. Given the spherical geometry of the problem, the space is organized on the basis of concentric spheres centered on the origin of the coordinate system, which does not limit the validity of the following. Thus the subspace Ω_2 is defined as the area within two spheres of radius R_1 et R_2 so that: $R_1 < |\vec{r}| = r < R_2$, where r is the radius associated to the listening point $\vec{r}$. The radius R_1 et R_2 are chosen in order to discard any sound source from Ω_2. The subspace Ω_1 is the remaining area (i.e. inner space of radius R_1 sphere and outer space of radius R_2 sphere). As pointed out in [3], this space dichotomy strongly resembles to that of the Kirchhoff Integral in the theory of Wave Field Synthesis (WFS).

Under these assumptions, the acoustic pressure $p(\vec{r},\omega)$ at any point $\vec{r}$ located inside Ω_2 is expressed as a weighted sum of the eigenfunctions:

$$\begin{aligned} p(\vec{r},\omega) = &\sum_{m=0}^{+\infty} i^m h_m^-(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} A_{mn}^{\sigma}(\omega) Y_{mn}^{\sigma}(\varphi,\theta) \\ &+ \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma}(\omega) Y_{mn}^{\sigma}(\varphi,\theta) \end{aligned} \quad (1)$$

The spherical harmonics are given by:

$$\begin{aligned} Y_{mn}^{\sigma}(\varphi,\theta) = &\sqrt{(2m+1)\epsilon_n \frac{(m-n)!}{(m+n)!}} P_{mn}(\sin\theta) \\ &\times \begin{cases} \cos(n\varphi) \; si \; \sigma = +1 \\ \sin(n\varphi) \; si \; \sigma = -1 \end{cases} \end{aligned} \quad (2)$$

where ϵ_n equals 1 if $n = 0$ and 2 if $n > 0$. The functions

[1] Spherical Bessel functions of the first kind.

[2] Spherical Bessel functions of the second kind or Neumann functions.

[3] Spherical Hankel functions of the first kind: wave travelling along decreasing r.

[4] Spherical Hankel functions of the second kind: wave travelling along increasing r.

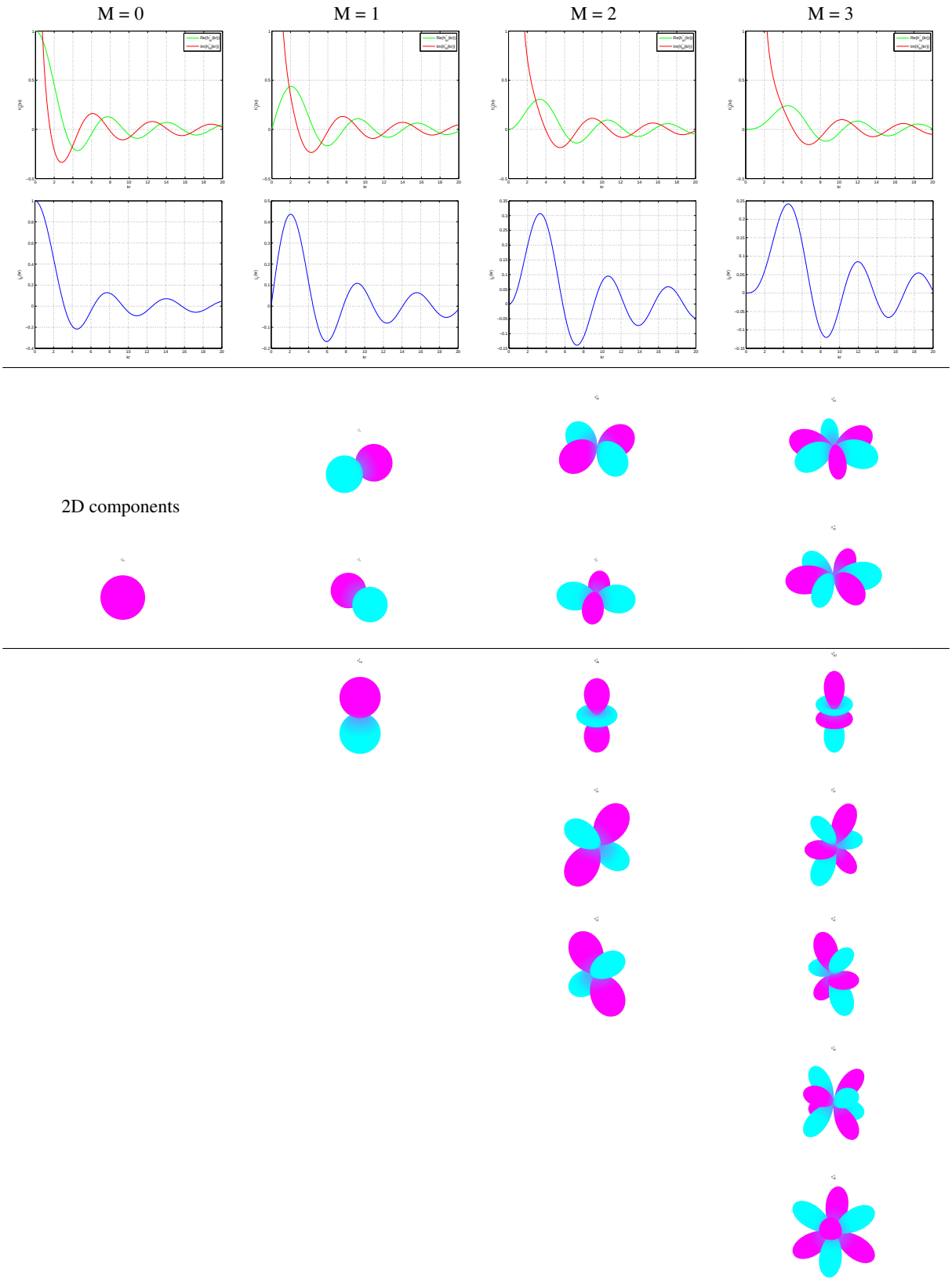


Figure 1: Illustration of the Bessel and Hankel spherical functions, and the spherical harmonics.

