Politecnico di Milano
University campus of Como

Master of Science in Computer Engineering

School of Industrial and Information Engineering

Denoising in the spherical harmonic domain of sound scenes acquired by compact arrays

Master Graduation Thesis of:
Clara Borrelli
Candidate Id: 840552

Supervisor:
Prof. Antonio Canclini

Assistant Supervisor:
Prof. Fabio Antonacci

Academic Year 2016/2017

Politecnico di Milano
Polo territoriale di Como

Corso di Laurea Specialistica in Ingegneria Informatica

Scuola di Ingegneria Industriale e dell'Informazione

Riduzione del rumore nel dominio delle armoniche sferiche di scene acustiche acquisite con schiere microfoniche compatte

Tesi di Laurea Specialistica di::
Clara Borrelli
Matricola: 840552

Relatore:
Prof. Antonio Canclini

Correlatore:
Prof. Fabio Antonacci

Anno Accademico 2016/2017

Abstract

This thesis aims at developing a new approach for denoising a soundfield that ensures the preservation of the spatial information. During the last decades the interest in capturing, manipulating and rendering 3D acoustic scenes has been grown thanks to the access to new acquisition devices, like spherical microphone arrays, and new methodologies for the representation of the spatial sound field. In some cases the recorded sound scene can be affected from noise or undesired sources. In this thesis is presented a methodology that approximates the sound field as if the noise were not present including its spatial informations. Moreover the overall denoising algorithm is indipendent from any rendering system, thanks to the adopted domain, the spherical harmonic domain, for both the representation and the manipulation of the acoustic scene. In detail, the soundfield acquired by a spherical microphone array is mapped in the spherical harmonic domain through a linear transformation. This domain is convenient for representing the spatial cues of the soundfields, in particular given the choice we have made for the acquisition technique. First source extraction is achieved through a properly designed spatial filter, also called beamformer. It is formulated in the spherical harmonic domain and combines linearly the spherical harmonic coefficients for enhancing the desired signal arriving from the desired source's direction of arrival, estimated in a precedent analysis stage. The spatiality is restored through the application of a bank of filters. In particular the beamformer result acts as input signal for an adaptive filter, while each coefficient of the acquired soundfield corresponds to the desired signal. The final representation of the enhanced soundfield is obtained from the results of this filtering operation. The properties of the adopted representation allow us to perform the adaptive filtering effectively and efficiently. The overall system has been tested simulating a set of simple soundscenes composed of a single source and two types of noise, directional and spatially diffuse, with different reverberation time and signal to noise ratio. The results of the tests suggest that this technique is well-founded and effective. Moreover the underlying principle can be exploited in the future for more complex applications.

Sommario

L'obiettivo di questa tesi è lo sviluppo di un nuovo approccio per il denoising di un campo sonoro disturbato da rumore garantendo la conservazione della spazialità. Negli ultimi decenni si è sviluppato sempre più l'interesse per la registrazione, manipolazione e riproduzione di scene acustiche tridimensionali grazie allo sviluppo di nuovi dispositivi di acquisizione, come array di microfoni, e di nuove tecnologie per la rappresentazione del campo sonoro tridimensionale. Se la registrazione non è effettuata in condizioni ottimali la scena sonora catturata potrebbe essere corrotta da un rumore o da fonti indesiderate. Questa tesi propone una metodologia che approssima il campo sonoro privo di rumore incluse tutte le informazioni spaziali. L'algoritmo proposto è indipendente da un possibile successivo sistema di rendering grazie all'utilizzo del dominio delle armoniche sferiche sia per la rappresentazione che per la manipolazione della scena sonora. Scendendo in dettaglio, il campo sonoro acquisito da una schiera sferica di microfoni è mappato nel dominio delle armoniche sferiche grazie ad una trasformazione lineare. Questo rappresentazione è particolarmente utile per rappresentare le proprietà spaziali di un campo sonoro, in particolar data la tecnica di acquisizione scelta. L'estrazione della sorgente viene eseguita da un filtro spaziale, anche detto beamformer. È formulato nel dominio delle armoniche sferiche e combina linearmente i coefficienti ottenuti dalla precedente trasformazione per estrarre il segnale desiderato proveniente dalla direzione di arrivo della sorgente desiderata, stimata in una precedente fase di analisi. La spazialità viene ripristinata attraverso una serie di filtri adattivi. Il risultato del beamformer è presentato all'ingresso di ogni filtro adattativo, mentre ciascun coefficiente del campo acquisito agisce come segnale desiderato. Le proprietà del dominio delle armoniche sferiche consentono di eseguire il filtraggio adattivo in modo efficace ed efficiente. Il risultato finale, cioè una approssimazione della rappresentazione spaziale del campo privo di rumore, è ottenuto a partire dai risultati Il sistema generale è stato testato simulando una serie di scene acustiche composte da una singola sorgente e da due tipi di rumore, direzionale e spazialmente diffuso, variando i parametri per il riverbero e il rapporto segnale-rumore. I risultati dei test suggeriscono che questa tecnica è valida e che il principio delineato può essere sfruttato in applicazioni più complesse.

Ringraziamenti

Questa tesi è stata svolta presso l'ISPL del Politecnico di Milano.

Il primo ringraziamento va al mio relatore prof. Antonio Canclini per la sua infinita pazienza e disponibilità dimostrate lungo tutto il percorso di tesi. Vorrei anche ringraziare il corelatore prof. Fabio Antonacci per i preziosi consigli che hanno guidato questo lavoro.

Ringrazio il gruppo ISPL/ANTLAB per avermi accolta e ringrazio i miei amici Francesco, Luca e Stella per avermi aiutata ad affrontare le difficoltà grandi e piccole e per tutti i caffè rubati.

Ringrazio Giacomo per avermi fatta ridere quando sembrava tutto nero e per avermi stoicamente sopportata durante la stesura della tesi. E' stata dura ma ora arriva il bello.

Infine grazie ai miei genitori, per avermi sempre aiutata e spronata a seguire le mie passioni, e alle mie sorelle Mari ed Emma alle quali dedico questa tesi.

Contents

List of Figures

List of Tables

Acronyms

ANC Adaptive Noise Canceller. 36, 37

DFT Discrete Fourier Transform. 35

DOA Direction of Arrival. xiv, 19–21, 27–29, 40, 41, 46–48, 50, 53, 54, 71, 72

FDAF frequency-domain adaptive filter. 48–50, 54, 57

FFT Fast Fourier Transform. 35, 36

FIR Finite Impulse Response. 30

HOA Higher Order Ambisonic. xii, xiii, 54, 55

LCMV Linearly Constrained Minimum Variance. 21, 24, 26, 72

LMS Least Mean Squares. 30, 32–34, 36, 37

LTI Linear Time-invariant. 30

MSE Mean Squared Error. xiv, 20, 21, 25, 30, 31, 55

MUSIC Multiple Signal Classification. 27–29, 46, 47, 50, 54, 71

MVDR Minimum Variance Distortionless Response. 21, 24–27

NMSE Normalized Mean Squared Error. xiv, 55–57, 61, 62, 64, 66, 68, 72

RTF relative transfer function. 25, 26

SHT Spherical Harmonic Transform. 6, 15–17, 21, 22, 25, 39, 41, 46, 48, 50, 54, 55, 58, 59, 68

SNR Signal-to-Noise Ratio. xiv, 20, 21, 41–43, 49, 52, 56–61, 64, 65, 72

SRP steered response power. 27

STFT Short Time Fourier Transform. 21, 22, 46, 48, 50, 54

WNG White Noise Gain. 20

Introduction

In the last decades the recording, processing and reproduction of spatial sound have earned an important role in the audio research and industry community. Spatial audio reproduction enables the listener to be virtually immersed in the soundscape, expanding and improving the sound experience. Given its potentialities, immersive audio has been employed in many fields, like audio-visual art, domestic and movie theathers audio systems, virtual-reality applications and medical aid.

The goal of this thesis is to develop a denoising methodology for 3D audio acquired by means of a spherical microphone array. The devised algorithm can be exploited for enhancing recordings acquired in non ideal conditions, like in presence of noise sources, without loosing in the operation the spatial cues of the sound field.

Several paradigms have been presented for the recording and reproduction of the spatial cues of an acoustic scene throughout the years.

The first and very popular method is stereophony, developed by Alan Dower Blumlein and patented in 1931. Stereophony aims at recreating the spatiality of sound scene by means of two loudspeakers, usually located at $\pm$ 30 degrees from the listener position. The sound scene is acquired through a specific recording technique. Among the several proposed solutions, the more popular are the X-Y technique (also called Blumlein pair) which is based only on the amplitude difference, and the A-B technique, based only on the time-of-arrival difference. Alternatively a virtual source can be synthetized at any position between the two loudspeakers by applying a specific panning to a monoaural signal. From the stereophony paradigm many others methods have been devised, like quadrophony, octophony and the popular Dolby Digital 5.1. Note that these methods are not fully periphonic, i.e. do not allow the reproduction of a source coming from any position in the 3D acoustic space. In 1997 Ville Pulkki introduced the Vector Base Amplitude Panning (VBAP) method in [2]. It can be considered a further extension of the stereophonic techniques, since it aims at creating virtual sources extending sterophonic panning to the entire horizontal plane in 2D VBAP, or to vertical and horizontal plane in 3D VBAP.

Another class of surround-sound reproduction systems aim at recreating the entire sound field by means of large loudspeaker arrays rather than the creation of virtual sources in specific positions. The Ambisonic

Figure 1: One of the first examples of art installation that aimed at creating an immersive sound experience created by Karlheinz Stockhausen for the 1970 World Expo in Osaka. The audience sat on a sound-permeable grid below the centre of the sphere on which 50 groups of loudspeakers were arranged.

system, ideated by Micheal Gerzon, is the milestone of this alternative approach and has been the first method to allow full periphony. In Ambisonic real 3D sound scene is recorded by means of a soundfield microphone, which is composed of four closely spaced cardioid and hypercardioid microphone capsules (in Figure 2).

Each capsule feeds one of 4 channels, called W,X,Y and Z, which together compose the so-called B-Format. The signals W,X,Y and Z can be also synthetized starting from a monaural signal. Then the four Ambisonic signals are combined linearly for obtaining the driver signals for a given loudspeakers configuration. The separation between the encoding and the decoding stage ensures the portability of the Ambisonic method. Moreover B-format representation of the spatial sound scene is powerful and allows to easily manipulate or process the soundfield.

The Ambisonic approach has been further extended, the result is Higher Order Ambisonic (HOA). For illustrating HOA a further analisys must be engaged. Ambisonic representation in general is based on solving the wave equation for a central listening spot under the assumption that both sound sources and louspeakers emit plane waves. The solution is a linear combination of spherical harmonic functions up to a maximum order N. The coefficients of this linear combination corresponds to the Ambisonic channels: the B-Format corresponds to the coefficients obtained up to order $N = 1$. In HOA orders $N > 1$ are considered for the extraction of the Ambisonic channels. These coefficients can not be obtained directly from microphones, as in firt-order Ambisonic, but must be derived from the signals of a spherical microphone array through an encoding stage. Given this representation of the soundfield, further processing or manipulation can be applied and then decoded for a loudspeaker setup. Spherical microphone arrays are choosed since they provide equal response for all angles of incidence.

Figure 2: A soundfield microphone

In this thesis a soundfield corrupted by noise is considered, captured by a rigid spherical microphone array and represented in the spherical harmonic domain. Many processing and denoising techniques for spherical microphone array signals have been proposed during the last years. A large part of them are designed specifically for speech signals, for supporting human-to-human and human-to-machine interaction and teleconferencing. Typically the spatiality of the input is not preserved at the ouput, since it is not necessary for the target application. On the other hand spherical microphone array processing for immersive music recording and reproduction is still an open challenge. Only few solutions for signal enhancing or noise suppression are available in literature for the considered context. This work aims at deepening this topic and proposes an innovative approach.

The signal enhancing methodology proposed is able to suppress undesired interferences while preserving all the spatial informations of the soundscape. The result is an estimation of the spatial description of the noise-free sound field, i.e. the soundfield as if the noise were not present. Moreover while the acquisition stage corresponds to the one defined by HOA methods, the denoising system result is indipendent from the rendering phase, hence it can be decoded and played on any reproduction system.

More precisely, the proposed solution makes use of two classical signal processing techniques applied in the adopted domain: beamforming and adaptive filtering. A sketch of the overall system is shown in Figure 3

As already stated, the input of the system is the spherical harmonic domain representation of the noisy soundfield, retrieved from the microphone signals of a spherical array. First a beamformer, formulated in the same domain, extracts the desired source from the noisy soundfield.

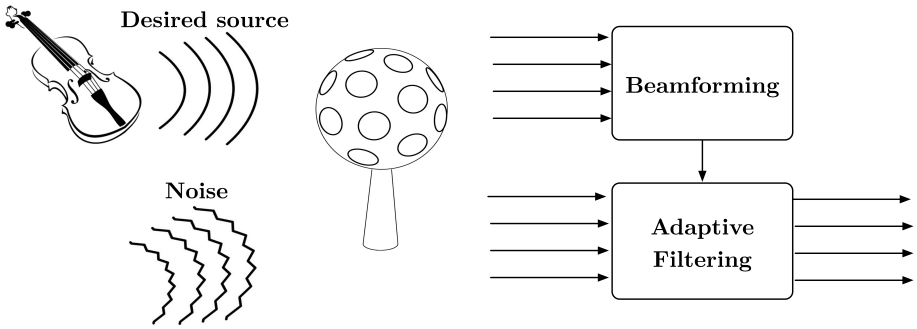

Figure 3: A simple sketch of the proposed solution

More specifically the beamformer implemented is the well-known Maximum Directivity beamformer pointed toward the previously estimated Direction of Arrival (DOA) of the desired source. The channels are linearly combined with the beamformer weights to obtain the estimation of the desired source signal, but doing so the spatial cues of the soundfield are lost. The goal of the successive step, adaptive filtering, is, indeed, to restore the spatiality. In particular a bank of frequency domain adaptive filters (FDAF) are designed to iteratively minimize the Mean Squared Error (MSE) between the input, which corresponds to the beamformer output, and the desired signal, which corresponds to each spherical harmonic coefficient of the acquired soundfield. The results produced by all the frequency domain adaptive filters form together a complete representation of the noise-free soundfield. Moreover the properties of the adopted domain can be exploited to further improve the efficiency of the designed adaptive filtering. This solution has been tested for a directional and for a spatially diffuse noise soundfield.

In Chapter 1 an overview on the theoretical notions for this work is presented, focusing on the spherical harmonic domain and spherical microphone array. In Chapter 2 the state of the art for both signal processing in the spherical harmonic domain and adaptive filtering is investigated. Moreover, a work that share our application context and purpose is presented. In Chapter 3 the proposed system decomposed in two blocks is described in detail. Moreover some useful properties of the noise soundfield represented in the spherical harmonic domain are highlighted. In Chapter 4 the overall denoising technique is validated starting from a set of simulations performed. More specifically two metrics, Signal-to-Noise Ratio (SNR) and Normalized Mean Squared Error (NMSE), are first presented and then their behaviours are analyzed for several setups. Then in Chapter 5 conclusions on the work are outlined

and possible future works are presented.

Chapter 1

Theoretical background

This chapter will introduce the theoretical aspects at the base of this thesis work. The focus will be the spherical harmonic analisys and some related notions of acoustics. This introduction is useful for understanding why spherical harmonic transform is a valid tool for the analysis of sound fields captured from rigid spherical microphone array and, modelling this device as a rigid sphere, how it interacts with the sound field itself.

1.1 Spherical coordinate system

For dealing with spherical microphone array and spherical harmonic analysis the spherical coordinate system is the more suitable domain. Hence we are going to define it starting with the Cartesian coordinate system.

The standard Cartesian coordinate system is given by

$$\mathbf{x} \equiv (x, y, z) \in \mathbb{R}^3 \tag{1.1}$$

where $\mathbb{R}^3$ is the three-dimensional space of real numbers.

The spherical coordinate system is an alternative coordinate system for three-dimensional space and represents all the positions on a sphere or radius r. Therefore a point is represented by the triplet

$$\mathbf{r} \equiv (r, \theta, \phi) \tag{1.2}$$

where θ is inclination angle, measured downwards from the z-axis and ϕ is the azimuth angle, measured from the x-axis towards the y-axis, as shown in Figure 1.1.

To define a unique set of coordinates we set a range of values for each

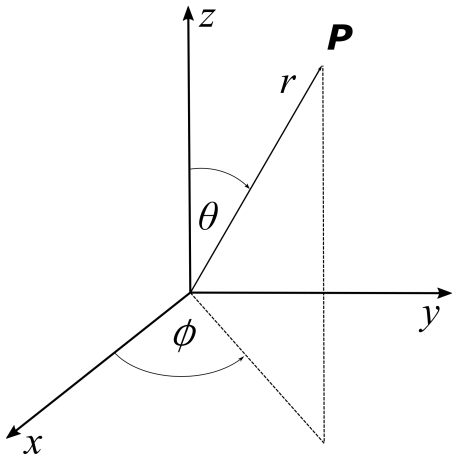

Figure 1.1: Spherical coordinate system

component:

$$\begin{aligned} r &\geq 0 \\ 0 \leq \theta &\leq \pi \\ 0 \leq \phi &< 2\pi). \end{aligned} \tag{1.3}$$

The position of a point P expressed in spherical coordinates as $\mathbf{r} = (r, \theta, \phi)$ is represented in Cartesian coordinates $\mathbf{x} = (x, y, z)$ where

$$\begin{aligned} x &= r \sin\theta \cos\phi \\ y &= r \sin\theta \sin\phi \\ z &= r \cos\theta. \end{aligned} \tag{1.4}$$

On the other hand the spherical coordinates can be computed from Cartesian ones with

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2} \\ \theta &= \arccos\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right) \\ \phi &= \arctan\left(\frac{y}{x}\right). \end{aligned} \tag{1.5}$$

1.2 Acoustical background

In this section some basic notions of acoustics for the analysis of sound fields are presented. Wave and Helmtoz equation will be presented for both Cartesian and spherical coordinates and the solution will be given for spherical coordinates only. After that two specific cases will be ad-

dressed: sound pressure generated from a plane wave and sound pressure in presence of a rigid sphere, both formulated with the help of spherical harmonic functions. This set of notions creates a theoretical framework that will be useful for dealing with spherical microphone array processing.