$P_{mn}(\sin\theta)$ are the associated Legendre functions defined by:

$$P_{mn}(\sin\theta) = \frac{d^n P_m(\sin\theta)}{d(\sin\theta)^n} \tag{3}$$

where P_m is the Legendre polynomial of the first kind of degree m.

It should be born in mind that the spherical harmonics Y^σ_{mn} form an orthonormal basis in terms of the scalar product over the sphere of radius $r = 1$, i.e.:

$$\begin{aligned}\frac{1}{4\pi}\int_{\varphi=0}^{2\pi}\int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} Y^\sigma_{mn}(\varphi,\theta)Y^{\sigma'}_{m'n'}(\varphi,\theta)\cos\theta\, d\theta d\varphi \\ = \delta_{mm'}\delta_{nn'}\delta_{\sigma\sigma'}\end{aligned} \tag{4}$$

The coefficients A^σ_{mn} and B^σ_{mn} of the spherical harmonics expansion (cf. Equ. 1) are therefore obtained by computing the scalar product of the acoustic pressure with the spherical harmonics and by applying their property of orthonormality (cf. Equ. 4). The knowledge of both the acoustic pressure $p(R,\varphi,\theta,\omega)$ and velocity $v_n(R,\varphi,\theta,\omega)$ over a sphere[5] of arbitrary radius R is required, as shown in [3]. If U^σ_{mn} and V^σ_{mn} represent the product scalar of respectively the pressure and the velocity with the spherical harmonics:

$$\begin{aligned}U^\sigma_{mn}(\omega) =& \frac{1}{4\pi R^2}\int_{\varphi=0}^{2\pi}\int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} p(R,\varphi,\theta,\omega) \\ &\times Y^\sigma_{mn}(\varphi,\theta)\cos\theta\, d\theta d\varphi \\ V^\sigma_{mn}(\omega) =& \frac{1}{4\pi R^2}\int_{\varphi=0}^{2\pi}\int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} v_n(R,\varphi,\theta,\omega) \\ &\times Y^\sigma_{mn}(\varphi,\theta)\cos\theta\, d\theta d\varphi\end{aligned} \tag{5}$$

the coefficients A^σ_{mn} and B^σ_{mn} are derived as:

$$\begin{aligned}\frac{A^\sigma_{mn}(\omega)}{i^{-m}} &= \frac{{j'}_m(kR)U^\sigma_{mn}(\omega) - icRj^-_m(kR)V^\sigma_{mn}(\omega)}{{j'}_m(kR)h^-_m(kR) - j_m(kR){h'}^-_m(kR)} \\ \frac{B^\sigma_{mn}(\omega)}{i^{-m}} &= \frac{{h'}_m(kR)U^\sigma_{mn}(\omega) - icRh^-_m(kR)V^\sigma_{mn}(\omega)}{j_m(kR){h'}^-_m(kR) - {j'}_m(kR)h^-_m(kR)}\end{aligned} \tag{6}$$

The coefficients A^σ_{mn} et B^σ_{mn} define the HOA representation of the acoustic wave. In the previous equations, these signals are in the frequency domain since they originate from the pressure and velocity expressed as in the frequency domain (cf. Equ. 5). Alternatively, signals in the time domain may be considered.

2.2. The HOA "wavelets"

The term "wavelet" refers here to the reconstruction of the primary wave as a sum of elementary wavelets by WFS. The general form of the HOA wavelets is defined (cf. Equ. 1) by $j_m(kr)Y^\sigma_{mn}(\varphi,\theta)$ or $h^-_m(kr)Y^\sigma_{mn}(\varphi,\theta)$. Each elementary wavelet is characterized by specific spatial features which depend on the order m of the associated HOA component. The angular variation (as a function of the direction described by the angles φ and θ) becomes faster and faster as the order m increases. As for the variation along the radius r, two types of wavelet are distinguished [5]:

- waves travelling outwards from the origin (radial dependency described by h^-_m),
- waves travelling inwards to the origin (radial dependency described by j^-_m).

The former represent the contribution of sound sources located inside the sphere of radius R_1, whereas the latter correspond to sound sources outside the sphere of radius R_2. Thus the two sets of HOA components, i.e. the A^σ_{mn} set and the B^σ_{mn} one, discriminate inner sources ($r < R_1$) from outer sources ($r > R_2$), in a way very similar to WFS [4].

Consequently, if no inner sound sources is present inside the sphere of radius R_1, all the coefficients A^σ_{mn} are null and Equ. 1 turns out to be:

$$p(\vec{r},\omega) = \sum_{m=0}^{+\infty} i^m j_m(kr)\sum_{n=0}^{m}\sum_{\sigma=\pm1} B^\sigma_{mn}(\omega)Y^\sigma_{mn}(\varphi,\theta) \tag{7}$$

The sphere of radius R_1 is then useless, so that the listening area Ω_2 includes all the inside of the sphere of radius R_2. The well-known "B format" proposed by Gerzon [1] is a particular case of Equ. 7, consisting in truncating the expansion up to the order m=1. Therefore only the first four components ($B^1_{00} \equiv W$, $B^1_{11} \equiv X$, $B^{-1}_{11} \equiv Y$, $B^1_{10} \equiv Z$) are considered. The HOA format generalizes this representation by adding the orders greater than 1 up to a maximal order $M > 1$, which leads to $(M+1)^2$ components B^σ_{mn} (m=0, 1, ..., M; n=0, 1, ..., m; $\sigma = \pm1$) for a full 3D encoding and $(2M+1)$ components for a 2D restriction[6] (horizontal plane). Equ. 7 is an exact representation of the acoustic wave under the condition that all the terms up to the order $M = +\infty$ are kept, which means an infinite number of components. As soon as the expansion is truncated to a finite order, the reconstruction of the soundfield becomes erroneous and the validity of the representation is limited. The validity should be analyzed in terms of the product *kr* where k is the wave number and r is the radius. Usually the reconstruction error is assumed to be negligable as soon as:

$$kr \leq M \tag{8}$$

which means that for a given frequency f_0 the reconstruction is erroneous inside a circle of radius $r_{max} = \frac{M}{k_0}$ (i.e. the **sweet spot**) or, alternatively, at a fixed radius r_0 the reconstruction is valid for frequencies up to $f_{max} = \frac{M}{r_0}$. Indeed the expansion involved in Equ. 7 can be interpreted as an asymptotic development around the origin (kr=0).