1.2.1 Wave equation

Let $p(\mathbf{x}, t)$ be the sound pressure in position $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ at time t, then it satisfies [3]

$$\nabla^2_{\mathbf{x}} p(\mathbf{x}, t) - \frac{1}{c^2} \frac{\partial^2 p(\mathbf{x}, t)}{\partial t^2} = 0 \tag{1.6}$$

where c is the sound speed constant (343 m/s in typical conditions) and $\nabla^2_{\mathbf{x}}$ is the Laplacian operator in Cartesian coordinates, defined for a function $f(x, y, z)$ as

$$\nabla^2_{\mathbf{x}} f \equiv \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}. \tag{1.7}$$

If we consider a single plane wave sound field, then the pressure can be expressed as [4]

$$p(\mathbf{x}, t) = P(\mathbf{x}) e^{i\omega t} \tag{1.8}$$

where ω is the radial frequency and $P(\mathbf{x})$ can be seen as the amplitude of sound pressure in the position $\mathbf{x}$. Defining the wave number $k = \omega/c$ it is possible to change notation in $p(k, \mathbf{x})$ to explicit the dependence on wave number: this notation represents the result of the Fourier transform for the frequency $\omega = k/c$ assuming we are describing a stationary broadband sound field. The Helmoltz equation is obtained by substituting (1.8) in (1.6), the result is :

$$\nabla^2_{\mathbf{x}} p(k, \mathbf{x}) + k^2 p(k, \mathbf{x}) = 0. \tag{1.9}$$

The wave equation can be reformulated in the spherical coordinates $\mathbf{r} = (r, \theta, \phi)$, starting from the definition of the Laplacian in the spherical coordinates for a function $f(r, \theta, \phi)$

$$\nabla^2_{\mathbf{r}} f \equiv \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \frac{\partial^2 f}{\partial \phi^2}. \tag{1.10}$$

Then the wave equation will be

$$\nabla^2_{\mathbf{r}} p(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial p(\mathbf{r}, t)}{\partial t^2} = 0. \tag{1.11}$$

Also the Helmoltz equation can be written in spherical coordinates as

$$\nabla^2_{\mathbf{r}} p(\mathbf{r}, k) + k^2 p(\mathbf{r}, k) = 0. \tag{1.12}$$

The wave equation in the spherical coordinates can be solved by applying separation of variables: defining $p(\mathbf{r}, t) = p(\mathbf{r})e^{i\omega t}$ the amplitude of pressure, it can be written as the product of four terms [3, 4]

$$p(\mathbf{r}, t) = R(r)\Theta(\theta)\Phi(\phi)T(t) \tag{1.13}$$

Substituing (1.13) in (1.11) the equation is decomposed in four partial equations, each one in a separate variable R, Θ, Φ and T. The equation that expresses the time dependence is

$$\frac{d^2T}{dt^2} + \omega^2 T = 0 \tag{1.14}$$

and has solution

$$T(t) = e^{i\omega t}, \quad \omega \in \mathbb{R}. \tag{1.15}$$

The term dependent on ϕ is

$$\frac{d^2\Phi}{d\phi^2} + m^2\Phi = 0 \tag{1.16}$$

and the fundamental solution is

$$\Phi(\phi) = e^{im\phi}, \quad m \in \mathbb{Z} \tag{1.17}$$

where m is an integer to represent the periodicity of Φ and $\phi \in [0, 2\pi)$. The term dependent on θ is

$$\frac{d}{d\mu}\left[(1-\mu^2)\frac{d}{d\mu}\Theta\right] + \left[n(n+1) - \frac{m^2}{1-\mu^2}\right]\Theta = 0 \tag{1.18}$$

where $\mu = \cos\theta$. This is also called the associated Legendre differential equation. We select as solution a function called Legendre function of the first kind (see Section 1.2.3):

$$\Theta(\theta) = P_n^m(\cos\theta) \quad n \in \mathbb{N}, \quad m \in \mathbb{Z}. \tag{1.19}$$

The term dependent on r is

$$\rho^2 \frac{d^2V}{d\rho^2} + 2\rho \frac{dV}{d\rho} + \left[\rho^2 - n(n+1)\right] V = 0 \tag{1.20}$$

where $\rho = kr$ and $V(\rho) \equiv R(r)$. This is the spherical Bessel equation that has as solutions spherical Bessel functions of first kind $j_n(kr)$ or spherical Hankel of the first kind $h_n(kr)$ or both.

Finally combining the solutions for r, t, θ and ϕ we obtain a solution of the wave equation in spherical coordinates:

$$p(\mathbf{r}, t) = j_n(kr) Y_n^m(\theta, \phi) e^{i\omega t} \tag{1.21}$$

or

$$p(\mathbf{r}, t) = h_n^1(kr) Y_n^m(\theta, \phi) e^{i\omega t} \tag{1.22}$$

or a combination of these two. In the previous equation $Y_n^m(\theta, \phi)$ are the spherical harmonic functions, $j_n(kr)$ is the Bessel function of the first kind and $h_n^{(1)}$ is the Hankel function of the first kind: all these functions will be presented in detail in the next functions.

1.2.2 Spherical Harmonics

Spherical harmonic functions are special functions defined over the surface of a sphere: they form a complete set of orthonormal functions, hence any function defined on a sphere can be described in terms of spherical harmonics and expansion coefficients. They are defined and used in many contexts, the definition given for acoustic applications is:

$$Y_n^m(\theta, \phi) \equiv \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi} \tag{1.23}$$

where $P_n^m(\cdot)$ are the associated Legendre functions, $m \in \mathbb{Z}$ is the function degree and $n \in \mathbb{N}$ is the function order (with $|m| < n$). Note that n expresses the dependence of the spherical harmonic on θ while m the dependence on ϕ with the exponential term.

Due to properties of P_n^m (explained later in Section 1.2.3) we have that:

$$Y_n^{-m}(\theta, \phi) = (-1)^m Y_n^m(\theta, \phi)^* \tag{1.24}$$

The spherical harmonics are orthonormal, which means:

$$\int_0^{2\pi} d\phi \int_0^{\pi} Y_n^m(\theta, \phi) Y_{n'}^{m'}(\theta, \phi)^* \sin\theta d\theta = \delta_{nn'} \delta_{mm'} \tag{1.25}$$

where the Kroenecher delta is defined as

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{1.26}$$

Since they form a complete set of orthonormal functions the closure relation is valid:

$$\sum_{n=0}^{\infty} \sum_{m=-n}^{n} Y_n^m(\theta, \phi) Y_n^m(\theta', \phi')^* = \delta(\phi - \phi')\delta(\cos\theta - \cos\theta'). \tag{1.27}$$

The most interesting property of spherical harmonics is that any function on a sphere can be decomposed in terms of them, meaning

$$g(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_{nm} Y_n^m(\theta, \phi) \tag{1.28}$$

where the coefficients A_{nm} are complex values. For the orthonormality of these functions we have that

$$A_{nm} = \int_{\Omega \in S^2} Y_n^m(\theta, \phi)^* g(\theta, \phi) d\Omega \tag{1.29}$$

where Ω is the solid angle defined as

$$\int d\Omega = \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta d\theta. \tag{1.30}$$

Equations (1.28) (1.29) form the Spherical Harmonic Transform (SHT), which will be, together with Fourier transform, in their discrete formulation the basic blocks of the processing chain in this thesis.

Note that for $m = 0$

$$Y_n^0(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi}} P_n(\cos\theta) \tag{1.31}$$

where P_n is the Legendre polynomial.

Is possible to define a real version of the SHT as

$$g(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} B_{nm} R_n^m(\Omega) d\Omega \tag{1.32}$$

and

$$B_{nm} = \int_{\Omega \in S^2} R_n^m(\Omega) g(\theta, \phi) d\Omega. \tag{1.33}$$

The real-valued spherical harmonic functions $R_n^m(\Omega)$ are expressed in terms of the imaginary and real parts of the complex ones as

$$\begin{aligned} R_n^m(\Omega) &= \begin{cases} \sqrt{2}(-1)^m \Im(Y_n^{-m}(\Omega)) & m < 0 \\ Y_n^0 & m = 0 \\ \sqrt{2}(-1)^m \Re(Y_n^m(\Omega)) & m > 0 \end{cases} \\ &= \begin{cases} \frac{i}{\sqrt{2}}(Y_n^m(\Omega) - (-1)^m Y_n^{-m)}(\Omega)) & m < 0 \\ Y_n^0(\Omega) & m = 0 \\ \frac{i}{\sqrt{2}}(Y_n^{-m}(\Omega) + (-1)^m Y_n^m(\Omega)) & m > 0 \end{cases} \end{aligned} \tag{1.34}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denotes the real and the imaginary parts of a complex number. The real spherical harmonic functions can be rewritten in terms of the associated Legendre polynomials defined in (1.36) as:

$$\begin{aligned} R_n^m(\Omega) = (-1)^m \sqrt{\frac{(2n+1)!}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^m(\cos\theta) \times \\ \begin{cases} \sqrt{2}\sin(|m|\phi) & m < 0 \cdot \\ 1 & m = 0 \\ \sqrt{2}\cos(m\phi) & m < 0 \end{cases} \end{aligned} \tag{1.35}$$

The real spherical harmonic functions mantain the same properties of the complex one, including orthogonality.

1.2.3 Legendre Functions

In this section the associated Legendre functions and the Legendre polynomial will be presented in detail. The associated Legendre polynomials are the canonical solutions of the general Legendre equation and they can be defined as derivatives of the Legendre polynomial [4]:

$$P_n^m(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m P_n(x)}{dx^m}, \quad x \in [-1, 1] \tag{1.36}$$

where the $(-1)^m$ is known as the Condon-Shortley phase, n is the order, m is the degree of the polynomial and $P_n(x)$ is the Legendre polynomial defined as:

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n \tag{1.37}$$

Note that the associated Legendre polynomials for negative values of m are given by

$$P_n^{-m}(x) = (-1)^m \frac{(n-m)!}{(n+m)!} P_n^m(x) \tag{1.38}$$

and this property is reflected in the behaviour of spherical harmonics functions over the inclination angle θ. Associated Legendre polynomials of different order n and same degree m are orthogonal since they satisfy

$$\int_{-1}^{1} P_n^m(x) P_{n'}^m(x) dx = \frac{2}{2n+1} \frac{(n-m)!}{(n+m)!} \delta_{nn'}, \quad -n \leq m \leq n \tag{1.39}$$

and this is responsible for the orthogonality of spherical harmonics when integrating along θ. Combining this property and the orthogonality of exponential functions orthogonality of spherical harmonics can be directly derived.

The values of associated Legendre function for $m = 0$ (equals to $Y_n^0(\theta, \phi)$) are exactly the Legendre polynomial:

$$P_n(x) = P_n^0(x). \tag{1.40}$$

1.2.4 Spherical Hankel and Bessel functions

In the previous section a solution for the wave equation is derived as function of spherical Hankel and Bessel functions, which will be presented in this section.

The spherical Bessel functions of first kind $j_n(x)$ and of the second kind $y_n(x)$ can be written as [4]

$$j_n(x) = (-1)^n x^n \left(\frac{1}{x} \frac{d}{dx} \right)^n \frac{\sin(x)}{x} \tag{1.41}$$

$$y_n(x) = -(-1)^n x^n \left(\frac{1}{x} \frac{d}{dx} \right)^n \frac{\cos(x)}{x}. \tag{1.42}$$

The spherical Hankel functions of the first kind $h_n^{(1)}(x)$ and of the second kind $h_n^{(2)}(x)$ can be written as

$$h_n^{(1)}(x) = -i(-1)^n x^n \left(\frac{1}{x} \frac{d}{dx} \right)^n \frac{e^{ix}}{x} \tag{1.43}$$

$$h_n^{(2)}(x) = i(-1)^n x^n \left(\frac{1}{x} \frac{d}{dx} \right)^n \frac{e^{-ix}}{x}. \tag{1.44}$$

The relationship between spherical Hankel functions and spherical Bessel function is:

$$h_n^{(1)}(x) = j_n(x) + iy_n(x) \tag{1.45}$$

and

$$h_n^{(2)}(x) = j_n(x) - iy_n(x). \tag{1.46}$$

The spherical Hankel and spherical Bessel functions are linked to the Bessel function $J_\alpha(x)$ and the Hankel function $H_\alpha(x)$ by the relationship:

$$j_n(x) = \sqrt{\frac{\pi}{2x}} J_{n+\frac{1}{2}}(x) \tag{1.47}$$

$$h_n^{(1)}(x) = \sqrt{\frac{\pi}{2x}} H_{n+\frac{1}{2}}(x). \tag{1.48}$$

In the Tables 1.1 and 1.2 the first 4 spherical Hankel and Bessel functions are presented.

$j_0(x) = \frac{\sin x}{x}$
$j_1(x) = -\frac{\cos x}{x} + \frac{\sin x}{x^2}$
$j_2(x) = -\frac{\sin x}{x} - \frac{3\cos x}{x^2} + \frac{3\sin x}{x^3}$
$j_3(x) = \frac{\cos x}{x} - \frac{6\sin x}{x^2} - \frac{15\cos x}{x^3} + \frac{15\sin x}{x^4}$

Table 1.1: Spherical Bessel functions of the first kind $j_n(x)$ for $n = 0, 1, 2, 3$

$h_0^{(1)}(x) = \frac{e^{ix}}{ix}$
$h_1^{(1)}(x) = -\frac{e^{ix}(i+x)}{x^2}$
$h_2^{(1)}(x) = \frac{ie^{ix}(-3+3ix+x^2)}{x^3}$
$h_3^{(1)}(x) = \frac{e^{ix}(-15i-15x+6ix^2+x^3}{x^4}$

Table 1.2: Spherical Hankel functions of the first kind $h_n^{(1)}(x)$ for $n = 0, 1, 2, 3$

The behaviour of spherical Bessel and Hankel functions can be further analyzed. We approximate $j_n(x) \approx \frac{x^n}{(2n+1)!!}$ for $x << 1$, hence toward 0 the 0-th order amplitude is constant while amplitude of functions with

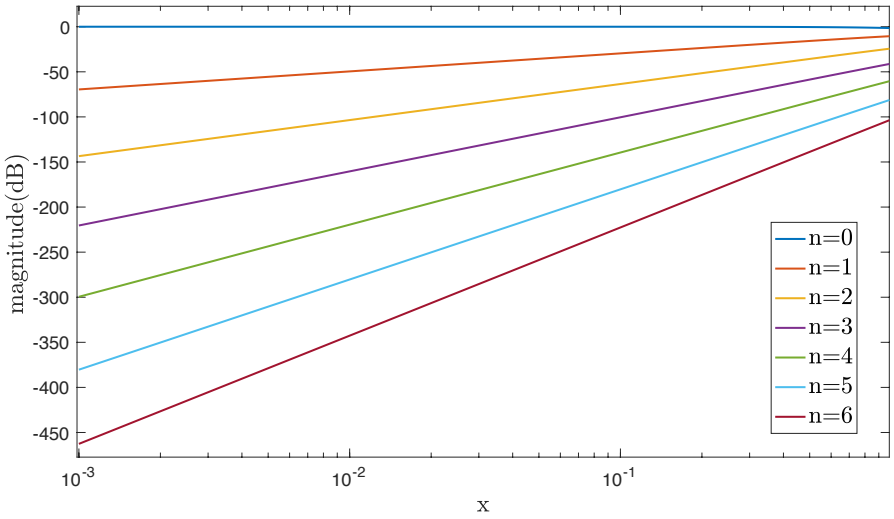

Figure 1.2: Magnitude of the spherical Bessel function of the first kind, $|j_n(x)|$, for $n = 0, ..., 6$ and for $x < 1$

$n > 0$ goes asymptotically to $-\infty$ as shown in Figure 1.2. When $x >> n$ the amplitude of $j_n(x)$ decays as $\frac{1}{x}$ approximately for all the orders since they can be approximated by $j_n(x) \approx \frac{1}{x}\sin(x - n\pi/2)$. Moreover spherical Bessel functions have zeros: for $j_0(x)$ zeros are for $\pm k\pi$ with $k \in \mathbb{N}$, for higher orders the first zeros are positioned from $x > \pi$, like shown in Figure 1.3.