2.3. HOA as a format for Spatial Audio

The B^σ_{mn} signals or HOA components define a new format to represent and encode a spatial sound scene. This format is really universal since it is able to represent any soundfield, including plane or spherical waves. For instance, for a plane wave of magnitude O_p and originating from the direction (φ_p,θ_p), the HOA

[5] In other words, the primary wave to be reproduced is described by the acoustic pressure and velocity recorded along a close surface, which is strongly similar to the representation of the spatial information by WFS [4]. What's more the choice of the radius R of the sphere where the soundfield is recorded is arbitrar in both HOA and WFS, at least theoretically. However it will be seen that, in practice, this choice is constrainted by the Bessel and Hankel spherical functions.

[6] In this case, it is more relevant to apply a Fourier-Bessel expansion [2]. However, it should be noticed that, for an appropriate restriction to the horizontal plane, any wave of non-null elevation should be discarded from the B^σ_{mn} components, which is not straightforward.

components are given by:

$$B^{\sigma}_{mn}(\omega) = \frac{O_p}{4\pi} Y^{\sigma}_{mn}(\varphi_p, \theta_p) \tag{9}$$

In the same way, a spherical wave of magnitude O_s and emitted by a sound source located at $\vec{r}_s(r_s, \varphi_s, \theta_s)$ with $|\vec{r}_s| > R_2$ is represented by:

$$B^{\sigma}_{mn}(\omega) = \frac{O_s}{4\pi} i^{-(m+1)} \frac{h^{-}_m(kr_s)}{k} Y^{\sigma}_{mn}(\varphi_s, \theta_s) \tag{10}$$

The HOA format has several advantages:

- It is intrinsically **discrete**, i.e. based on a discrete series of components, which should, at least theoretically, avoid spatial sampling[7],
- It is a **scalable** format which relies on a **hierarchic** representation, which means that even the lower order components are in themselves sufficient to provide a full description of the sound scene. The contribution of the higher order components is only to enhance the spatial accuracy and results mainly in increasing the validity limit $(kr)_{max} = M$, which means enlarging the sweet spot or increasing the frequency bandwidth $[0 - f_{max}]$. Scalability allows one to discard the higher components whenever required by the limited capacity of storage, transmission, or rendering.
- The representation is straightforwardly (and even intuitively) **readable** in terms of spatial structure, firstly since radial and angular variations are separated. Secondly each HOA component is identified to a spatial scanning of sound components, with an increasing resolution as a function of the order m. Historically the HOA format is linked to empirical representations of spatial audio which were developped by sound engineers. Indeed the B format is somehow a generalization of the M-S (*Mitte-Seite*) stereophonic format which combines a omnidirectional and a bidirectional microphones, in order to record separately the omnidirectional and the left-right informations.
- The expansion of spherical harmonics can be interpreted as a dual transform between spatial coordinates and spatial frequency. The B^{σ}_{mn} components thus define a **spatial spectrum** which may be interpreted in terms of spatial frequency or variation.

3. HOA ENCODING

Spatial encoding is the first step which consists in recording the sound scene and aims at delivering the B^{σ}_{mn} signals. In the case of a plane wave (cf. Equ. 9), the B^{σ}_{mn} signals take the form of the spherical harmonics which can be identified to the directivity function of microphones (cf. Fig. 1). Thus the first component (B^1_{00}) corresponds to an omnidirectional microphone, whereas the 3 components of order 1 (B^1_{11}, B^{-1}_{11}, B^1_{10}) are the output of bidirectional (i.e. "figure of eight") microphones aligned with the x, y and z-axis. Therefore up to the first order, the B^{σ}_{mn} signals could be recorded by a set of conventional microphones. For higher orders, the directivity becomes more and more complex and the equivalent microphones do not exist. Anyway, even though the desired directivities would have been available, this solution is not feasable because it requires to put all the microphones at the same point.

[7]However it will be shown that, in practice, the microphone arrays used for HOA recording imply spatial sampling.

3.1. Spherical array of microphones

To get the B^{σ}_{mn} signals, the alternative [6] is to record the sound wave (i.e. the acoustic pressure $p(r_M, \varphi, \theta, \omega)$) over a sphere of radius r_M and to compute its scalar product with the spherical harmonics according to Equ. 5:

$$U^{\sigma}_{mn}(\omega) = \frac{1}{4\pi r_M^2} \int_{\varphi=0}^{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} p(r_M, \varphi, \theta, \omega) \times Y^{\sigma}_{mn}(\varphi, \theta) \cos\theta \, d\theta d\varphi \tag{11}$$

The B^{σ}_{mn} signals are deduced from the U^{σ}_{mn} signals by a formula similar to Equ. 6, but which takes into account the fact that the expansion contains only the B^{σ}_{mn} components under the assumption that the A^{σ}_{mn} are null (cf. Equ. 7):

$$B^{\sigma}_{mn}(\omega) = EQ(kr_M) U^{\sigma}_{mn}(\omega) \tag{12}$$

where the equalization term $EQ(kr_M)$ is defined as:

$$EQ(kr_M) = \frac{1}{i^m j_m(kr_M)} \tag{13}$$

From above, it should be realized that two different audio formats are considered:

- the **recording** format (i.e. the microphone outputs),
- the HOA format (i.e. B^{σ}_{mn} signals).

These two formats are fully independent, which means that the HOA format is not determined by the recording setup. This is a fundamental property of HOA technology.

3.2. Cardioid microphones

This solution of HOA recording raises a first problem. Whenever the spherical Bessel function is null, the equalization term (cf. Equ. 13) is null and the B^{σ}_{mn} signals can not be computed. In order to overcome this problem, it is shown that if not only the pressure, but also the velocity are recorded, the denominator of the equalization term contains both the spherical Bessel function and its first derivative, which are never null simultaneously [5]. In practice, this result is obtained by replacing the pressure microphones ($p(r_M, \varphi, \theta, \omega)$) by cardioid[8] sensors. The output of such microphones $c(r_M, \varphi, \theta)$ is a linear sum of the acoustic pressure and its gradient (i.e. the acoustic velocity, except for a multiplying factor):

$$c(r_M, \varphi, \theta) = p(r_M, \varphi, \theta) - \frac{\vec{\nabla} p(v, \varphi, \theta).\vec{n}}{ik} \tag{14}$$

These signals are developed in spherical harmonics (cf. Equ. 5):

$$C^{\sigma}_{mn}(\omega) = \frac{1}{4\pi r_M^2} \int_{\varphi=0}^{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} c(r_M, \varphi, \theta, \omega) \times Y^{\sigma}_{mn}(\varphi, \theta) \cos\theta \, d\theta d\varphi \tag{15}$$

Thus the B^{σ}_{mn} are deduced from the C^{σ}_{mn} signals by the formula:

$$B^{\sigma}_{mn}(\omega) = EQ(kr_M) C^{\sigma}_{mn}(\omega) \tag{16}$$

where the equalization term is now:

$$EQ(kr_M) = \frac{1}{i^m [j_m(kr_M) + k j'_m(kr_M)]} \tag{17}$$

Another solution consists in putting the microphones around a sphere and the resulting diffraction modifyies the equalization term in order that it is never null [6].

[8]HOA encoding becomes then very similar to WFS encoding.

3.3. Spatial Sampling

The second problem is that in theory the sound wave ($p(r_M, \varphi, \theta)$ and $\vec{\nabla} p(r_M, \varphi, \theta)$) is supposed to be recorded continuously along the sphere, whereas in practice it is recorded by a discrete array of sensors, which implyies spatial sampling. We consider now a microphone array which is composed of N_M transducers. The qth mictrophone's location is defined by $\vec{r}_{M,q}(r_M, \varphi_{M,q}, \theta_{M,q})$. The question to be solved is to determine the optimal positioning of the N_M microphones, under the following constraints:

- to minimize the estimate error of the B^{σ}_{mn} signals,
- to minimize the number N_M of necessary sensors,
- a feasible geometry of the microphone array.

This problem has an exact solution as soon as the spatial spectrum of the sound wave is band-limited, i.e. the B^{σ}_{mn} components are all null for any order greater than a maximal order m_{max}. This is a generalization of the Nyquist-Shannon sampling theorem for functions defined over a sphere. Indeed Driscoll & Healy [7] have shown that it is only required to uniformly sample the azimuth and elevation angles. Under this condition, the B^{σ}_{mn} signals can be accurately interpolate from the N_M microphone signals $c_q(\omega) = c(r_M, \varphi_{M,q}, \theta_{M,q}, \omega)$ [6]. The disadvantage of this solution is that the number of microphones is suboptimal (i.e. $N_M = 4(M+1)^2$ for a recording up to order M).

In order to decrease the number of microphones, an approximate solution (i.e. $\hat{B}^{\sigma}_{mn}$) is preferred. Starting from Equ. 14, the acoustic pressure and its derivative are developed in spherical harmonics (cf. Equ. 7) for each output microphone:

$$\begin{aligned} c_q(\omega) = & \sum_{m=0}^{M} i^m [j_m(kr_M) + kj'_m(kr_M)] \\ & \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B^{\sigma}_{mn}(\omega) Y^{\sigma}_{mn}(\varphi_{M,q}, \theta_{M,q}) \\ & for\ q = 1, ..., N_M \end{aligned} \tag{18}$$

which yields a set of N_M equations with $(M+1)^2$ unknowns which are the B^{σ}_{mn} signals. This problem can be reformulated into a matrix equation:

$$\mathbf{c} = \mathbf{Y}_M \, \mathbf{W}_M \, \mathbf{b} \tag{19}$$

where the vector $\mathbf{c}$ contains the microphone outputs c_q and the vector $\mathbf{b}$, the B^{σ}_{mn} signals. The matrix $\mathbf{Y}_M$ and $\mathbf{W}_M$ are given by:

$$\mathbf{Y}_M = \begin{bmatrix} Y^1_{00}(\varphi_{M,1}, \theta_{M,1}) & Y^1_{10}(\varphi_{M,1}, \theta_{M,1}) & \dots & Y^{-1}_{MM}(\varphi_{M,1}, \theta_{M,1}) \\ Y^1_{00}(\varphi_{M,2}, \theta_{M,2}) & Y^1_{10}(\varphi_{M,2}, \theta_{M,2}) & \dots & Y^{-1}_{MM}(\varphi_{M,2}, \theta_{M,2}) \\ \vdots & \vdots & \vdots & \vdots \\ Y^1_{00}(\varphi_{M,N_M}, \theta_{M,N_M}) & Y^1_{10}(\varphi_{M,N_M}, \theta_{M,N_M}) & \dots & Y^{-1}_{MM}(\varphi_{M,N_M}, \theta_{M,N_M}) \end{bmatrix}$$

$$\mathbf{W}_M = \begin{bmatrix} [j_0(kr_M) + kj'_0(kr_M)] & 0 & \dots & 0 \\ 0 & [j_1(kr_M) + kj'_1(kr_M)] & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & i^M [j_M(kr_M) + kj'_M(kr_M)] \end{bmatrix}$$