Spherical Hankel functions on the contrary diverge toward the origin since the small argument approximation is $h_n^{(1)}(x) \approx -i\frac{(2n-1)!!}{x^{n+1}}$ for $x << 1$ as shown in Figure 1.4; for large values of x $h_n^{(1)}(x)$ decays for all the orders with the same trend since can be approximated with $h_n^{(1)}(x) \approx (-i)^{n+1}\frac{e^{ix}}{x}$ for $x >> \frac{n(n+1)}{2}$ as we can see in Figure 1.5

1.2.5 Plane waves

In this section we are going to analyze the scenario of a single plane wave from direction (θ_a, ϕ_a) with wavevector $\mathbf{k} = (k, \theta_a, \phi_a)$ in free field. Since plane wave is a solution of homogenous wave equation, the sound pressure given by a plane wave can be written as a combination of generic solutions of the wave equation in spherical coordinates, i.e. spherical Bessel and spherical harmonics functions [4]

$$p(k, r, \theta, \phi) = e^{i\mathbf{k}\cdot\mathbf{r}} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi i^n j_n(kr)[Y_n^m(\theta_a, \phi_a)]^* Y_n^m(\theta, \phi). \quad (1.49)$$

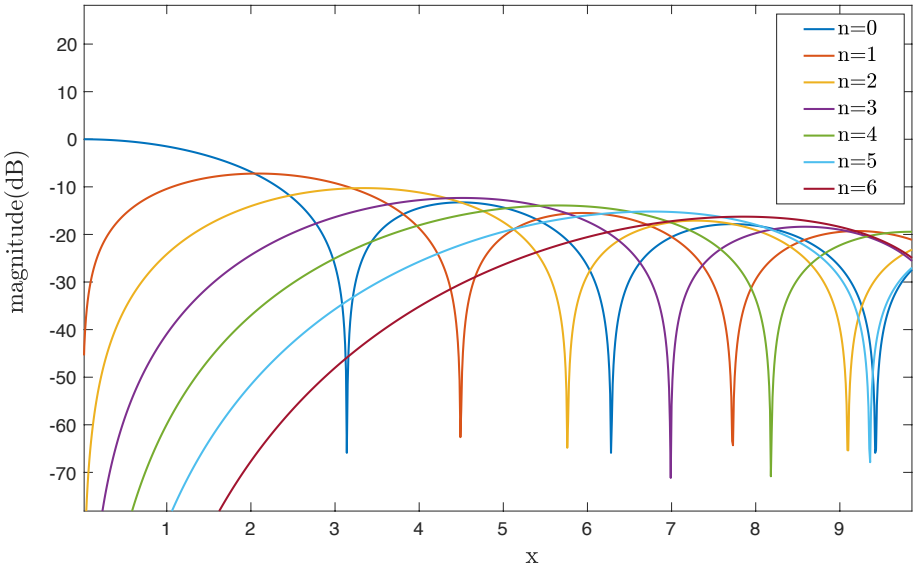

Figure 1.3: Magnitude of the spherical Bessel function of the first kind, $|j_n(x)|$, for $n = 0, ..., 6$

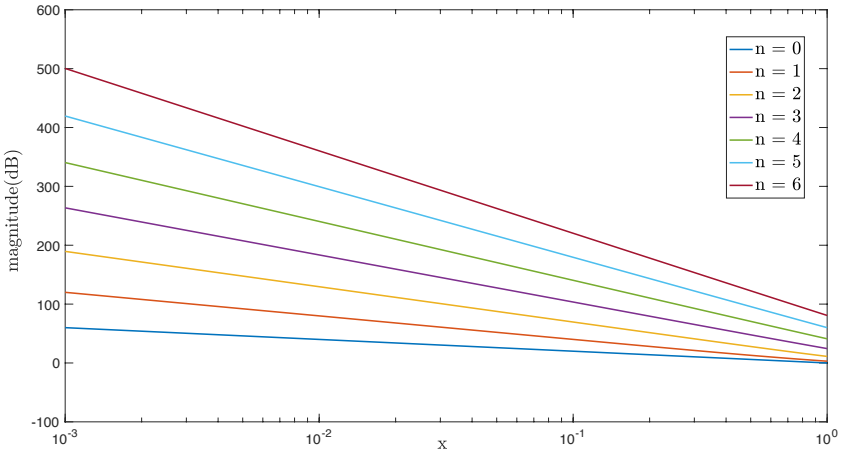

Figure 1.4: Magnitude of the spherical Hankel function of the first kind, $|h_n^{(1)}(x)|$, for $n = 0, ..., 6$ and for $x < 1$

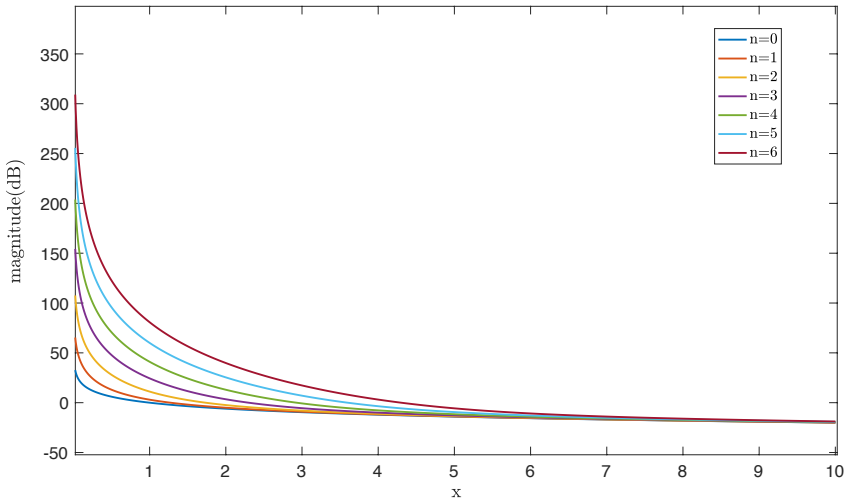

Figure 1.5: Magnitude of the spherical Hankel function of the first kind, $|h_n^{(1)}(x)|$, for $n = 0, ..., 6$

By writing the scalar product $\mathbf{k} \cdot \mathbf{r} = kr\cos\Theta$ and exploting spherical harmonic functions properties (1.49) can be further reduced to

$$p(k, r, \Theta) = e^{ikr\cos\Theta} = \sum_{n=0}^{\infty} i^n j_n(kr)(2n+1)P_n(\cos\Theta) \tag{1.50}$$

called plane wave expansion expression.

In pratice infinite summation will be truncated to order N with

$$p(k, r, \theta, \phi) \approx \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi i^n j_n(kr)[Y_n^m(\theta_a, \phi_a)]^* Y_n^m(\theta, \phi). \tag{1.51}$$

Sound pressure $p(k, r, \theta, \phi)$ is a function defined over a sphere of radius r therefore spherical harmonic expansion defined in (1.28) can be applied:

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm}(kr) Y_n^m(\theta, \phi) \tag{1.52}$$

where, from previous equations,

$$p_{nm}(k, r) = 4\pi j_n(kr)[Y_n^m(\theta_a, \phi_a)]^*. \tag{1.53}$$

This equation shows that magnitude of p_{nm} has the same behaviour of

$j_n(kr)$, hence it decays for $n > kr$: this means that truncated expansion is still a good approximation if $kr < N$ and that the truncation error will be negligible. The sound pressure given by a plane wave is an order limited function defined on a sphere, which means that for $n > N$ $p_{nm} \approx 0$. In general this condition provide us a upper limit for accurate reconstruction of soundfield using spherical harmonics functions.

1.2.6 Sound pressure for rigid sphere

In this section we illustrate the sound pressure modifications in presence of a rigid sphere. Since the acquisition device of our work will be a rigid spherical microphone array, this derivation will be useful in the next chapters.

The sound pressure in this case is composed of two components, the incident sound field and the scattered sound field. The first one is the sound field that would be present in the free-field, the second component is given by the scattering of the incident sound field on the surface of the rigid sphere.

The presence of a perfectly rigid sphere of radius r_a adds a condition on the radial component of velocity on its surface which is [4]

$$u_r(k, r_a, \theta, \phi) = 0. \tag{1.54}$$

Using the Euler equation in spherical coordinates and decomposing the sound pressure in the two components, incident $p_i(k, r, \theta, \phi)$ and scattered $p_s(k, r, \theta, \phi)$, we obtain

$$\frac{\partial}{\partial r}[p_i(k, r, \theta, \phi) + p_s(k, r, \theta, \phi)] = 0. \tag{1.55}$$

Applying the spherical harmonic expansion for the incident component the result is

$$p_i(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}(k) 4\pi i^n j_n(kr) Y_n^m(\theta, \phi) \tag{1.56}$$

assuming the sound field is composed of multiple plane waves. Note that this is obtained combining the single plane wave solution seen in Section 1.2.5 and the properties of spherical harmonics. The term $a_{nm}(k)$ is the spherical harmonic decomposition of directional amplitude density $a(k, \theta_a, \phi_a)$.

The spherical harmonic expansion for scattered pressure component is

$$p_s(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} c_{nm}(k) h_n^{(2)}(kr) Y_n^m(\theta, \phi). \tag{1.57}$$

Substituting (1.56) and (1.57) in (1.55) we obtain

$$c_{nm}(k) = -a_{nm} 4\pi i^n \frac{j'_n(kr_a)}{h'^{(2)}_n(kr_a)}. \tag{1.58}$$

If the sound pressure is rewritten adding (1.56) and (1.57) and substituing $c_{nm}(k)$ with (1.58) then the result is

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}(k) 4\pi i^n \left[j_n(kr) - \frac{j'_n(kr_a)}{h'^{(2)}_n(kr_a)} h^{(2)}_n(kr) \right] Y^m_n(\theta, \phi). \tag{1.59}$$

By denoting

$$b_n(kr) = 4\pi i^n \left[j_n(kr) - \frac{j'_n(kr_a)}{h'^{(2)}_n(kr_a)} h^{(2)}_n(kr) \right] \tag{1.60}$$

the so called *mode strength coefficients* for rigid sphere are obtained (tipically used for $r = r_a$) and the pressure outside rigid sphere can be written as

$$p_{nm}(kr) = a_{nm}(k) b_n(kr). \tag{1.61}$$

In Figure 1.6 the magnitude of $b_n(kr)$ is plotted on kr axis. As in the free field case, the spherical harmonics coefficients decay for $n > kr$ hence the truncation at $kr < N$ of the infinite summation produce a negligible error. Similarly to the plane waves case, the sound pressure around a rigid sphere is an order limited function on sphere.

1.3 Spatial sampling using spherical microphone array

In this thesis a sound field is going to be captured and elaborated using a spherical microphone array. Recording the sound pressure with a set of microphones corresponds to sampling it, hence the quality of reconstruction depends on the used sampling configuration, i.e. couples of positions and weights. The classic sampling functions (Gaussian sampling, Uniform and Nearly Uniform sampling,ecc) will not be presented in this first section but it will be explained how to compute spherical harmonic coefficients starting from an arbitrary configuration of microphones. Then we will present how in practice spatial sampling is implemented with a spherical microphone array, since sampling on a sphere is an intuitive sampling method, and which problems is necessary to deal with.

In the previous section, analyzing the plane wave sound pressure and rigid sphere sound pressure, the concept of order-limited function

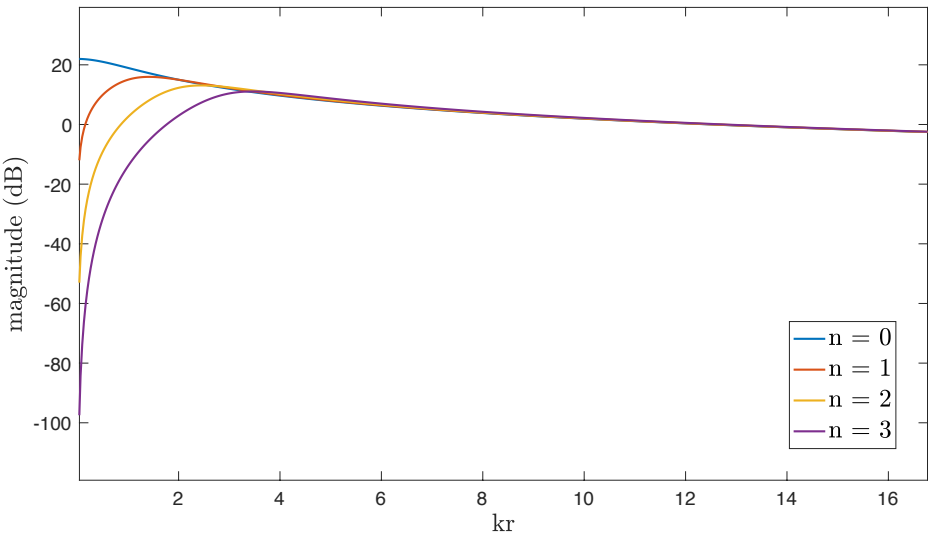

Figure 1.6: Magnitude of the mode strength coefficients for a rigid sphere, $|b_n(kr)|$ for $n = 0, 1, 2, 3$

has been introduced. Similarly to the Nyquist sampling theorem, which states that the function must be bandlimited to guarantee perfect reconstruction from the samples, the sampling methods on the sphere require the functions to be order limited, such that it can be represented with a finite numer of basis functions. Note that in the following exposition the focus will be on spherical harmonic analysis only, without taking in account the frequency domain representation.

Consider a function $g(\theta, \phi)$ defined over a unit sphere, sampling formulation can be derived starting from a general quadrature problem. In fact, given a set of samples on the sphere (θ_q, ϕ_q) and sampling weigths α_q, our goal is to approximate the integral of the function as a sum of the samples $g(\theta_q, \phi_q)$ with $q = 1, .., Q$, which means [4]:

$$\int_0^{2\pi} \int_0^{\pi} g(\theta, \phi) \sin\theta d\theta d\phi \approx \sum_{q=1}^{Q} \alpha_q g(\theta_q, \phi_q). \tag{1.62}$$

Starting from (1.29) and defining $g(\theta, \phi) = f(\theta, \phi)[Y_n^m(\theta, \phi)]^*$ (which means traducing the quadrature problem in a reconstruction problem), (1.62) can be reformulated as an estimate of the SHT coefficients of f_{nm} starting from the function samples:

$$f_{nm} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi)[Y_n^m(\theta, \phi)]^* \sin\theta d\theta d\phi \approx \sum_{q=1}^{Q} \alpha_q f(\theta_q, \phi_q)[Y_n^m(\theta_q, \phi_q)]^*. \tag{1.63}$$

This approximation in case of order limited functions and Q big enough

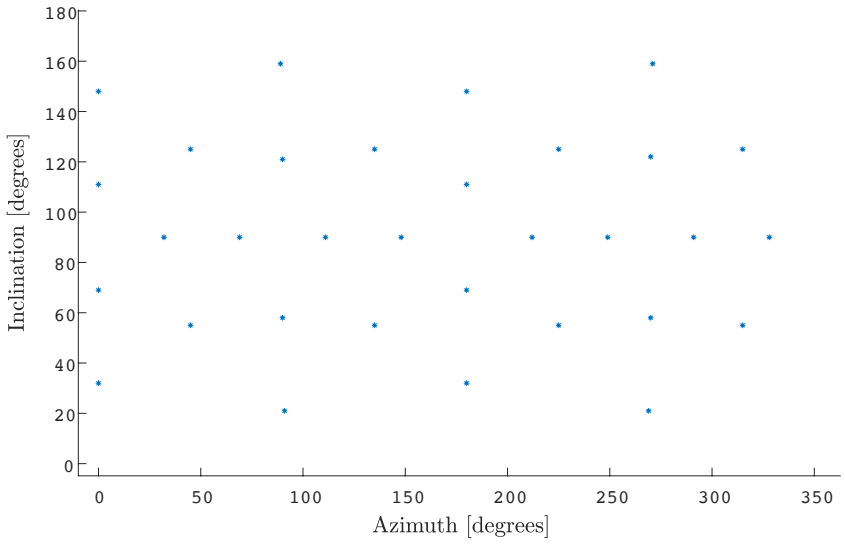

Figure 1.7: Sampling distribution for Eigenmike spherical microphone array

becomes an equality, which means that the reconstruction will be exact. Moreover it can be shown that orthogonality of spherical harmonics is mantained at least for a range of orders, which is one of the basic properties of ideal sampling.

Given an arbitrary set of positions of sampling (in Figure 1.7 the microphones positions for an Eigenmike are plotted) our aim is to compute first the weigths α_q and then use (1.63) for computing the SHT coefficients. Therefore consider an order limited function $f(\theta, \phi)$ such that $f_{nm} = 0$ for $n > N$. If this function is sampled at positions (θ_q, ϕ_q), the result is a set of samples $f(\theta_q, \phi_q)$. Applying the SHT we obtain

$$f(\theta_q, \phi_q) = \sum_{n=0}^{N} \sum_{m=-n}^{n} f_{nm} Y_n^m(\theta_q, \phi_q) \quad q = 1, ..., Q \tag{1.64}$$

that can be written in matricial form as

$$\mathbf{f} = \mathbf{Y}\mathbf{f_{nm}} \tag{1.65}$$

where

$$\mathbf{f} = [f(\theta_1, \phi_1), f(\theta_2, \phi_2), ..., f(\theta_q, \phi_q)]^T \tag{1.66}$$

$$\mathbf{f_{nm}} = [f_{00}, f_{1(-1)}, f_{10}, ..., f_{NN}]^T \tag{1.67}$$

$$\mathbf{Y} = \begin{bmatrix} Y_0^0(\theta_1,\phi_1) & Y_1^{-1}(\theta_1,\phi_1) & Y_1^0(\theta_1,\phi_1) & ... & Y_N^N(\theta_1,\phi_1) \\ Y_0^0(\theta_2,\phi_2) & Y_1^{-1}(\theta_2,\phi_2) & Y_1^0(\theta_2,\phi_2) & ... & Y_N^N(\theta_2,\phi_2) \\ \vdots & \vdots & \vdots & ... & \vdots \\ Y_0^0(\theta_Q,\phi_Q) & Y_1^{-1}(\theta_Q,\phi_Q) & Y_1^0(\theta_Q,\phi_Q) & ... & Y_N^N(\theta_Q,\phi_Q) \end{bmatrix}. \tag{1.68}$$

When $Q = (N+1)^2$ the solution is obtained as

$$\mathbf{f_{nm}} = \mathbf{Y^{-1}f} \tag{1.69}$$

that requires $\mathbf{Y}$ to be invertible. When $Q > (N+1)^2$, i.e. oversampling is performed, the solution is obtained in least-squares sense

$$\mathbf{f_{nm}} = \mathbf{Y^{\dagger}f} \tag{1.70}$$

since the systems is over-determined, with $\mathbf{Y}^{\dagger} = (Y^H Y)^{-1} Y^H$. If $Q < (N+1)^2$ then the sound field is undersampled and (1.65) does not lead to a unique solution. Equations (1.65) and (1.70) are the decomposition and the expansion formulas of the discrete SHT.

Reformulating the problem in terms of sampling weights it follows that:

$$f_{nm} = \sum_{q=1}^{Q} \alpha_q^{nm} f(\theta_q, \phi_q) \tag{1.71}$$

where α_q^{nm} are the elements of the matrices Y^{-1} or $Y^{\dagger}$ with index $[(n^2 + n + m), q]$.

In this thesis the acquisition device is a rigid spherical microphone array, which is a set of microphones arranged on a sphere made of rigid and reflective material. Only in the last few years spherical microphone array have begun to be available on the market hence algorithms for analysis and processing are still a research theme.