The first requirement to solve Equ. 19 is that the number of equations is greater than or equal to the number of unknowns, which means that the minimal number of microphones is $(M+1)^2$ (instead of $4(M+1)^2$ for the exact solution). Generally the solution is obtained by least-square minimization. The B^{σ}_{mn} signals are estimated from the microphone outputs thanks to the Moore-Penrose pseudoinverse of $\mathbf{Y}_M$:

$$\hat{\mathbf{b}} = \mathbf{E}_M \, (\mathbf{Y}^t_M \, \mathbf{Y}_M)^{-1} \, \mathbf{Y}^t_M \, \mathbf{c} \tag{20}$$

where $\mathbf{Y}^t_M$ refers to the transpose conjugate of $\mathbf{Y}_M$. The matrix $\mathbf{E}_M$ is defined by:

$$\mathbf{E}_M = \begin{bmatrix} \frac{1}{[j_0(kr_M) + kj'_0(kr_M)]} & 0 & \dots & 0 \\ 0 & \frac{1}{i[j_1(kr_M) + kj'_1(kr_M)]} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1}{i^M [j_M(kr_M) + kj'_M(kr_M)]} \end{bmatrix}$$

The primary cause of estimate error relies on the sensitivity of the system to any approximation, concerning mainly the microphone outputs (i.e. signal-to-noise ratio or mispositioning of the microphones). The equalization term contained in the matrix $\mathbf{W}_M$ is another factor of instability. Regularization is therefore recommended. The diagonal terms of the matrix $\mathbf{E}_M$ are then replaced by [6]:

$$F_m(kr_M) = \frac{|i^m [j_m(kr_M) + kj'_m(kr_M)]|^2}{|i^m [j_m(kr_M) + kj'_m(kr_M)]|^2 + \lambda^2} \tag{21}$$

where λ is the regularization parameter. The term $F_m(kr_M)$ is a regularization filter.

3.4. Orthonormality condition

In Equ. 20, let us focus on the term:

$$\mathbf{Y}^t_M \, \mathbf{Y}_M \, .$$

If the spatial distribution of the microhones over the sphere satisfyies the orthonormality property of the spherical harmonics (cf. Equ. 4), this term reduces to:

$$\mathbf{Y}^t_M \, \mathbf{Y}_M = \mathbf{1} \tag{22}$$

where **1** is the identity matrix. The $\hat{B}^{\sigma}_{mn}$ signals become :

$$\hat{B}^{\sigma}_{mn}(\omega) = \frac{1}{j_m(kr_M) + kj'_m(kr_M)} \frac{1}{N_M} \times \sum_{q=1}^{N_M} c_q(\omega) Y^{\sigma}_{mn}(\varphi_{M,q}, \theta_{M,q}) \quad (23)$$

This result is the discrete (i.e. sampled) version of Equ. 16.

If the microphone locations are chosen arbitrarly, the orthonormality property is not satisfied in general. The estimate of the $\hat{B}^{\sigma}_{mn}$ signals include then the term $\mathbf{Y}^t_M \, \mathbf{Y}_M$ which originates from the aliasing of spherical harmonics. The matrix ϵ quantifyies this "non-orthonormality" error:

$$\epsilon = \mathbf{1} - \frac{1}{N_M} \mathbf{Y}^t_M \, \mathbf{Y}_M \quad (24)$$

Therefore it should be realized that, in our problem, spatial sampling affects not only the soundfield (i.e. $c(r_M, \varphi, \theta)$), but also the spherical harmonics basis (i.e. $Y^{\sigma}_{mn}(\varphi, \theta)$). That's why, besides all the requirements which were previously presented, the choice of the array geometry must satisfy an additional requirement in order to satisfy the orthonormality property. But this is very difficult in practice. A limited choice of regular polyhedrons provides geometries which satisfy this condition for lower orders. Semi-regular polyhedrons are convenient solutions for order greater than 2 [6]. More generally, for a given order M, the spatial distribution of the microphones is optimized by minimizing the matrix ϵ.

3.5. Conclusion

To summarize, the process of HOA encoding is characterized by three main parameters:

- the **number** of microphones N_M,
- the **location** of each microphone $(\varphi_{M,q}, \theta_{M,q})$ over the sphere,
- the **radius** of the microphone array r_M.

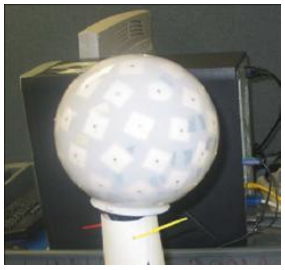
The first question is to choose the maximal order M which is expected. The value of M determines the minimal number of microphones:

$$N_M \geq (1 + M)^2.$$

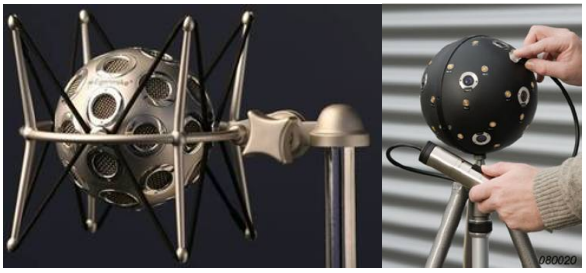
The next step is to find the regular or semi-regular polyhedron which is composed of at least $(M+1)^2$ vertices and which minimizes the non-orthonormality error (cf. Equ. 24) for orders $m \leq M$. The selected polyhedron imposes the number of microphones and their location. The last issue concerns the value of the radius r_M which determines the values of the Bessel functions $j_m(kr_M)$ and $j'_m(kr_M)$. These latter have a double effect [6]:

- On the one hand, in Equ. 18, it can be shown that the Bessel functions act as a low-frequency filter which reduces spatial sampling like an anti-aliasing filter [6]. The cutoff frequency increases with r_M. Therefore decreasing the radius will minimize the spatial aliasing.
- On the other hand, the Bessel functions are also present in the equalization term (Equ. 17). It is observed that a small radius leads to the ill-conditioning of the problem (Equ. 20), especially for low frequencies. Thus, decreasing r_M disadvantaged the accuracy of the estimate of the $\hat{B}^{\sigma}_{mn}$ signals at low frequencies.