The array configuration can determine many aspects of the sound field analysis [4], in particular determines the spatial aliasing behaviour. In the previous section the concept of order-limited functions has been presented and it is a necessary assumption for perfect reconstruction. Unfortunately in real applications higher orders harmonics can be different from zero and hence spatial aliasing will be present, especially at high frequencies. In general spatial aliasing determines an erroneous representation of the spatial soundfield and leads to a performance degradation in array processing. The condition necessary for having small errors caused by spatial aliasing is that $kr << N$ where N is the maximum reconstructed order and r is the radius of the spherical microphone array. This leads to setting an upper limit for operating frequency which

Figure 1.8: Eigenmike spherical microphone array

is $f << \frac{Nc}{2\pi r}$ [1].

For example to have good reconstruction using an Eigenmike microphone array [5] of $Q = 32$ microphones and radius $r = 0.042\,\text{m}$ the maximum order achievable is $N = 5$ since $Q \geq (N+1)^2$ and the operating frequency is limited by $f = 6.5\,\text{kHz}$. Note that if for some reasons is necessary to reach a very high order N then the radius r will be so large that dealing with the microphone array can be inconvenient in practice and the scattered part of the sound field can interact with other elements of the room (like walls). Note that the analysis of the interaction of a sound field with a rigid sphere addressed in Section 1.2.6 provides a theoretical model for the rigid spherical microphone array in the spherical harmonic domain.

Chapter 2

State of the Art

In this chapter an outline will be given for two signal processing wide areas that will be exploited in the solution proposed in this thesis. First processing for spherical microphone array will be addressed, focusing on beamformers and their application in signal enhancement for noisy enviroments. The second topic is adaptive filtering: we start with the classical Wiener formulation and continue with an overview of adaptive tecniques both in time and frequency domain. Finally an interesting denoising system proposed recently will be presented.

2.1 Spherical microphone array processing

Spherical microphone array processing has a key role in this thesis proposal. A number of methods for manipulating the spatial representation of the sound field captured from the array or extracting acoustic parameters have been proposed in the literature in the last decade. Typical applications are signal enhancing, dereverberation or DOA estimation of a source in a sound scene. Most of them define or include in more complex systems a spatial filter, also called beamformer.

Beamforming is a classic signal processing tecnique that allows to enhance the signal coming from a specific direction and attenuate the noise signal coming from an undesired direction: the principal operation is the linear combination of a set of weights, designed for achieving a specific goal, with the microphone signals.

In this section we are going to review its more populare usage, i.e. signal enhancement. All the tecniques for signal enhancement will be proposed using the spherical harmonic domain, given its effectiveness in dealing with signals captured by a spherical microphone array.

The goal of signal enhancement is to isolate a desired source from one or more interfering sources. Typically the spherical microphone array records a sound scene where a mixture of signals with different spatial characteristic is present, as exemplified in Figure 2.1 (from [1]).

Signal enhancement tecniques in spherical harmonic domain can be

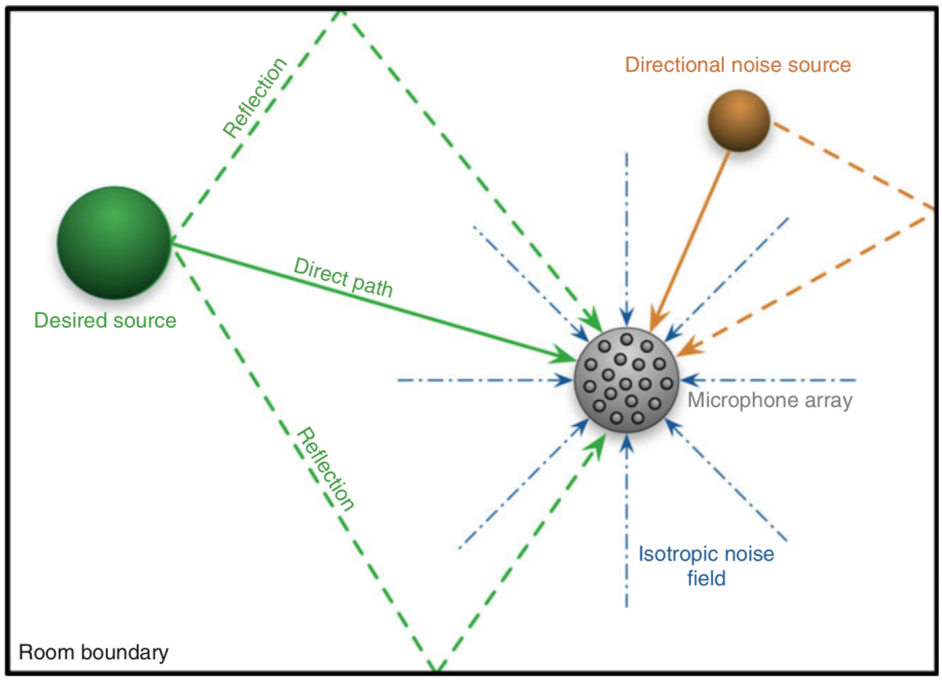

Figure 2.1: Illustration of a typical scenario in microphone array signal processing [1]

roughly divided in three categories: signal-indipendent beamformers, signal-dependent beamformers and informed beamformers.

The first class of noise reduction tecniques includes signal-indipendent beamformers: the filter weights are derived considering only the DOA of the desired source while they do not depend on the desired signal statistics. In practice this is obtained by setting a distortionless constraint on the DOA and optimizing with respect to some performance metrics. Classical signal-indipendent beamformers formulated in the spherical harmonic domain are Maximum White Noise Gain (WNG) and Maximum Directivity: they both optimize two different performance metrics, Directivity and WNG, that represent respectively spatial selectivity and noise amplification. A third alternative is Multiply Constrained Beamformer which tries to minimize the side lobes of the spatial response of the beamformer. All these beamformers are formulated for both farfield and nearfield conditions in [1]. These spatial filters are easy to implement but the major inconvenient is the necessary tradeoff between spatial selectivity and robustness. If background or diffuse noise is prominent then the ability of suppressing the noise is reduced.

The second class of signal enhancement is that of signal-dependent beamformers. These spatial filters adaptively try to achieve both selectivity and noise reduction exploiting an estimate of the statistics of both noise and desired signals. Often the performance measures, like SNR or MSE, in this case are formulated for frequency subbands of the signals. Note that, since this spatial filters are designed for speech processing,

the 0-th order spherical harmonic coefficients of the target source is considered the desired signal, that corresponds to the signal captured by an omnidirectional microphone positioned at the centre of the sphere. The most general signal-dependent beamformer is the maximum SNR Filter, that maximizes the subband output SNR. Other examples are the Wiener filter, which is designed minimizing the MSE respect to the desired signal, and the Minimum Variance Distortionless Response (MVDR) also called Capon beamformer, that maximizes the noise reduction factor while imposing a distortionless constraint ont the desired signal. Using a tuning parameter, a cost function that combines both noise reduction and speech distortion can be defined. Consequently the tradeoff beamformer is designed: given the tuning parameter $\in 0, 1$, when the tuning parameter is equal to 0 corresponds to a Wiener filter, when is equal to 0 corresponds to a MVDR filter. Another used beamformer is the Linearly Constrained Minimum Variance (LCMV) filter, which is a generalization of the MVDR beamformer where distortionless constraint can be imposed linearly on multiple directions. A detailed derivation of these beamformers can be found in [1].

Signal-dependent beamformers can be further improved incorporating additional informations, like DOA of the target source or signal-to-diffuse ratio: this allows the so-called informed spatial filters to adapt to changing acoustic conditions or moving sources. An example of informed beamformer is presented in [6], where a signal-dependent spatial filter incorporates instantaneous narrowband estimations of the DOA. In practice the noise and desired power spectral density, necessary for the filter weights computation, are obtained iteratively using as smoothing factor the combination of an a-posteriori multichannel Speech Presence Probability with a DOA-based probability. A very similar approach is proposed in [7] and [8], except for the choice of the beamformer.

In the following sections first a general design for spherical harmonic domain processing will be shown, differentiating among real and complex domain. Then we will present in detail a general formulation of spatial filtering and an example of signal-indipendent beamformer. Finally a method for acoustic parameters estimation, specifically DOA estimation, will be addressed.

2.1.1 Pre-processing

In this section we are going to present a necessary pre-processing stage common to both beamformers and acoustic parameters estimation. In Section 1.2.2 we have defined the classic SHT using the equations (1.28) and (1.29), also called complex SHT, and the real SHT using (1.32) and (1.33). In our processing chain we are going to combine these two definition of SHT with the Short Time Fourier Transform (STFT): by doing so, the processing phase will deal with a narrowband version of the signal.

The STFT of a discrete time signal $x(t)$, where t is the discrete time index, is defined as

$$X(l,\nu) = \sum_{t=0}^{K} x(t+lR)w(t)e^{-i\frac{2\pi}{K}\nu t} \tag{2.1}$$

where $w(t)$ is the analysis window of length M, R is the hop size between a time window and the successive and l is the window index. Note that ν, the frequency index, is such that $0 \leq \nu \leq K-1$ and is related to continuous frequency $f = \frac{\nu}{K} f_s$, where f_s is the frequency sampling. The inverse short time Fourier transform is defined as

$$x(t) = \sum_{l} \sum_{\nu=0}^{K-1} X(l,\nu)\hat{w}(t-lR)e^{i\frac{2\pi}{K}\nu(t-lR)} \tag{2.2}$$

where $\hat{w}(t)$ is the synthesis window.

We assume the soundfield to be sampled by a spherical microphone array of radius r with Q microphones located at positions $\mathbf{r_q} = (\theta_q, \phi_q, r)$ with $q = 1, ..., Q$ and acquiring for each position the signal $x(t, \mathbf{r_q})$. By changing the order of application of STFT and SHT we define two possible processing chains, as depicted in Figure 2.2 [1]:

Complex spherical harmonic domain pre-processing: to the q-th microphone signal $x(t, \mathbf{r_q})$ is applied STFT obtaining a signal $X(l, \nu, \mathbf{r_q})$. After that the complex SHT is applied obtaining the coefficients $X_{nm}(l, \nu)$: these coefficents are usually called eigenbeams. The overall structure is described in the left part of Figure 2.2.

Real spherical harmonic domain pre-processing: here, SHT is applied to microphone signals before computing the STFT. The main difference is that, since microphone signals are real, real SHT as defined in 1.2.2 is first applied to the q-th microphone signal $x(t, \mathbf{r_q})$ obtaining $x_{nm}(t)$ and then STFT is applied, obtaining $X_{nm}(l, \nu)$. The overall structure is described in the right part of Figure 2.2.

In general all the methods proposed are valid for both complex and real pre-processing stage. The only difference will be that if spherical harmonic functions are used in any expression then they will be the complex or the real one, depending on which one has been choosed for the pre-processing

2.1.2 Spherical Harmonic Domain Beamforming

For the following derivations it is useful to elaborate the notation. Considering the sound pressure after the pre-processing stage described in Section 2.1.1 $P_{nm}(l, \nu)$, we substitute the frequency bin index ν with the

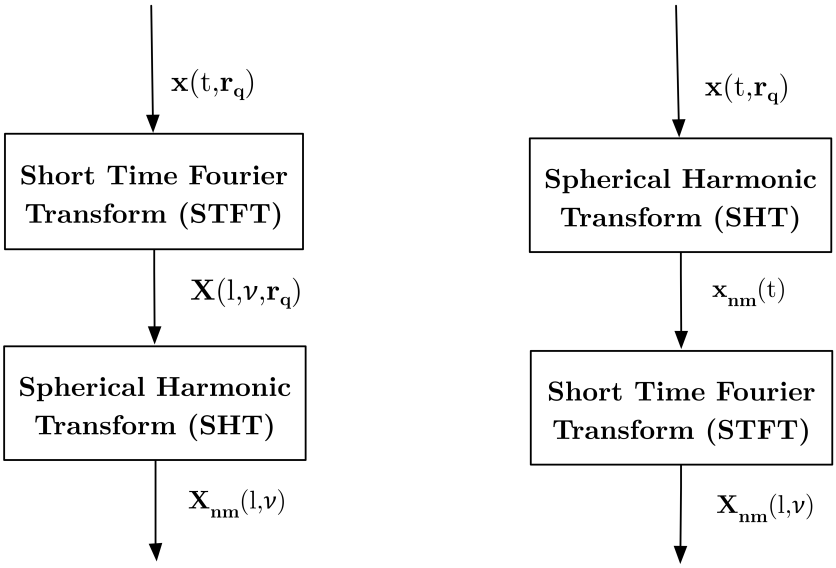

Figure 2.2: Complex (left) and real (right) spherical microphone array processing chain

wave number $k = \frac{2\pi\nu f_s}{cK}$ hence obtaining $P_{nm}(l,k)$. To further simplify the notation the time index is omitted obtaining as final notation $p_{nm}(k)$. A beamformer in the spherical harmonic domain is defined by a set of weights $w^*_{nm}(k)$ which are combined linearly with the spherical harmonic coefficients of sound pressure. The output of the beamformer is therefore

$$z(k) = \sum_{n=-\infty}^{\infty} \sum_{m=-n}^{n} w^*_{nm}(k) p_{nm}(k). \tag{2.3}$$

Truncating the summation at N then it can be reformulated as:

$$z(k) = \mathbf{w^H}(k)\mathbf{p}(k) \tag{2.4}$$

where

$$\mathbf{w}(k) = [w_{00}(k), w_{1(-1)}(k), w_{10}(k), ..., w_{NN}(k)]^T \tag{2.5}$$

and

$$\mathbf{p}(k) = [p_{00}(k), p_{1(-1)}(k), p_{10}(k), ..., p_{NN}(k)]^T. \tag{2.6}$$

Given these framework, now it is possible to analyze in detail the formulations of some interesting beamformers: we are going to present

the maximum directivity beamformer, MVDR beamformer and LCMV beamformer.

Maximum Directivity beamformer The maximum directivity beamformer aims at maximizing the directivity while imposing the distortionless constraint in the steering direction. It is a signal indipendent beamformer and this definition assumes far field condition.

Directivity is a measure of how much the beamformer is spatially selective, i.e. it evaluates how much is good in extracting only the sound field received from a specific direction: it is defined as the ratio of the power of the beamformer output coming from the steering direction Ω_s and the power of the beamformer output averaged on all the directions.

The weights for the maximum directivity beamformer are estimated through a maximization problem: [1]

$$\max_{\mathbf{w}(k)} D(k) \quad \text{subject to} \quad \mathbf{w}(k)\mathbf{B}(k)\mathbf{y}^*(\Omega_s) = 1 \tag{2.7}$$

where Ω_s is the steering direction, $D(k)$ is the directivity and

$$\mathbf{B}(k) = \text{diag}([b_0(k), b_1(k), b_1(k), b_1(k), b_2(k), ..., b_N(k)]) \tag{2.8}$$

which is built by repeating each $b_n(k)$, the mode strength coefficients defined in (1.60), for $(n+1)^2$ times on the diagonal and $\mathbf{y}^*(\Omega_s) = [Y_0^0(\Omega_s), Y_1^{-1}(\Omega_s), Y_1^0(\Omega_s), Y_1^1(\Omega_s), ..., Y_N^N(\Omega_s)]^T$ is the vector of the spherical harmonic functions computed for the steering direction.

This can be also expressed as

$$\min_{\mathbf{w}(k)} ||\mathbf{B}(k)\mathbf{w}^*(k)||^2 \quad \text{subject to} \quad \mathbf{w}(k)\mathbf{B}(k)\mathbf{y}^*(\Omega_s) = 1. \tag{2.9}$$

The solution, whose derivation is detailed in [1], is

$$\mathbf{w}_{\text{maxD}}(k) = \frac{[\mathbf{B}^*(k)]^{-1}\mathbf{y}^*(\Omega_s)}{||\mathbf{y}(\Omega_s)||^2}. \tag{2.10}$$

or, using the Unsold's theorem [9]

$$\mathbf{w}_{\text{maxD}}(k) = \frac{4\pi}{(N+1)^2}[\mathbf{B}^*(k)]^{-1}\mathbf{y}^*(\Omega_s). \tag{2.11}$$

Note that the Maximum Directivity is very similar, apart from the

multiplicative constant, to the plane-wave decomposition beamformer, also called regular beamformer described in detail in [10].

It can be shown that the beam pattern of the maximum directivity beamformer approximates asimptotically a delta function $\delta(\Omega_s)$ as N, the maximum order for the SHT, grows.

MVDR beamformer The MVDR beamformer aims at minimizing the residual power noise while imposing a distortionless constraint on the desired signal. It was first proposed in [11]. The signal model assumes that $p_{nm}(k) = x_{nm}(k) + v_{nm}(k)$ where $x_{nm}(k)$ are the spherical harmonic coeffients of the received source signal and $v_{nm}(k)$ are the spherical harmonic coefficients of the noise source: the desired signal is $x_{00}(k)$, which corresponds to the signal that would be captured by a omindirectional microphone placed at the center of the spherical microphone array. The relationship between $x_{00}(k)$ and $x_{nm}(k)$ is expressed by a relative transfer function (RTF) $\mathbf{d}(k)$, meaning that

$$\mathbf{x}(k) = \mathbf{d}(k)x_{00}(k) \tag{2.12}$$

where $\mathbf{x}(k) = [x_{00}(k), x_{1(-1)}(k), x_{10}(k), ..., x_{NN}(k)]^T$ and $\mathbf{d}(k) = [d_{00}(k), d_{1(-1)}(k), d_{10}(k), ..., d_{NN}(k)]^T$.

For this application is necessary first to perform mode strength compensation for eliminating the dependences on the array configuration, in practice the coefficients $p_{nm}(k)$ will be divided by the correspondent mode strength coefficients $b_n(k)$, i.e. $\tilde{p}_{nm}(k) = p_{nm}(k)/b_n(k)$. The compensated coefficients vectors of the signals will be indicated with $\tilde{\mathbf{x}}(k)$, $\tilde{\mathbf{v}}(k)$ and $\tilde{\mathbf{p}}(k)$.