(a) Orange Labs prototypes [6]



(b) University of Maryland [8]



(c) EigenMike™(mh-acoustics) (d) Brüel & Kjær™

Figure 2: Example of possible or existing HOA microphones.

As a result, the optimal radius is a difficult compromise between spatial sampling and low-frequency recording.

This section has described all the steps from the microphone outputs to the estimate of the $\hat{B}^{\sigma}_{mn}$ signals, i.e. the HOA components. Recording is performed by a microphone array (cf. Fig. 2). It was explained why cardioid sensors are preferred. The rules to optimize the array geometry as a function of the recording contraints were clarified. Before investigating HOA decoding, it should be mentioned that the Soundfield™microphone is the first example of a HOA recording setup which fully agrees with the previous requirements, namely a spherical array of cardioid microphones which are distributed at the vertices of a regular polyhedron which satisfyies the orthonormality property up to order 1 (i.e. tetrahedron).

4. HOA DECODING

4.1. Overview

The decoding step aims at reconstructing the primary acoustic wave by a loudspeaker setup. The loudspeakers are fed by signals which are appropriately derived from the B^{σ}_{mn} sig-

nals. In the following, an array of N_L emitters is considered. The location of the lth loudspeaker is described by a vector $\vec{r}_{L,l}(r_{L,l}, \varphi_{L,l}, \theta_{L,l})$. If $s_l(\omega)$ refers to the lth loudspeaker input, the secondary wave $\hat{p}$ (i.e. the wave synthesized by the loudspeaker array) is:

$$\hat{p}(\vec{r}, \omega) = \sum_{l=1}^{N_L} s_l(\omega) p_l(\vec{r}, \omega) \tag{25}$$

where p_l is the acoustic pressure induced by the lth loudspeaker at $\vec{r}$. This latter can be developed in spherical harmonics (Equ. 7):

$$p_l(\vec{r}, \omega) = \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} {L_l}^{\sigma}_{mn}(\omega) Y^{\sigma}_{mn}(\varphi, \theta) \tag{26}$$

The ${L_l}^{\sigma}_{mn}$ signals are the HOA components for the acoustic wave radiated by the lth loudspeaker. This wave may be of any kind: spherical, plane or more complex [9]. It is also possible that any loudspeaker has a specific acoustic radiation which differs from the other ones. Theoretically the geometry of the loudspeaker array is arbitrar and is not limited to a sphere or a circle (2D rendering).

The primary wave to be synthesized is described by its HOA components (i.e. the B^{σ}_{mn} signals obtained at the encoding output). To derive the s_l signals to feed the loudspeaker, the spherical harmonic expansion of the primary (Equ. 7) and the synthesized (Equ. 25 & 26) waves are matched (*mode-matching* principle):

$$B^{\sigma}_{mn} = \sum_{l=1}^{N_L} s_l(\omega) {L_l}^{\sigma}_{mn}(\omega) \tag{27}$$

which yields a set of $(M+1)^2$ equations with N_L unknowns which are the loudspeaker input s_l. This problem can be reformulated into a matrix equation:

$$\mathbf{b} = \mathbf{L}\mathbf{s} \tag{28}$$

where the vector $\mathbf{s}$ contains the louspeaker input signals s_l. The matrix $\mathbf{L}$ is composed of the ${L_l}^{\sigma}_{mn}$ components for each loudspeaker:

$$\mathbf{L} = \begin{bmatrix} {L_1}^{1}_{00}(\varphi_{M,1}, \theta_{M,1}) & {L_2}^{1}_{00}(\varphi_{M,2}, \theta_{M,2}) & \dots & {L_{N_L}}^{1}_{00}(\varphi_{M,N_L}, \theta_{M,N_L}) \\ {L_1}^{1}_{10}(\varphi_{M,1}, \theta_{M,1}) & {L_2}^{1}_{10}(\varphi_{M,2}, \theta_{M,2}) & \dots & {L_{N_L}}^{1}_{10}(\varphi_{M,N_L}, \theta_{M,N_L}) \\ \vdots & \vdots & & \vdots \\ {L_1}^{-1}_{MM}(\varphi_{M,1}, \theta_{M,1}) & {L_2}^{-1}_{MM}(\varphi_{M,2}, \theta_{M,2}) & \dots & {L_{N_L}}^{-1}_{MM}(\varphi_{M,N_L}, \theta_{M,N_L}) \end{bmatrix}$$

The solution of the problem depends on the number N_L of loudspeakers in respect with the order M. Three cases should be distinguished [10]:

- $N_L < (M+1)^2$: The problem is overdetermined. No exact solution exists. An approximate solution can be found by least-square minimization.
- $N_L = (M+1)^2$: The matrix $\mathbf{L}$ is square. If its inverse exists, the solution is given by:

$$\mathbf{s} = \mathbf{L}^{-1}\mathbf{b} \tag{29}$$

- $N_L > (M+1)^2$: The problem is underdetermined, which means that it has an infinity of solutions. The solution which minimizes the energy of the loudspeaker signals is obtained by using the pseudoinverse of $\mathbf{L}$:

$$\mathbf{s} = \mathbf{L}^t(\mathbf{L}\mathbf{L}^t)^{-1}\mathbf{b} \tag{30}$$

The optimal number N_L of loudspeakers is a worthwhile issue. The value $N_L = (M+1)^2$ ($N_L = 2M+1$ for a 2D rendering) can be considered as optimal as it minimizes the reconstruction error [10]. However, it is observed that in this case, when a virtual sound source is located in the direction of a loudspeaker, only this latter is switched on, which causes audible artefacts in terms of loudspeaker transparency and homogeneity of the sound rendering [11]. This effect is eliminated as soon as N_L becomes greater than $(M+1)^2$. Nevertheless listening tests confirm that increasing the number of loudspeakers beyond $(M+1)^2$ is detrimental to the audio quality, since the soundfield is likely to become instable, especially when the listener moves his(her) head [12].

4.2. Decoding matrix

The loudspeaker signals are derived from the B^{σ}_{mn} signals thanks to the decoding matrix $\mathbf{D}$:

$$\mathbf{s} = \mathbf{D}\mathbf{b} \tag{31}$$

From Equ. 29 or 30, the decoding matrix is defined by:

$$\mathbf{D} = \mathbf{L}^{-1} \text{ or } \mathbf{L}^t(\mathbf{L}\mathbf{L}^t)^{-1} \tag{32}$$

This matrix performs a kind of transcoding of the B^{σ}_{mn} signals to the loudspeaker space [5]. The HOA format (i.e. the B^{σ}_{mn} signals, cf. Section 2.3) is thus not only independent of the recording format, as previously mentionned, but also independent of the rendering format (i.e. the loudspeaker format or D format [13]).

The decoding matrix is determined by the loudspeaker layout (i.e. number and geometry). In theory it is able to account for any setup. Nevertheless it is recommended to prefer a regular layout, such an uniform sampling of the sphere, which yields a simple and stable matrix. The matrix given by Equ. 32 follows a particular rule of decoding which is called "basic decoding" and which aims at the pure reconstruction of the acoustic wave. Other strategies exist, which include additional constraints while solving the decoding problem (in the case $N_L > (M+1)^2$) [11]. One strategy is to maximize the signals of the loudspeaker which are the closest to the virtual source location. Another one is to minimize the signals of the loudspeakers which are opposite of the virtual source.

4.3. Regular array

Even though the loudspeaker layout can be chosen arbitrarly, which is a remarkable advantage of HOA technology, it is preferred, whenever there is no constraint, to opt for a regular setup. For 3D rendering, it consists in placing the loudspeaker over the surface of a sphere of radius r_L. The loudspeaker distribution is considered as regular if it satisfyies the orthonormality property of the spherical harmonics (Equ. 22), i.e.:

$$\mathbf{L}^t\mathbf{L} = \mathbf{1} \tag{33}$$

In a similar way as in Section 3.4, a convenient strategy is to locate the loudspeakers at the vertices of a regular or semi-regular polyhedron. Since such a polyhedron with exactly $N_L = (M+1)^2$ ($N_L = 2M+1$ in the 2D case) vertices is seldom found, the polyhedron for which the number of vertices is the closest to $(M+1)^2$ with $N_L > (M+1)^2$ is chosen. The decoding[9] matrix then reduces to:

$$\mathbf{D} = \mathbf{L}^t \tag{34}$$

In the case of 2D rendering, a regular layout is obtained by equally spacing the loudspeakers along a circle of radius r_L.

If the loudspeakers radiate spherical waves, the matrix $\mathbf{L}$ is defined by:

$$\mathbf{L} = \mathbf{W}_L \mathbf{Y}_L \tag{35}$$

where:

$$\mathbf{Y}_L = \begin{bmatrix} Y_{00}^1(\varphi_{L,1},\theta_{L,1}) & Y_{00}^1(\varphi_{L,2},\theta_{L,2}) & \dots & Y_{00}^1(\varphi_{L,N_L},\theta_{L,N_L}) \\ Y_{10}^1(\varphi_{L,1},\theta_{L,1}) & Y_{10}^1(\varphi_{L,2},\theta_{L,2}) & \dots & Y_{10}^1(\varphi_{L,N_L},\theta_{L,N_L}) \\ \vdots & \vdots & & \vdots \\ Y_{MM}^{-1}(\varphi_{L,1},\theta_{L,1}) & Y_{MM}^{-1}(\varphi_{L,2},\theta_{L,2}) & \dots & Y_{MM}^{-1}(\varphi_{L,N_L},\theta_{L,N_L}) \end{bmatrix}$$

and:

$$\mathbf{W}_L = \begin{bmatrix} (-i) & 0 & \dots & 0 \\ 0 & -\frac{h_1^-(kr_L)}{k} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{h_M^-(kr_L)}{k} i^{-(M+1)} \end{bmatrix}$$

Under the assumption that the loudspeakers emit plane waves, the term $\mathbf{W}_L$ is omitted.

4.4. Conclusion

With Section 3 and 4, the overall processing from the recording of the sound scene (primary wave p) to its rendering (synthesized wave $\hat{p}$) has been described. The quality of the reconstruction of the acoustic wave is mainly determined by the encoding order M and the estimate error of the signals $\hat{B}_{mn}^{\sigma}(\omega)$. This issue of how to assess this quality is examined now.

5. ASSESSMENT OF THE SYNTHETIC SOUND WAVE

In order to assess the reproduction quality of a virtual sound scene, various tools and criteria are available. A first strategy is to observe the acoustic waves (either measured or simulated) [4, 11, 14]. The synthetic wave (i.e. the wave $\hat{p}$ reproduced by the loudspeaker array) is compared with the target wave (i.e. the primary soundfield p), which allows one to judge to what extent the wavefront and other macroscopic spatial properties are faithfully reconstructed. Another way is to filter all the information provided by the acoustic waves in order to focus on the information which is used and analysed by the auditory system. This can be done by listening tests, either localization tests [15, 16, 17, 12] or multi-criteria assessments [18, 13, 19]. An alternative is to process the signals perceived at the entrance of the listener's ear by a perception model in order to compute perceptive attributes which account for how the soundfield is perceived [16, 12, 9, 20]. For instance, it is possible to compute the localization cues (for instance the Interaural Time Difference, ITD, and the Interaural Level Difference, ILD). Models of sound localization allow then one to estimate the perceived location of the virtual sound source for comparison with the target location [16, 12]. The *velocity* and *energy* vectors which were introduced by Gerzon [1] are first examples of such models. The source timbre is another attribute which is of interest for the assessment of the reproduction quality [9].

A comprehensive toolkit of objective assessment include:

- the **acoustic pressure** $\hat{p}(\omega, \vec{r})$ measured or computed for a fine spatial sampling of the listening area (i.e. at the center of the loudspeaker array, off-centered location, and at the neighborhood of the loudspeakers),
- the **loudspeaker outputs** (magnitude and phase), in order to quantify the reconstruction effort and its "reasonableness" in terms of energy and loudspeaker distribution,
- the **localization cues**, in order to estimate the accuracy of the perceived location of the virtual sound sources,
- **listening tests** which are helpful to investigate perceptive attributes other than localization.