The MVDR beamformer estimation problem can be formulated in two ways: the first one is

$$\min_{\mathbf{w}(k)} \phi_{\tilde{v}_r(k)} \quad \text{subject to} \quad v_{sd}[\mathbf{w}(k)] = 0 \tag{2.13}$$

where $\phi_{\tilde{v}_r(k)}$ is the power of the residual noise at the beamformer output $\mathbf{v}_r(k)$ and $v_{sd}[\mathbf{w}(k)]$ is the speech distortion index, computed as the normalized MSE between the filtered signal and the desired signal.

The second equivalent expression of the problem is

$$\min_{\mathbf{w}(k)} \mathbf{w}^H(k)\mathbf{\Phi}_{\tilde{\mathbf{v}}(k)}\mathbf{w}(k) \quad \text{subject to} \quad \mathbf{w}^H(k)\mathbf{d}(k) = 1 \tag{2.14}$$

where $\mathbf{\Phi}_{\tilde{\mathbf{v}}(k)}$ is the power spectral density of the noise.

The weights obtained from the solution of the minimization prob-

lem are

$$\mathbf{w}(k) = \frac{\mathbf{\Phi}^{-1}_{\tilde{\mathbf{v}}(k)}\mathbf{d}(k)}{\mathbf{d}^H(k)\mathbf{\Phi}^{-1}_{\tilde{\mathbf{v}}(k)}\mathbf{d}(k)}. \tag{2.15}$$

As evident from (2.15) the MVDR beamformer relies on an estimate of both the RTF $\mathbf{d}(k)$ and the power spectral density of the noise $\mathbf{\Phi}_{\tilde{\mathbf{v}}(k)}$. For estimating the RTF many methods have been proposed, since the performance of the MVDR beamformer is really sensitive to this approximation. Some of them can be found in [1]. An alternative distortionless formulation has been proposed called minimum power distortionless response (MPDR) beamformer that minimizes the total power at the beamformer's output, avoiding the estimation of $\mathbf{\Phi}_{\tilde{\mathbf{v}}(k)}$: it can be shown that for MVDR and MDPR beamformers are equivalent if the RTFs are exact and equal.

Moreover it can be proved that in case of spatially diffuse noise, in anechoic enviroment and plane wave incidence, the MVDR beamformer reduces to the maximum directivity beamformer.

LCMV beamformer The LCMV filter is a generalization of the MVDR beamformer that seeks to suppress a noise signal while mantaining multiple distortionless contraints. In this scenario I directional sources are present instead of one, as in the MVDR beamformer case. The desired signals is represented by $\tilde{\mathbf{x}}_{00}(k)$ while the RTFs between the received desired sources signals and the desired signals are defined as $\mathbf{D}(k) = [\mathbf{d}_1(k)|\mathbf{d}_2(k)|...|\mathbf{d}_I(k)]$. The underlying problem is the formulated as [12]

$$\min_{\mathbf{w}(k)} \mathbf{w}^H(k)\mathbf{\Phi}_{\tilde{\mathbf{v}}(k)}\mathbf{w}(k) \quad \text{subject to} \quad \mathbf{w}^H(k)\mathbf{D}(k) = \mathbf{q}^T(k) \tag{2.16}$$

where $\mathbf{q}(k) = [Q^{(1)}(k), Q^{(2)}, ..., Q^{(I)}(k)]$ represents the desired responses, meaning $Q^{(i)}(k) = 0$ if the i-th source has to be suppresed, $Q^{(i)}(k) = 1$ if the i-the source has to be preserved. The solution is the known LCMV beamformer, which is

$$\mathbf{w}_{LCMV}(k) = \mathbf{\Phi}^{-1}_{\tilde{\mathbf{v}}}(k)\mathbf{D}(k)[\mathbf{D}^H(k)\mathbf{\Phi}^{-1}_{\tilde{\mathbf{v}}}(k)\mathbf{D}(k)]^{-1}\mathbf{q}(k). \tag{2.17}$$

For this solution the columns of the matrix $\mathbf{D}(k)$ must be linearly indipendent and $\mathbf{\Phi}^{-1}_{\tilde{\mathbf{v}}}(k)$ must be full rank. Alike the MVDR beamformer, it is possible to define an alternative beamformer which requires to minimize the power of the output, leading to the linearly constrained minimum power (LCMP) beamformer. Again an accurate estimation of $\mathbf{D}(k)$ is necessary and if the exact RTFs are available then LCMV and LCMP are equivalent.

2.1.3 Acoustic Parameter Estimation

Spherical microphone array processing includes the estimation of acoustic parameters of the soundfield, for example the DOAs of the sources that are present in the soundfield. This parameter is useful for example in case of signal-indipendent beamformers, where the steering direction of the beamformer coincides with the DOA of the source we desire to extract.

A first approach to solve this problem is to determine the output's power of a beamformer on a grid of values of Ω, producing the so-called steered response power (SRP). Any beamformer can be used for this procedure, for example in [13] the computation of the SRP is carried on using the MVDR beamformer. The values that maximize locally the SRP correspond to the DOAs of the sources. This method assumes only one source is present for each time-frequency bin: this assumption can be considered too strong in practice and moreover this approach leads to high computational cost.

Another class of DOA estimator includes the subspace-based methods, where the vector space of the covariance matrix of the noisy signal $\mathbf{\Phi}_{\tilde{\mathbf{p}}}(k)$ is decomposed in two orthogonal subspaces, the noise subspace and the signal subspace: from these subspaces a pseudospectrum and consequently an estimate of the DOA of a desired source can be extracted. A detailed formulation of a subspace-based method, the Multiple Signal Classification (MUSIC) DOA estimation, will be given after presenting a useful signal model.

The soundfield is assumed to be composed of I plane-wave sources incident on the spherical microphone array from directions $\mathbf{\Omega} = [\Omega_1, \Omega_2, ...\Omega_I]^T$ and a noise signal.

Consider the sound pressure after the pre-processing stage described in Section 2.1.1 $P_{nm}(l, \nu)$: similarly to Section 2.1.2, we substitute the frequency bin index ν with the wave number $k = \frac{2\pi\nu f_s}{cK}$ obtaining $P_{nm}(l, k)$. To further simplify the notation the time index is omitted obtaining as final notation $p_{nm}(k)$.

Hence the overall soundfield can be described in terms of spherical harmonic domain coefficients as

$$p_{nm}(k) = \sum_{i=1}^{I} x_{nm}(k, \Omega_i) s_i(k) + v_{nm}(k) \tag{2.18}$$

where $p_{nm}(k)$ is the coefficient of order n degree m for the noisy signal acquired from the spherical microphone array, $v_{nm}(k)$ for the noise signal, $x_{nm}(k, \Omega_i)$ is the spherical harmonic coefficient obtained for a unit-amplitude plane wave coming from the source direction Ω_i and $s_i(k)$ is the amplitude of i-th the plane wave.

Since these coefficients are dependent on the mode strenght compensation coefficients $b_n(k)$ defined in (1.60), hence from the array configuration, it's necessary to cancel this dependency perfoming mode strenght

compensation, i.e. $\tilde{p}_{nm}(k) = p_{nm}(k)/b_n(k)$.

The mode strength compensated coefficients can be grouped in vectors obtaining $\tilde{\mathbf{p}}(k)$ and $\tilde{\mathbf{v}}(k)$. Moreover we define $\mathbf{s}(k)$ as a vector where each element i corresponds to the amplitude of the i-th plane wave.

Note that for a unit amplitude plane wave incident from direction Ω_i the coefficients of the sound pressure are related to the mode strenght coefficients by the relationship $a_{nm}(k) = b_n(k)[Y_n^m(\Omega_i)]^*$. Then, after mode strengt coefficients, it will result that $\tilde{x}_{nm}(k) = [Y_n^m(\Omega_i)]^*$: furthermore since this term is frequency independent we can change the notation in $\tilde{\mathbf{x}}(\mathbf{\Omega_i})$. The manifold matrix is then defined as

$$\begin{aligned} \tilde{\mathbf{X}}((\Omega)) &= [\tilde{\mathbf{x}}(\Omega_1)|\tilde{\mathbf{x}}(\Omega_2)|...|\tilde{\mathbf{x}}(\Omega_I)] \\ \tilde{\mathbf{x}}(\Omega_i) &= [\tilde{x}_{00}(\Omega_i), \tilde{x}_{1(-1)}(\Omega_i), \tilde{x}_{10}(\Omega_i), ..., \tilde{x}_{NN}(\Omega_i)]^T \\ \mathbf{\Omega} &= [\Omega_1, \Omega_2, ..., \Omega_I]^T \end{aligned} \tag{2.19}$$

The final model takes form of

$$\tilde{\mathbf{p}}(k) = \tilde{\mathbf{X}}(\mathbf{\Omega})\mathbf{s}(k) + \tilde{\mathbf{v}}(k). \tag{2.20}$$

MUSIC DOA estimation The MUSIC subspace-based DOA estimation method exploits specifically the orthogonality between the manifold vectors $\tilde{\mathbf{x}}(\mathbf{\Omega_i})$ and the noise subspace $\mathbf{U}_v(k)$, meaning:

$$\mathbf{U}_v(k)^H \tilde{\mathbf{x}}(\mathbf{\Omega_i}) = 0 \quad i = 1, ..., I. \tag{2.21}$$

The noise subspace is extracted perfoming first eigenvalue decomposition of the covariance matrix of the noisy signal $\mathbf{\Phi}_{\tilde{\mathbf{p}}}(k)$. Then eigenvalues are arranged in decreasing order and only the last $(N+1)^2 - I$ eigenvalues are selected: the corresponding eigenvectors are the columns of $\mathbf{U}_v(k)$.

The MUSIC algorithm is performed in 3 steps:

1. an estimate of the covariance matrix of the noisy vector of spherical harmonic coefficients is extracted using a time-average process obtaining $\mathbf{\Phi}_{\tilde{\mathbf{p}}}(k)$;
2. the eigenevalues decomposition of $\mathbf{\Phi}_{\tilde{\mathbf{p}}}(k)$ is performed and the noise subspace $\mathbf{U}_v(k)$ is assembled;
3. the pseudospectrum defined as

$$\mathrm{P}_{MUSIC}(k, \Omega) = \frac{1}{\tilde{\mathbf{x}}(\Omega)^H \mathbf{U}_v(k) \mathbf{U}_v(k)^H \tilde{\mathbf{x}}(\Omega)} \tag{2.22}$$

is computed for a grid of values of Ω; the I higher peaks of $\mathrm{P}_{MUSIC}(k, \Omega)$ will correspond to the DOA of the desired source for the k-th frequency subband.

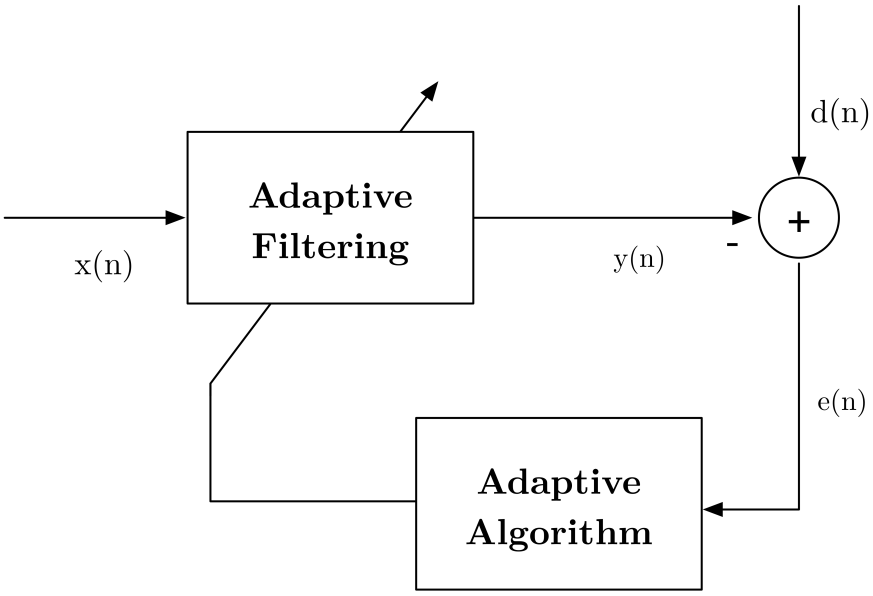

Figure 2.3: General scheme for adaptive filtering

It's interesting to highlight that peak finding in two dimensions, required from all the DOA estimation methods, is not of easy solution. A possible approach is to use a Von Mises peak-finding proposed in [14]: the azimuth and inclination angles are modeled as Von Mises distributions. The two dimensional map (which for example in the MUSIC is the pseudospectrum) is scanned looking for an absolute maxima. When a peak is found, an inverse mask built starting from the Von Mises distribution centered in the position $(\theta_{MAX}, \phi_{MAX})$ is applied on the map, flattening it around those point. A new peak is found repeating this procedure on the modified map.

2.2 Adaptive Filtering

In signal processing adaptive filtering has been widely used for many applications like echo and noise cancellation, background noise removal or signal prediction. The design of a digital filter with fixed coefficients assumes the knowledge of a complete characterization of the input signal and of the reference signal, that is necessary for the definition of the performance metric. If these or other specifications are not available or time changing then is useful to design a filter that adapts its transfer function trying to meet a performance requirement online: this is the main motivation beyond the design of adaptive filters. A general scheme for adaptive filtering is presented in Figure 2.3

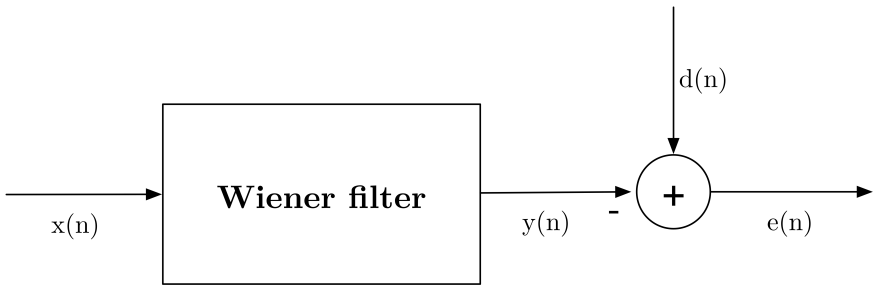

Figure 2.4: General scheme for Wiener filter

In the following sections, starting with the classic formulation of Wiener filter, we will present first the Least Mean Squares (LMS) filter and then the frequency-domain adaptive filtering tecniques.

2.2.1 Classical Wiener filter

The Wiener filter was proposed in 1949 by Norbert Wiener and its main purpose is to produce an estimate of a stationary signal of interest by applying a Linear Time-invariant (LTI) filter to a noisy observed signal: in practice the task is performed through the minimization of a MSE-based cost function. In Figure 2.4 the block diagram for Wiener filter is presented.

In this section we are going to introduce the Finite Impulse Response (FIR) formulation of Wiener filter. The signals $x(n)$, $y(n)$, $d(n)$ and $e(n)$ are assumed to be complex-valued, stationary, random and zero-mean. Input signal and filter coefficients are defined as

$$\begin{aligned} \mathbf{x}(n) &= [x(n), x(n-1), ..., x(n-M+1)]^T \\ \mathbf{w} &= [w(0), w(1), ..., w(M-1)]^T \end{aligned} \tag{2.23}$$

where M is the number of coefficients of the filter. The output $y(n)$ is given by

$$y(n) = \sum_{k=0}^{M-1} w^*(k)x(n-k) \tag{2.24}$$

and the error signal is defined as

$$e(n) = d(n) - y(n). \tag{2.25}$$

To compute the coefficients of the filter it is necessary to solve

$$\mathbf{w_0} = \arg\min_{\mathbf{w_0}} J \tag{2.26}$$

where $\mathbf{w_0}$ is the vector of optimum coefficients and J is the Minumum MSE cost fuction, defined as

$$J = E\{e(n)e^*(n)\} = E\{|e(n)^2|\} \tag{2.27}$$

where $E\{\cdot\}$ is the expectation operator. Substituting the definition of $e(n)$ we obtain

$$\begin{aligned} J &= E\{e(n)e^*(n)\} \\ &= E\{|d(n)|^2\} - \sum_{k=0}^{M-1} w^*(k)E\{x(n-k)d^*(n)\} - \sum_{k=0}^{M-1} w(k)E\{x^*(n-l)d(n)\} \\ &+ \sum_{k=0}^{M-1}\sum_{i=0}^{M-1} w^*(k)w(i)E\{x^*(n-k)x(n-1)\} \\ &= \sigma_d^2 - \sum_{k=0}^{M-1} w^*(k)p(-k) - \sum_{k=0}^{M-1} w(k)p^*(-k) + \sum_{k=0}^{M-1}\sum_{i=0}^{M-1} w^*(k)w(i)r(i-k) \end{aligned} \tag{2.28}$$

where $\sigma_d = E\{|d(n)|^2\}$ is the variance of $d(n)$, $p(-k)$ is the crosscorrelation between the input and the desired signal and $r(i-k)$ is the autocorrelation function of the input. The cost function J is convex and positive, and presents a single minimum that can be found by setting its gradient to zero. This condition, computing the gradient starting from (2.28), is expressed as

$$\nabla J = -2\mathbf{p} + 2\mathbf{R}\mathbf{w} = 0 \tag{2.29}$$

where $\mathbf{p}$ is the crosscorrelation vector defined as

$$\mathbf{p} = E\{\mathbf{x}(n)\mathbf{d}^*(n)\} = [p(0), p(-1), ..., p(1-M)] \tag{2.30}$$

and $\mathbf{R}$ is the autocorrelation matrix defined as

$$\begin{aligned} \mathbf{R} &= E\{\mathbf{x}(n)\mathbf{x}^H(n)\} \\ &= \begin{bmatrix} r(0) & r(1) & ... & r(M-1) \\ r^*(1) & r(0) & ... & r(M-2) \\ \vdots & \vdots & ... & \vdots \\ r^*(M-1) & r^*(M-2) & ... & r(0) \end{bmatrix}. \end{aligned} \tag{2.31}$$

Therefore the so called Wiener-Hopf equations can be formulated in matrix form as:

$$\mathbf{R}\mathbf{w_0} = \mathbf{p}. \tag{2.32}$$

The solution is obtained by matrix inversion as

$$\mathbf{w_0} = \mathbf{R}^{-1}\mathbf{p}. \tag{2.33}$$

This solution is called the Wiener solution.