The theory of auditory localization teaches us that at least three main localization cues should be considered [21]:

- the ITD and the ILD which govern the lateralization (i.e. perception of the source azimuth),
- the Spectral Cues (SC) which are spectral features present in the ear signals (mainly over the band [4-13kHz] [22]) and which are responsible for the perception of elevation.

These localization cues can be estimated from the signals at the entrance of the listener's ear. The signals are either measured by a dummy-head or by microphones inserted in the earcanal of a subject, or simulated. In the latter case, Head Related Transfer Function (HRTF) are used in order that the modelling takes into account the interaction of the acoustic wave with the listener's morphology. The ITD is for instance estimated from the difference between the mean low-frequency phase delay [0-2kHz] of the left and right ear signals [23]. The ILD is computed as the ratio of the high-frequency [1-5kHz] power spectrum (dB) of the left and right ear signals [24]. As for the SC, instead of examining the spectrum pattern, it is proposed to use the Inter-Subject Spectral Difference (ISSD) which was introduced by Middlebrooks to compare two HRTFs [25]. The ISSD is defined as the variance of the difference of two power spectra. In the present problem, it is applied to measure the dissimilarity between the SC reproduced at each ear by the primary wave and the synthetic one. Therefore, with the ITD, the ILD and the ISSD, it is expected to assess the rendering of localization cues in the virtual sound scene, by comparison with those present in the original scene. For a comprehensive analysis, the localization cues should be examined for various locations in the overall listening area. For each location, the orientation of the listener's head may

[9] In [10], Poletti proposed a new interpretation of this solution.

also be varied in order to observe the effect of listener's rotation.

6. CONCLUSION

HOA technology was investigated from several points of view: recording, rendering, format and assessment of spatial audio. One objective was to fill the gap between several issues: theory and practice, encoding and decoding, acoustics and psychoacoutics. Even though theoretical issues of HOA technology are now better understood and mastered, the practice of HOA recording and reproduction still needs further investigation to answer questions such as: what is the optimal encoding order M beyond which no quality improvement can be perceptually expected?, what is the audible effect of increasing the number of loudspeakers for a fixed order M?, what is observed when the listener moves (off-center listening position) or turns his(her) head?...

7. REFERENCES

[1] M. A. Gerzon, "General metatheory of auditory localisation," in *Proceedings of the A.E.S. 92nd Convention*, 1992.

[2] P. Morse and K. Ingard, *Theoretical Acoustics*. McGraw-Hill, 1968.

[3] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging," in *114th AES Convention*, no. 5788, 2003.

[4] R. Nicol, "Restitution sonore spatialisée sur une zone étendue: Application à la téléprésence," Ph.D. dissertation, Université du Maine, 1999.

[5] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging," in *114th AES Convention*, no. 5788, 2003.

[6] S. Moreau, "Etude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore higher order ambisonics: microphone 3d et contrôle de la distance," Ph.D. dissertation, Université du Maine, Le Mans, France, 2006.

[7] J. R. Driscoll and D. M. Healy, "Computing Fourier transforms and convolutions on the 2-sphere," *Adv. Appl. Math.*, vol. 15, pp. 202–250, 1994.

[8] D. Zotkin, R. Duraiswami, and N. Gumerov, "Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, 2010.

[9] A. Solvang, "Representation of high quality spatial audio," Ph.D. dissertation, Norwegian University of Science and Technology, 2009.

[10] A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1024, 2005.

[11] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris 6, 2000.

[12] S. Bertet, "Formats audio 3d hiérarchiques: Caractérisation objective et perceptive des systèmes ambisonics dordres supérieurs," Ph.D. dissertation, INSA Lyon, 2008.

[13] F. Rumsey, *Spatial audio*. Focal Press, 2001.

[14] S. Spors, "Comparison of wave field synthesis and higher-order ambisonics," in *Ambisonics Symposium*, 2009.

[15] A. Sontacchi, M. Noisternig, P. Majdak, and R. Hldrich, "Subjective validation of perception properties in binaural sound reproduction systems," in *21st Int. Audio Eng. Soc. Conference (St Petersburg, Russia)*, 2002.

[16] V. Pulkki and T. Hirvonen, "Localization of virtual sources in multichannel audio reproduction," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 105–119, 2005.

[17] A. Capra, S. Fontana, F. Adiaensen, A. Farina, and Y. Grenier, "Listening tests of the localization performance of stereo dipole and ambisonic systems," in *123rd AES Convention*, 2007.

[18] A. Farina and E. Ugolotti, "Subjective comparison between stereo dipole and 3d ambisonic surround systems for automotive applications," in *16th Int.AES Conf. on Spatial Sound Reproduction (Rovaniemi, Finland)*, 1999.

[19] C. Guastavino and B. Katz, "Perceptual evaluation of multi-dimensional spatial audio reproduction," *J. Acoust. Soc. Am.*, vol. 116, pp. 1105–11 115, 2004.

[20] A. Baskind, "Modles et mthodes de description spatiale de scnes sonores: Application aux enregistrements binauraux," Ph.D. dissertation, Universit Paris 6, 2003.

[21] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*. The MIT Press, Cambridge, Massachusetts, 1983.

[22] P. Guillon, "Individualisation des indices spectraux pour la synthèse binaurale: recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF," Ph.D. dissertation, Université du Maine, Le Mans, France, 2009.

[23] A. Kulkarni, S. Isabell, and H. Colburn, "Sensitivity of human subjects to head-related transfer function phase spectra," *J. Acous. Soc. Am.*, vol. 105, no. 5, pp. 2821–2840, 1999.

[24] V. Larcher, "Techniques de spatialisation des sons pour la réalité virtuelle," Ph.D. dissertation, Université de Paris VI, 2001.

[25] J. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acous. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, 1999.