2.2.2 Least Mean Squares

The LMS algorithm is a widely used tecnique in adaptive filtering thanks to its low computational complexity and its robustness. LMS attempts to estimate the Wiener optimum filter by updating the filter coefficients in an iterative fashion following the steepest-descent approach. In this section hence we will present first the steepest-descent approach, then the LMS filtering and finally a variation of LMS called Block LMS

2.2.2.1 Steepest Descent

Solving the Wiener-Hopf equations presented in (2.32) requires a matrix inversion that can be computationally complex. A possible solution to this problem is to use an adaptive tecnique for the computation of the weights, which is the steepest-descent algorithm. Starting from an arbitrary value for the filter weights vector, the gradient of the cost function $\nabla J(n)$ is computed and the weights are updated iteratively following the steepest descent of the gradient. Since steepest-descent method provides a time-varying filter weights vector we will stress the time dependence in the notation adopted which is

$$\begin{aligned} \mathbf{x}(n) &= [x(n), x(n-1), ..., x(n-M+1)]^T \\ \mathbf{w}(n) &= [w_0(n), w_1(n), ..., w_{M-1}(n)]^T \\ J(n) &= \sigma_d^2 - \mathbf{w}^H(n)\mathbf{p} - \mathbf{p}^H\mathbf{w}(n) + \mathbf{w}^H(n)\mathbf{R}\mathbf{w}(n) \end{aligned} \tag{2.34}$$

where the expression for the cost function is obtained from (2.28) assuming the input $\mathbf{x}(n)$ and the desired response $d(n)$ are jointly stationary hence both the cross correlation vector and the autocorrelation matrix are constant. The update equation for the steepest-descent algorithm is

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{1}{2}\mu[-\nabla J(n)] = \mathbf{w}(n) + \mu[\mathbf{p} - \mathbf{R}\mathbf{w}(n)] \tag{2.35}$$

where μ is a parameter positive real-valued, the step size and the gradient expression is the one given in (2.29).

The condition for convergence of steepest-descent algorithm is obtained analyzing the natural modes of the cost fuction: the step size μ must be chosen following

$$0 < \mu < \frac{1}{\lambda_{\max}} \tag{2.36}$$

where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{R}$.

2.2.2.2 Least Mean Squares

The LMS algorithm is very similar to steepest-descent presented in Section 2.2.2.1 except for the computation of the gradient of the cost function. If the gradient of the cost function $\nabla J(n)$ as defined for steepest-descent algorithm is rewritten using the definition of $\mathbf{R}$ and $\mathbf{p}$ we obtain

$$\begin{aligned} \nabla J(n) &= -2\mathbf{p} + 2\mathbf{R}\mathbf{w}(n) = -2E\{\mathbf{x}(n)[d*(n) - \mathbf{x}^H(n)\mathbf{w}(n)]\} \\ &= -2E\{\mathbf{x}(n)[d^*(n) - y^*(n)]\} = -2E\{\mathbf{x}(n)e^*(n)\}. \end{aligned} \tag{2.37}$$

Dealing with expected value would require infinite length signals, that are not available in practice, hence another tecnique for gradient computation is required. In particular LMS considers the computation of the stochastic gradient rather than its exact value, which is

$$\nabla \tilde{J}(n) = -2\mathbf{x}(n)e^*(n). \tag{2.38}$$

Consequently the update equation for LMS algorithm is

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\mathbf{x}(n)e^*(n). \tag{2.39}$$

Note that the choice of computing the stochastic gradient leads to a substantially drop in computationally cost since it can be performed using multiplication instead of correlation.

As for steepest-descent, is possible to derive a condition of convergence for LMS algorithm:

$$0 < \mu < \frac{1}{tr[\mathbf{R}]} \tag{2.40}$$

where $tr\{\cdot\}$ denotes the trace operator. A complete derivation of convergence conditions for both steepest-descent and LMS algorithm can be found in [15].

2.2.2.3 Block Least Mean Squares

As explained, weights in LMS are updated for each pair of values $(x(n), d(n))$ received: alternatively the input can be grouped in blocks of length L and

the weights can be updated for each block. This version of the algorithm is called Block LMS algorithm.

Using k to indicate the block index, the time index n is expressed as

$$n = kL + i \quad i = 0, 1, ..., M-1 \quad k = 0, 1, ... \tag{2.41}$$

hence the notation for the weigths vector and input is

$$\begin{aligned} \mathbf{w}(k) &= [w_0(k), w_1(k), ..., w_{M-1}(k)]^T \\ \mathbf{x}(n) &= [x(n), x(n-1), ..., x(n-M+1)]^T. \end{aligned} \tag{2.42}$$

The filtering equation for the Block LMS algorithms is therefore

$$y(kL + i) = \mathbf{w^H}(k)\mathbf{x}(kL + i) \quad i = 0, 1, ..., M-1. \tag{2.43}$$

Differently from classical LMS, in the adaptation step the stochastic gradient is averaged on each block of L samples, obtaining a more accurate estimation of the exact gradient:

$$\nabla \hat{J}(k) = -\frac{2}{L} \sum_{i=0}^{L-1} \mathbf{x}(kL + i)e^*(kL + i). \tag{2.44}$$

Consequently the update equation is

$$\begin{aligned} \mathbf{w}(k+1) &= \mathbf{w}(k) - \frac{L}{2}\mu\nabla\hat{J}(k) = \\ &= \mathbf{w}(k) + \mu \sum_{i=0}^{L-1} \mathbf{x}(kL + i)e^*(kL + i). \end{aligned} \tag{2.45}$$

The derivation of convergence condition for wide stationary signals is similar to the one for classic LMS and it can be proved that the weights vector converges to the Wiener optimal solution as k approaches to infinity. The main difference is that the step size will have to satisfy a more stringent condition since the upper bound is scaled by a factor L. If the matrix $\mathbf{R}$ has large conditioning number, that intuitively corresponds to large power variation, the Block LMS may converge slower than the classic LMS.

The choice of the block length L is fundamental for the performance of the algorithm, generally $L = M$ is generally the preferred choice.

2.2.3 Frequency-Domain Adaptive Filtering

The frequency-domain adaptive algorithms can represent, under certain conditions, a good alternative to time-domain adaptive algorithms: in fact combining the block-update strategy and the efficiency of frequency-

domain in performing time-domain convolutions, a lower overall complexity can be obtained. A detailed framework for frequency-domain adaptive algorithms was first proposed in [16].

The first phase of frequency-domain adaptive filtering is to pre-process the input signal by performing Discrete Fourier Transform (DFT) or, if possible, its efficient implementation Fast Fourier Transform (FFT). The signals obtained are roughly uncorrelated: this allows to use a time-varying step size or, if the input has variable power hence convergence is slower, to help the algorithm to converge by choosing a smaller step size for the DFT bins where power is higher.

For this algorithm the frequency-domain weights vector and the input matrix are defined as

$$\begin{aligned} \mathbf{W}(k) &= [W_0(k), W_1(k), ..., W_{J-1}(k)]^T \\ \mathbf{X}(k) &= \text{diag}([X_0(k), X_1(k), ..., X_{J-1}]) \end{aligned} \tag{2.46}$$

where J is the length of the DFT depending on the considered FDAF configuration (usually $J = M$). The representation of the input as a matrix allows a simple expression of the output vector:

$$\mathbf{Y}(k) = \mathbf{X}(k)\mathbf{W}(k). \tag{2.47}$$

The update step for frequency domain adaptive algorithms is

$$\mathbf{W}(k+1) = \mathbf{W}(k) + 2\mathbf{G}\boldsymbol{\mu}(k)\mathbf{X}^{\mathbf{H}}(k)\mathbf{E}(k) \tag{2.48}$$

where

- the step size matrix $\boldsymbol{\mu}(k)$ is defined as

$$\boldsymbol{\mu}(k) = \text{diag}([\mu_0(k), \mu_1(k), ..., \mu_{J-1}(k)]). \tag{2.49}$$

  As mentioned above, the step size is time-varying according to the signal power within the considered frequency bin: this is implemented by setting

$$\mu_j(k) = \frac{\mu}{P_j(k)} \tag{2.50}$$

  where μ is a fixed scalar and $P_j(k)$ is an estimate of the signal power in the j-th bin. A possible iterative estimation of $P_j(k)$ is given by

$$P_j(k) = \lambda P_j(k-1) + \alpha |X_j(k)|^2 \tag{2.51}$$

  where $\lambda = 1 - \alpha$ denotes the forgetting factor. The inital value $P_j(0)$ is usually set to a positive initial power estimates of $X_j(k)$.

Note that if signals are stationary then $\boldsymbol{\mu} = \mu\mathbf{I}$ being $\mathbf{I}$ the identity matrix.

- the error vector $\mathbf{E}(k) = [E_0(k), E_1(k), ..., E_{J-1}(k)]^T$, depending on the algorithm, can be computed in the time domain and then frequency-domain transformed or computed directly in the frequency domain as the difference between $\mathbf{Y}(k)$ and the frequency-domain transformation of the desired response $\mathbf{D}(k)$.

- the matrix $\mathbf{G}$ represents the frequency-domain transformation of constraint on the gradient necessary for computing it as a linear correlation. Also the choice of this matrix depends on the specific algorithm.

Many algorithms are defined on this general form, for example changing the tecnique used for implementing linear convolution (or linear correlation) through FFT. As already mentioned, a complete framework for frequency-domain adaptive filtering is presented in [16].

2.3 Beamforming and Adaptive Noise Cancelling

In this section we will summarize the noise cancellation technique presented in [17], which takes advantages from both acoustical beamforming and adaptive filtering.

The goal of the method in [17] is the design of an Adaptive Noise Canceller (ANC) for 3D sound field that aims at preserving the spatiality of the sound field. This is accomplished by combining a beamformer and an adaptive filtering algorithm, specifically the Quaternion LMS.

The starting context is a noisy 3D sound field and the filtering section of the system is derived from the noise cancellation configuration. A scheme of this adaptive filtering application is shown in Figure 2.5. The main idea is to first obtain an estimate of the interfering source and then to subtract it from the noisy signal using it as input for the adaptive filter. The desired signal is a mixture of target and noise sources, the error signal $e(n)$ contains the target source.

In [17] the overall sound-field is acquired with a coincident array of 4 microphones following the B-format, whose signals are usually denoted as X,Y,Z and W, and then the noise signal is extracted using virtual microphones tecniques. A virtual microphone in a specific position is obtained combining and delaying the contributions of the capsules in a sensor array. This in practice consists in applying a sequence of transformations to X,Y,Z and W. The transformations applied in [17] are

- rotation in the direction of the source, defined following the classic Tait-Bryan angles [18];

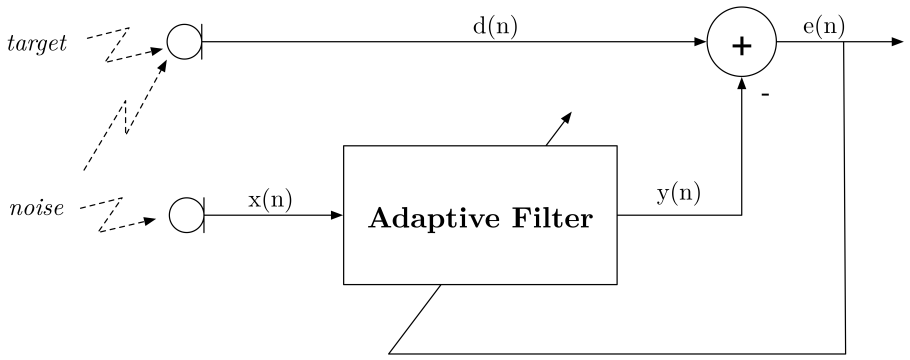

Figure 2.5: General scheme for noise cancellation

- a pattern shaping, achieved through a combination of the microphone signals with a directivity factor D, choosed depending on the desired polar pattern of the virtual microphone;
- a zooming operation, obtained through a matrix transformation based on the distance of the source from the array.

A set of virtual microphones can be created repeating this operation. Specifically in this proposal they are arranged in a Uniform Linear Array (ULA) configuration, and used to extract the noise source.

Then the Quaternion LMS is applied, assigning each B-format microphone signal to a quaternion component, exploiting the correlation among the microphone signals for improving the ANC process.

2.4 Discussion

In the previous sections we have shown the last spherical microphone array signal enhancement tecniques, well-known adaptive filtering tecniques from classic signal processing and a proposal that combines both for 3D soundscape denoising.

The signal spherical microphone array enhancement systems at the state-of-the-art are specifically designed for speech denoising, hence they exploit some specific properties of the speech signals and moreover they do not preserve in the ouput the spatiality of the signal acquired by the spherical microphone array.

The work [17] presented in Section 2.3 presents some similarities to the approach proposed in our work, like the overall purpose or the combined use of microphone array processing tecniques and adaptive filtering. The main differences are the assumption of the knowledge of the noise source's exact position, the lack of a diffuse noise case solution and the use of a

lower order for spherical harmonic processing imposed by the use of a B-format array.

In the next section, after having illustrated both the theoretical and application background, the core of the proposed solution will be presented.

Chapter 3

Proposed Solution

In this chapter we present a signal enhancement system formulated in the spherical harmonic domain. The goal of this work is to suppress interfering components in a noisy soundfield while preserving the spatiality of the recorded acoustic scene. This approach includes the use of two classic signal processing tecniques, beamforming and adaptive filtering. The beamformer performs the actual denoising operation extracting the desired source from the soundfield while the adaptive filtering recovers the original spatiality of the acoustic scene. The input of the system is a soundfield recorded from a rigid spherical microphone array and transformed in the spherical harmonic domain, the ouput is the enhanced sound field represented in the same domain. The target application of our proposal is for example elaboration of 3D recordings acquired in non-ideal settings.

Many techniques for signal enhancing have been proposed in literature using spherical microphone array processing tecniques. Most of these proposals aim at isolating a speech source from a noisy enviroment and typical target applications are human-human and human-machine speech communication systems. The spatiality provides an information redundancy that is exploited for achieving noise suppression. Note that the result of this process is a single channel signal, hence spatial informations are lost in the procedure. On the contrary, the goal of this work is to preserve a 3D representation of the acoustic scene, hence these tecniques singularly are not suitable for our application. Differently, the solution proposed in this thesis combines two stages. First, the signal of the desired source is extracted by means of a beamforming algorithm; then, an adaptive filter restores the original directionality information of the soundfield produced by the desired source.

A scheme of the algorithm is presented in Figure 3.1. The input are the signals acquired from the a spherical microphone array, indicated with $\mathbf{p}(t, \mathbf{r}_q)$, where t is the discrete time index and the Q microphones are located at positions $\mathbf{r_q} = (\theta_q, \phi_q, r)$ with $q = 1, ..., Q$. The first operation performed on the microphone signals is the SHT, which expresses

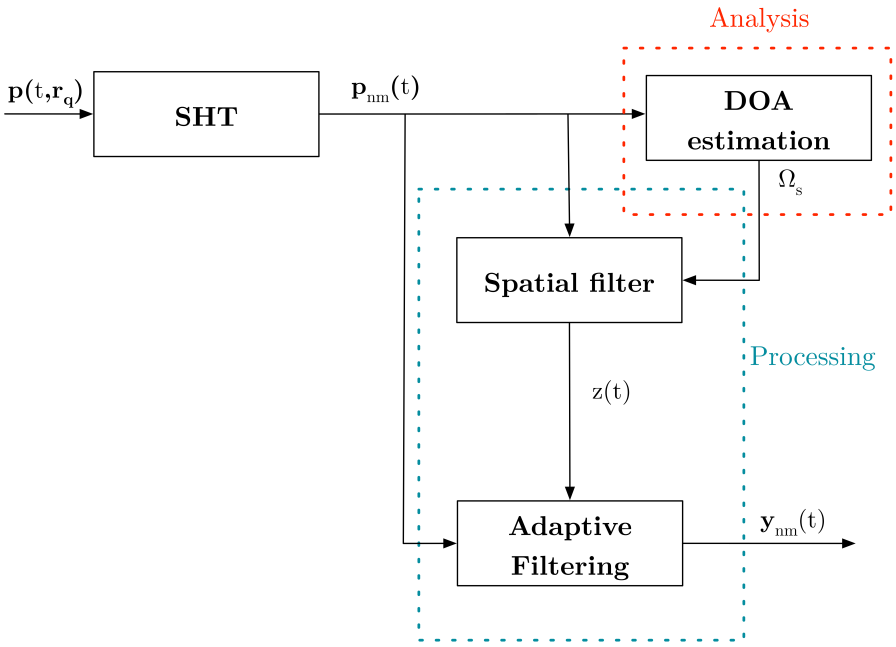

Figure 3.1: General scheme for the denoising system proposed

them in the spherical harmonic domain obtaining the coefficients $p_{nm}(t)$. The spherical harmonic coefficients are then fed to two blocks, the analysis stage and the processing stage. In the first analysis block the goal is to estimate some acoustic properties of the sound field that will act as parameters for the successive operations, specifically we are going to extract the DOA Ω_s of the desired source. In the second block, a processing stage is performed combining two elements: first the desired signal is extracted using a beamformer pointed towards the desired source DOA Ω_s, then the estimate of the source signal $z(t)$ and the noisy spherical harmonic coefficients $p_{nm}(t)$ are used as input and desired signals respectively for an adaptive filter. The output $y_{nm}(t)$ corresponds to the spherical harmonic coefficients of the enhanced signal, i.e. the soundfield as it would be if only the source were present.

3.1 Signals model

In this section we are going to present in detail how the sound field has been defined formally for our work.

Consider a spherical microphone array of radius r and Q microphones at positions $\mathbf{r_q} = (\theta_q, \phi_q, r)$ that acquires a soundfield composed of a desired directional source and a noise. Then the microphone signals so obtained can be expressed as

$$p(t, \mathbf{r_q}) = x(t, \mathbf{r_q}) + v(t, \mathbf{r_q}) \tag{3.1}$$

where t is the discrete time index.

In the spherical harmonic domain (3.1) becomes

$$p_{nm}(t) = x_{nm}(t) + v_{nm}(t). \tag{3.2}$$

In practice these coefficients are obtained by applying discrete SHT as defined in Section 1.3 for a maximum order N to the microphone signals.

The noise components in these types of applications are usually classified depending on their spatial behaviour. For our work we have considered two types of noise:

Directional noise field i.e.the sound field produced by a noise source incident on the spherical microphone array from a specific direction. If far field condition holds, a directional noise can be assumed to behave as a plane wave. It's reasonable to assume that the DOA of the source is different from the noise DOA and that the desired source is closer to the spherical microphone array than the noise source.

Diffuse noise field i.e. the soundfield produced by plane waves incident from all directions with equal probability. The signals received at the microphone are spatially distributed but still correlated among them. Typically diffuse sound field are generated in highly reverberant spaces where the direct components becomes negligible with respect to the reverberant components, hence there is not a privileged direction of propagation.

3.2 Spatial properties of the noise field

In this section we are going to analyze the behaviour in the spherical harmonic domain of the two types of noisy sound fields presented .

The main intuition is that the diffuse noise is distributed among the spherical harmonic coefficients in a different way with respect to the directional one. In particular, since the diffuse signal is incident from all directions, it will be in large part represented by the 0-th order spherical harmonic coefficient, which represents the signal as acquired from an ominidirectional microphone placed at the center of the spherical microphone array.

For validating this intuition we will study the behaviour of the SNR in the spherical harmonic representation. Consider the SNR of a single component $p_{nm}(t)$ defined as the ratio of the power of the desired component and the power of the noise component:

$$\mathrm{SNR}\{p_{nm}\} = \frac{E\{x_{nm}^2(t)\}}{E\{v_{nm}^2(t)\}} \tag{3.3}$$

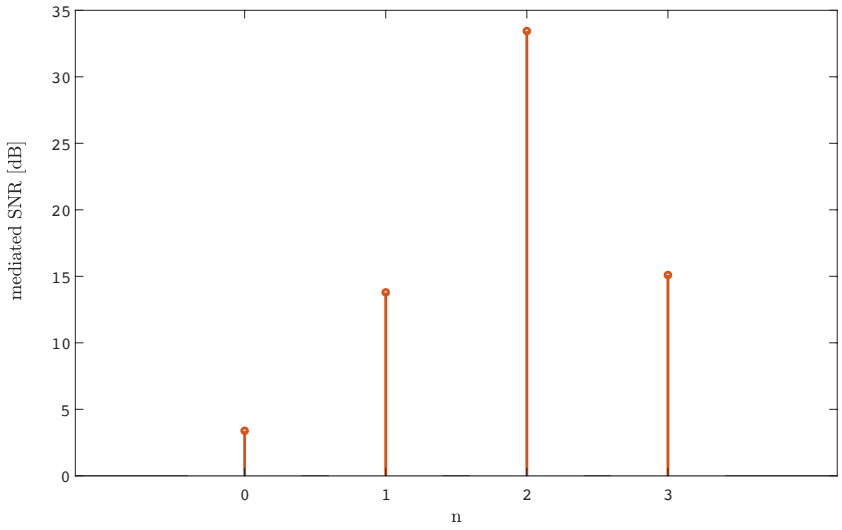

Figure 3.2: SNR in dB averaged over time and over the components of the same order n for diffuse noise and directional source

where $E\{\cdot\}$ is the expected value operator, which will be computed in practice as the sample mean of the signal over a rectangular window.

We have simulated a directional desired source, a diffuse noise and a directional noise in reverberant conditions. The desired source signal is a violin playing, while the noise is the one produced by an air conditioner, similar to a white noise. After that we have synthetized a soundfield combining the desired source first with the directional noise and then with the diffuse noise, setting at the microphone signals the SNR equal to 20 dB. More details about the specific setup used for the simulations are given in Chapter 4. Then we have computed the SNR as defined in (3.3) for the spherical harmonic coefficients of the total soundfield on windows of 1024 samples for both the cases with sampling frequency $f_s = 44100$Hz.

Consider the mean value over time of the computed SNR. If we group together the components having the same order n and we take the average over them we obtain an estimate of the SNR for the order n, which can be indicated as SNR_n. The values SNR_n are shown in 3.2 for the diffuse noise case and 3.3 for the directional noise case. In the first case the difference among the 0-th order SNR and the remaining components is higher with respect to the second case: this is given by the fact that a large part of the noise energy is concentrated in the first component hence the SNR will be lower for the order 0 with respect to the values for orders $n \neq 0$. Hence the assumed spatial behaviour of the diffuse sound field can be considered validated.

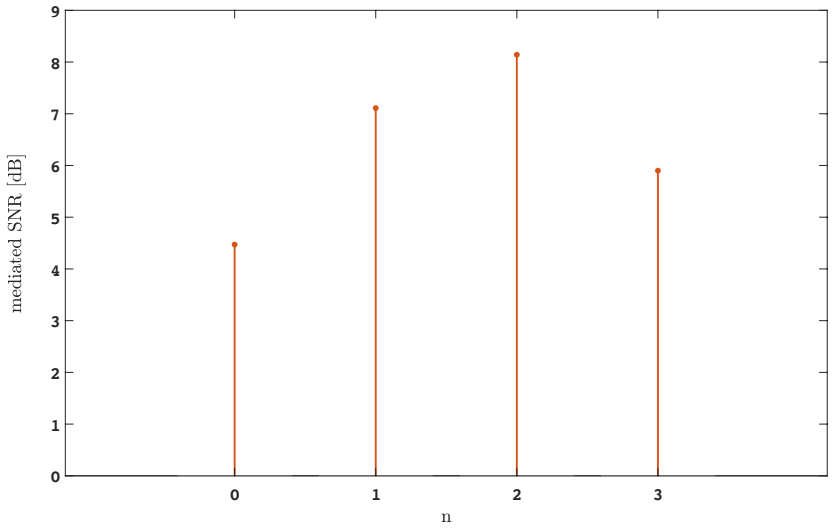

Figure 3.3: SNR in dB averaged over time and over the components of the same order n for directional noise and source

To show that this result is valid also as time changes the SNR for the components having order $n = 0$ and order $n = 1$ are plotted against time in Figures 3.4 and 3.5. Again the SNR for the 0-th order is globally lower with respect to the SNR for the other orders in the diffuse noise case. On the other hand in the directional case the SNR does not have a precise behaviour. More probably the SNR of the mn coefficient depends on the presence or not of energy in a specific direction.

A second osbervation can be done about the spherical harmonic coefficients of the directional noise. As the order grows, the energy of the v_{nm} coefficient decreases, since we are expressing narrower spatial portions of the sound field. The energy of the spherical harmonic coefficients v_{nm} is computed as the expected value of the square of the time signals; in practice is computed as the sample mean over rectangular windows. In Figure 3.6 each bar represents the energy normalized and averaged over all the time windows of the coefficient v_{nm}. Bars of the same color have the same order n. We can observe that the energy for the 0-th order is higher than the energies contained in orders $n > 0$. For completeness the same quantities are plotted for the diffuse noise in Figure 3.7. Here the gap between the 0-th order component and the 1-st order components is even larger with respect to the directional noise case, hence this is a further validation on the considerations on diffuse noise spatial properties.

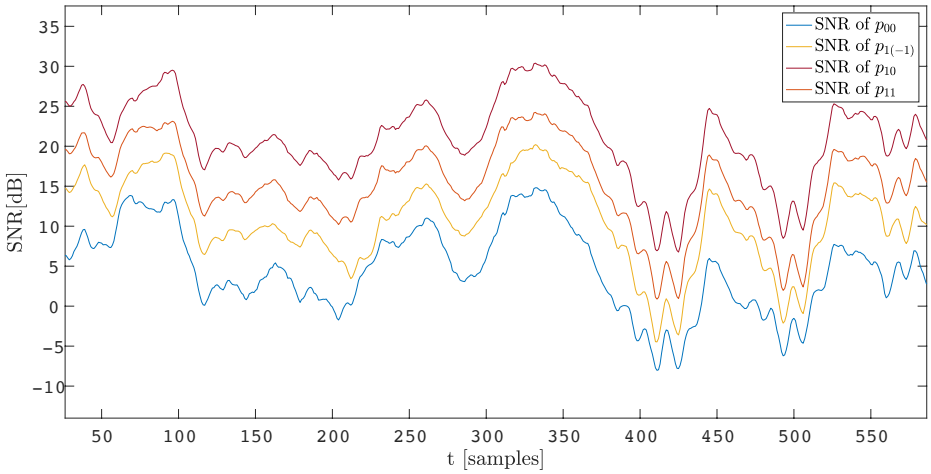

Figure 3.4: SNR of $p_{nm}(t)$ in dB produced by a soundfield with diffuse noise and directional source

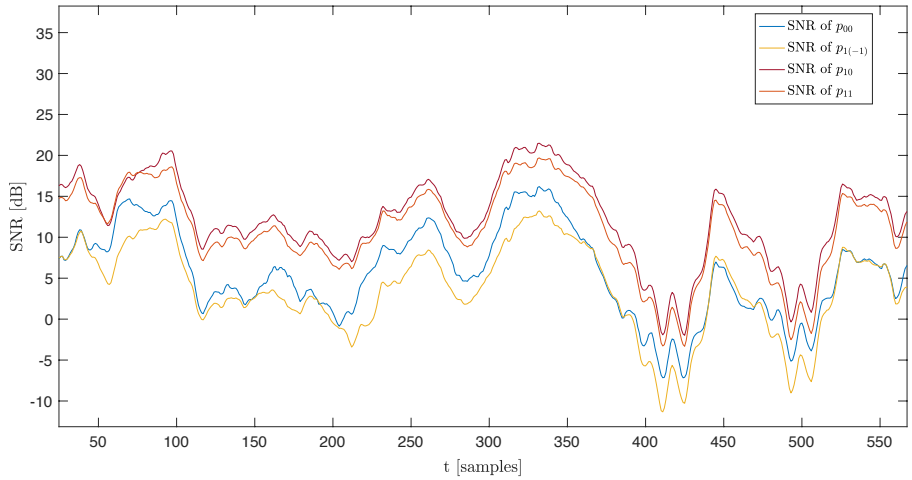

Figure 3.5: SNR of $p_{nm}(t)$ in dB produced by a soundfield with directional noise and source

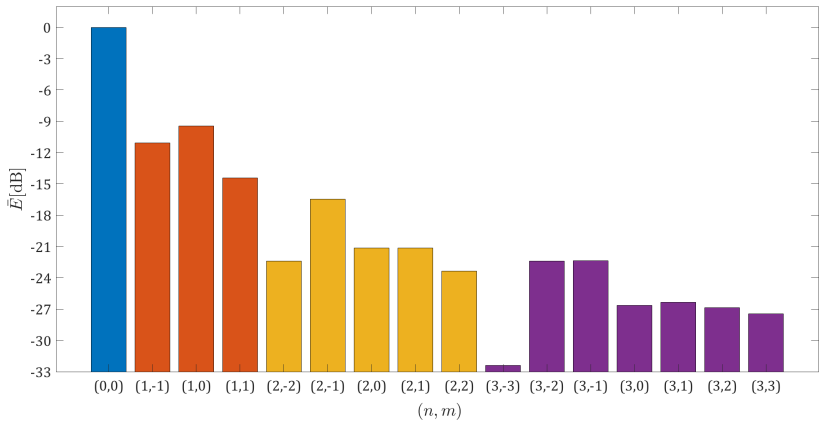

Figure 3.6: Energy of the directional noise in dB averaged over the time windows for each component (n, m) of the spherical harmonic decomposition

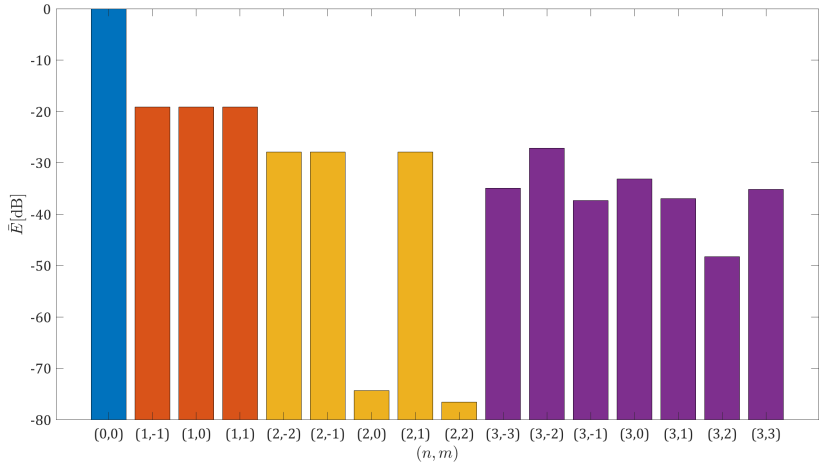

Figure 3.7: Energy of the diffuse noise in dB averaged over the time windows

3.3 Analysis stage

As explained in Figure 3.1 the beamforming operation relies on the knowledge of the DOA of the source. For this reason, we must perform an analysis stage, that consists in the extraction of the DOA of the desired source. Note that this information will be used as a parameter for the beamformer.

We decided to perform DOA estimation in the spherical harmonic domain as well since it is the more natural domain for processing spherical microphone array signals. The microphone array signals will be pre-processed following the scheme illustrated in Section 2.1.1. More specifically we decided to use the real spherical microphone array pre-processing. Note that when the total number of components of the discrete SHT $(N + 1)^2$ is smaller than the number of microphones Q is computational convenient to perform first the spherical harmonic domain transformation and then compute the STFT on a smaller number of signals respect to the complex pre-processing.

In our work DOA estimation is performed using the well-known MUSIC algorithm described in Section 2.1.3 for a single source. The expression for the narrowband pseudospectrum is given by expression (2.22). Note that we must take care of using the correct spherical harmonic functions in the pseudospectrum computation: since real-preprocessing has been choosed, then the real spherical harmonic functions has to be used instead of complex ones. The narrowband pseudospectrum allows to obtain an estimate of the DOA for a given frequency range. To obtain a broadband estimate, necessary for the succesive processing stage, we then compute the geometric mean of the narrowband pseudospectrums obtaining $\hat{P}(\Omega)$ and then perform peak-finding on the averaged pseudospectrum.

Referring to the model presented in Section 3.1, a sound field has been simulated for diffuse noise case and directional desired source (for further details on simulation setup see Chapter 4). In Figure 3.8 the pseudospectrum averaged over the frequency bins is presented. As shown in figure, where the red cross corresponds to the estimated DOA and the red circle indicate the actual DOA, the result is very accurate. In our simulations the error in the estimation is less than 2 degree.

The result of this estimation, which is indicated with Ω_s, will act as a parameter for the subsequent processing stage, specifically for the beamforming step.

If the desired source is assumed to be moving, hence the DOA is changing over time, it is possible to perform the DOA estimation on short overlapping time windows, on which the source is assumed to be fixed in space. As a consequence the beamformer will adapt its weights on each time window. This variation leads to a growth of the computational cost but does not modify the overall architecture.

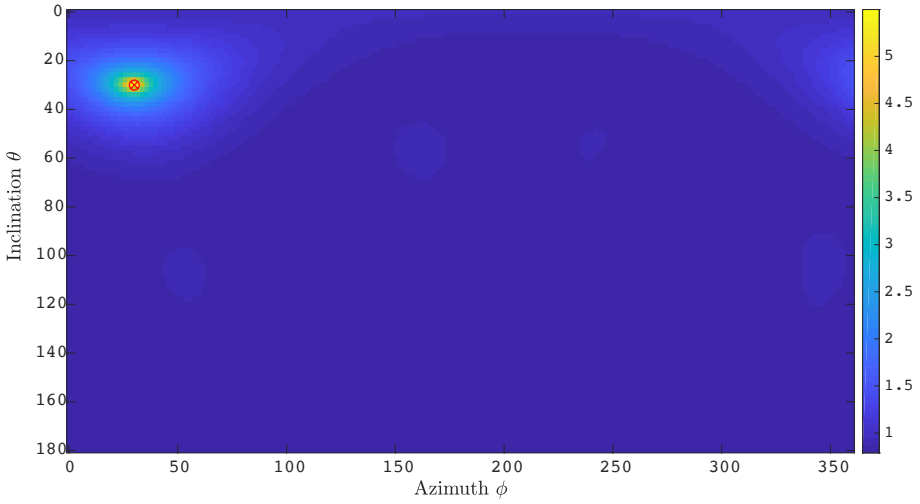

Figure 3.8: Pseudospectrum $\hat{P}(\Omega)$ over a grid of $\Omega = (\phi, \theta)$ for a directional desired source and a diffuse noise

3.4 Processing stage

In this section the actual denoising method is presented. Our objective is to subtract the noise components, generated by a diffuse or directional noise source, from the soundscape acquired with the spherical microphone array, preserving the acoustic spatiality in the output. This is achieved by means of two blocks, as shown in Figure 3.1. A beamformer, formulated and applied in the spherical harmonic domain, extracts the desired source signal; then an adaptive filtering block aims at recovering the spatiality of the soundfield and reconstructing the final result.

Beamformer A beamformer among the several proposed in literature should be choosed for isolating the desired source. For this purpose we have selected a signal-indipendent beamformer, the maximum directivity beamformer, presented in Section 2.1.2. This explains why in the analysis stage DOA estimation is performed: in general signal-indipendent beamformers need as a parameter the steering direction Ω_s, more specifically the maximum directivity beamformer imposes a distortionless constraint in the desired source direction while maximizing the directivity.

As for MUSIC DOA estimation we decided to apply the real spherical microphone array pre-processing showed in Section 2.1.1.

The pre-processing stage produces a time-frequency representation of the spherical harmonic coefficients $\mathbf{P}_{nm}(l, \nu)$ that will be fed to the beamformer as input: fixed a time window l and a frequency bin ν, a set of frequency-dependent weights are linearly combined with the input spherical harmonic coefficients, obtaining the output $Z(l, \nu)$. This result will be further transformed in the discrete time

domain using the inverse STFT, obtaining $z(t)$.

The beamformer weights are computed with the formula given in (2.11) and having as steering direction the DOA estimated from the analysis stage, Ω_s. Note that we must take care of using the right spherical harmonic functions in the weights computation: since real-preprocessing hase been choosed, then the real spherical harmonic functions will be used nstead of the complex ones.

The result will contain the soundfield components along a beam centered in direction Ω_s, therefore the desired source. The width of the beam, i.e. how much selective is the spatial filter in space and which corresponds intuitively to the directivity, depends on the maximum order N used for SHT hence on the number of microphones.

This beamformer has the advantage of being relatively easy to implement and does not require estimates of the signal statistics. As a matter of fact, more elaborate beamformers have been proposed for speech enhancement, but the maximum directivity beamformer is accurate enough for testing the overall system proposed.

Adaptive filtering In the second step an adaptive filter is applied to each spherical harmonic component indipendently.

The beamformer output $z(t)$ is used as input to a frequency-domain adaptive filter (FDAF), as presented in Section 2.2.3. The output $y_{nm}(t)$ is then compared with the desired signal, which corresponds to the noisy spherical harmonic coefficients $p_{nm}(t)$. The error $e_{nm}(t) = p_{nm}(t) - y_{nm}(t)$ is used for adaptively computing the filter weights. An overall scheme of this block is presented in Figure 3.9.

Note that this filtering block is not used in the noise cancelling configuration as in the work [17] presented in Section 2.3. The filter ouput $y_{nm}(t)$ is the final output of the system and it is the spherical harmonic coefficient of order n and degree m of the enhanced soundfield. The filtering block is rather designed to restore at the input $z(t)$ the spatial informations lost in spatial filtering while the effective denoising operation is implemented by the beamformer itself.

This operation is repeated for all the components up to order h with $h < N$ and N is the maximum order choosed for spherical harmonic representation. The remaining coefficients of order greater than h up to N will be forwarded as they are at the output.

The choice of h depends on the nature of the noise. If we are dealing with a diffuse noise, h can be very small respect to N. This implementative decision is justified from the considerations we have made in Section 3.2 about the diffuse noise. We have

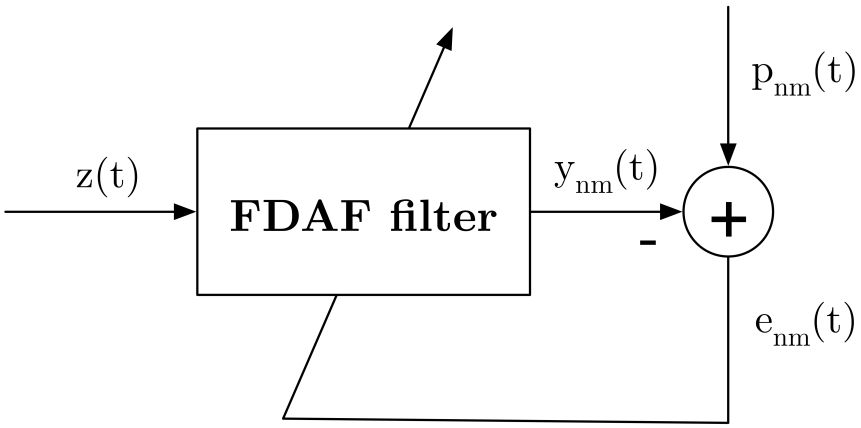

Figure 3.9: Block of adaptive filtering for a single spherical harmonic coefficient

shown that the noise signal representation in the spherical harmonic domain is not equally distributed on all the components from an energetic point of view. In the diffuse case, the 0-th order is more significant with respect to the components of order $n \neq 0$, where the diffuse noise contribution will be smaller with respect to the desired directional signal contribution. Consequently it is sufficient to reconstruct through the adaptive filtering stage only the first components and using as final result for the higher orders the noisy ones. In Chapter 4 will we show that this approach leads to a good approximation of the spherical harmonic coefficients of the noise-free soundfield, i.e. the sound field as if only the source was present. Moreover an analogue reasoning can be applied for the directional noise case. Since as showed in Section 3.2 the SNR for the higher orders is higher, the stop order h can be smaller of N and still the result is a good reconstruction of the denoised sounfield. Also this intuition will be confirmed by the results in Chapter 4.

This implementation allows to obtain a consistent drop in computational cost. In fact the filter block is applied only $(h+1)^2$ times instead of $(N+1)^2$, when all the components are filtered, or Q times, in case the adaptive filtering is perfomed for each microphone signal in the space-time domain.

Moreover having a small number of filters allows a finer tuning of the initialization parameters singularly for each FDAF filter, hence obtaining globally a more efficient filtering phase (more informations about the inizialization of the FDAF can be found in Section 2.2.3 and in [16]).

3.5 Algorithm review

In this section we will summarize the steps of the proposed algorithm.

Starting with the signals captured from a rigid spherical microphone array of Q microphones at positions $\mathbf{r_q} = (\theta_q, \phi_q, r)$ with $q = 1, ..., Q$:

1. The microphones signals are pre-processed using the transformations proposed in Section 2.1.1, in particular the real pre-processing chain. The maximum order of the SHT is N. The result is $P_{nm}(l, \nu)$ and is the input for both the DOA estimation and the beamformer.

2. From $P_{nm}(l, \nu)$ the narrowband pseudospectrum $\mathrm{P}_{MUSIC}(\nu, \Omega)$ is extracted following the MUSIC DOA estimation presented in Section 2.1.3. The source DOA Ω_s is computed by taking the average of the narrowband pseudospectrums and finding the peaks of the broadband pseudospectrum.

3. Having as parameter the source direction of arrival Ω_s and as input $P_{nm}(l, \nu)$, the maximum directivity beamformer described in Section 2.1.2 is applied to extract the desired source signal $z(\nu)$ in the frequency domain or $z(t)$ in the time domain, after inverse STFT.

4. A bank of indipendent $h < N$ FDAF filters are designed having as desired signal each coefficient of the SHT of the input p_{nm} up to $n = h$ and as input the beamformer output. The ouput coefficients approximate the SHT of the noise-free soundfield.

3.6 Final remarks

In this chapter we have proposed a signal enhancing system for 3D recordings that preserves the spatiality in the output. The system is composed of a first stage, where MUSIC DOA estimation is performed extracting the DOA of the desired source, and a second stage, where the desired source is extracted using a maximum directivity beamformer and the spatiality is restored through a bank of FDAFs. The result of this system is expressed in the spherical harmonic domain, hence it is indipendent from both the recording setup and any assumed reproduction system. In the next section we will present how simulations have been setted up and which results have been obtained using the proposed architecture.

Chapter 4

Simulations and Results

In this chapter we apply the proposed solution to a simulated soundfield and we study the behaviour of the evaluation metrics as the parameters of the simulation change. First the simulation setup is presented, explaining how soundfield has been simulated and the implementation parameters of each step. Then the evaluation metrics are presented for studying the effectiveness of the processing stage. Finally the results for the directional noise case and for the diffuse noise case will be presented compared to a benchmark approach, where the adaptive filtering is performed for each microphone signal.

4.1 Simulations Setup

The simulated soundfield assumes the presence of a single directional desired source and of a noise signal, that can be directional too or diffuse. The first step is to simulate the signals captured by an acquisition device, specifically a rigid spherical microphone array. Then we will present how the overall proposed system have been implemented in practice. Finally we will present a possible reproduction system, since the evaluation metrics will be estimated starting from the loudspeaker's signals.

4.1.1 Acquisition

The assumed acquisition device is a rigid spherical microphone array. For this simulation the microphones locations and the radius of the well-known MH Acoustics Eigenmike [5] has been considered.

The simulation of the microphone signals acquired from a specific sound field is obtained through two operations. First we need to compute the impulse response of each microphone given a set of specifications on the geometry of the field and on the reverberation of the room. Then we convolve each impulse response with a monoaural signal, that can be obtained from synthesis or can be the recording of a single microphone located in proximity of an acoustic source.

For our application we have selected two real monoaural signals: for the desired source we have choosed a violin sound, in particular an excerpt from an artistic performance, while for the noise the recording of an air-conditioner. The frequency sampling for both is $f_s = 44100$Hz.

For the simulation of the microphone impulse responses we used the Spherical Microphone array Impulse Response (SMIR) generator, an algorithm that generates the impulse responses between a source and a spherical microphone array, by modeling early reflections within a shoebox room of arbitrary dimensions.. An implementation for MATLAB is available at [19] and an accurate analysis of the algorithm is given in [20].

The SMIR generator has been employed for generating both a directional source and a diffuse source impulse responses. This tool allows to specify a number of parameters like for example the geometry of the sound scene or the reverberation of the assumed room: one of these is the reverberation time, specifically the T60 as empirically formulated by Sabine, presented in [21]. For testing the system in several conditions, the simulation of the sound field has been performed for five different values of T60. The parameters of the SMIR toolbox used for simulating a directional desired source and a directional noise can be found in Table 4.2 and 4.3. For the diffuse components simulation we positioned a source quite far from the microphone array and we set high reflection order and a long T60. By considering only the tail of the impulse response, a good approximation of the impulse response caused by a diffuse soundfield is given. A section of the selected parameters are shown in Table 4.1. The sampling frequency common to all the impulse responses is $f_s = 44100$Hz

Room dimensions	$\mathbf{L}_{room}$=[10 m,10 m,10 m]
Source-array distance	$d = 4.24$ m
Reflection order	$N_r = 30$
Room reflection coefficients	$\mathbf{C}_{refl}$ = [1,1,1,1,1,1]

Table 4.1: Setup for the SMIR generator in case of diffuse sound field

After the convolution step, the microphone signals for the directional desired source and directional noise or diffuse noise will be summed together, scaling the noise components by a constant factor. The scaling factor is computed depending on the desired global SNR: several SNR values have been used for this operation, specifically ten values from $\text{SNR}_m = 0$dB to $\text{SNR}_m = 25$dB.

4.1.2 Processing

In this section the implementative details for the pre-processing, analysis and processing stages are presented.

Room dimensions	$\mathbf{L}_{room}$ = [10 m, 15 m, 18 m]
Source-array distance	d = 3 m
Source DOA	$\boldsymbol{\Omega_n}$ = [2.09, 2.35]
Reflection order	$N_r = 1$
Room T60	**T60** = [0.5 s, 1 s, 1.5 s, 2 s, 2.5 s]

Table 4.2: Setup for the SMIR generator in case of directional noise sound field

Room dimensions	$\mathbf{L}_{room}$ = [10 m, 15 m, 18 m]
Source-array distance	d = 1 m
Source DOA	$\boldsymbol{\Omega_s}$ = [0.52, 0.52]
Reflection order	$N_r = 1$
Room T60	**T60** = [0.5 s, 1 s, 1.5 s, 2 s, 2.5 s]

Table 4.3: Setup for the SMIR generator in case of directional source sound field

The discrete SHT has been perfomed working with the real spherical harmonic functions, presented in Section 1.2.2 and up to a maximum order $N = 3$. A detailed explanation for this choice will be given later in Section 4.1.3.

The parameters for the STFT and inverse STFT operations are reported in Tables 4.4 and 4.5.

Sampling frequency	44100 Hz
Window type	Hamming
Window length	23 ms
Overlap	75%
FFT length	1024 1

Table 4.4: Parameters for the STFT

Sampling frequency	44100 Hz
Window type	Rectangular
Window length	23 ms
Overlap	75%
FFT length	1024

Table 4.5: Parameters for the inverse STFT

The MUSIC DOA estimation block has been implemented starting from the MATLAB library Acoustical Spherical Array Processing Library available at [22].

The FDAF used is the one implemented in the Matlab adaptive filtering toolbox[23]. The parameters and initialization setup are described in Table 4.6. Note that the same initial configuration is used for the FDAF of each spherical harmonic coefficient.

4.1.3 Reproduction

The evaluation metrics that we will propose in Section 4.2 are formulated starting from the speakers signals for a simulated reproduction system. More specifically, the result of the signal enhancement has to be coded in a specific format for a specific configuration of loudspeakers. Among various spatial audio reproduction approaches the HOA has been chosed,

Filter length	1024
Step size	$\mu = 0.025$
Averaging factor	$\lambda = 0.9$
Initial FFT input power	$P(0) = 0.01$

Table 4.6: Parameters for the FDAF

which is an extension of the Ambisonic tecnique. The term HOA refers to a set of tecniques for recording a real sound scene or encoding synthetized ones and generate the loudspeaker signals for a specific setup. The common factor of these tecniques specifically for HOA is the use of the spherical harmonic domain decomposition up to a maximum order $N > 1$, while for classic Ambisonic the decomposition is up to the first order. The mapping of the input in the ouput is divided in two indipendent stages, the encoding and the decoding stage. The encoding stage transforms the signals coming from a set of microphones in the spherical harmonic domain: an example of this stage starting with a rigid spherical microphone array can be found in Section 1.3. A deeper analysis about 3D recording using HOA can be found in [24]. The decoding stage is designed for a specific configuration of loudspeakers and given the maximum order of the SHT N. In its simplest implementation, decoding is performed through a matrix multiplication between a decoding matrix $\mathbf{D}$ and the spherical harmonic coefficients of the soundfield for a given time t, indicated as $\mathbf{x}(t) = [x_{00}(t), x_{1(-1)}(t), x_{10}(t), ..., x_{NN}(t)]^T$. The result of the decoding is a set of S loudspeakers signals, each one will be indicated as $x_s(t)$.

For the computation of the evaluation metrics we have used a decoding matrix computed for an arbitrary loudspeaker configuration following the AllRAD algorithm proposed in [25].

The positions of the loudspeakers, arranged on a sphere of radius $r_d = 1.44$ m, are shown in Figure 4.1. This reproduction system is the one present at the acoustic laboratories of the Museo del Violino in Cremona.

4.2 Evaluation Metrics

For studying the behaviour of the proposed system we have choosed a set of metrics. The first one is the classic NMSE. The NMSE measures how precisely an estimator approximates a variable. It corresponds to the MSE normalized for the energy of the variable to be estimated. Given an approximation $\hat{y}(t)$ of a time dependent signal $y(t)$, the NMSE is defined

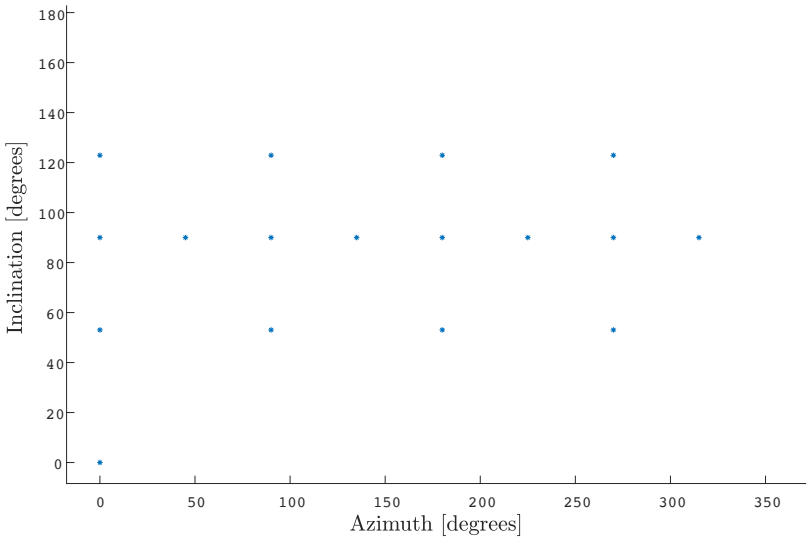

Figure 4.1: Loudspeaker distribution for the evaluation metrics computation

as

$$\text{NMSE}(t) = \frac{E\{(\hat{y}(t) - y(t))^2\}}{E\{y(t)^2\}} \tag{4.1}$$

where $E\{\cdot\}$ is the expected value operator. In practice the time signal will be divided in J rectangular windows of length K over which $y(t)$ and its estimation $\hat{y}(t)$ are assumed to be stationary. The NMSE will be approximated for the j-th window as

$$\text{NMSE}_j = \frac{\frac{1}{K}\sum_{i=0}^{K} |(\hat{y}(i+jK) - y(i+jK))|^2}{\frac{1}{K}\sum_{i=0}^{K} |y(i+jK)|^2}. \tag{4.2}$$

The NMSE is useful to study the performances of the proposed system. More precisely it will express how similar is the reconstructed denoised soundfield to the soundfield as if only the source was present.

Another useful measure for our purpose is the SNR. Given a desired signal $x(t)$ and a noise signal $n(t)$ the SNR is defined as:

$$\text{SNR}(t) = \frac{E\{x(t)^2\}}{E\{n(t)^2\}}. \tag{4.3}$$

Using the same methodology used for the NMSE an approximation of

the SNR for the j-th window is given by

$$\mathrm{SNR}_j = \frac{\frac{1}{K}\sum_{i=0}^{K}|x(i+jK)|^2}{\frac{1}{K}\sum_{i=0}^{K}|y(i+jK)|^2}. \tag{4.4}$$

In our application we will study the relationship between SNR in input and the SNR after the denoising process has been applied. If the last one is higher than the first one then the processing block is enhancing the signal and suppressing the noise.

4.3 Results

In this section we will show the results obtained from the simulations illustrated in Section 4.1 and measured in terms of the metrics defined in Section 4.2. Both SNR and NMSE are computed starting from the signals decoded for a reproduction setup. More precisely, the reproduction system used is the one described in Section 4.1.3. In the following sections first a benchmark approach will be proposed. Then an analysis of the performance metrics will be engaged for both the two types of noise soundfield, directional and spatially diffuse.

4.3.1 Adaptive filtering in the space domain

The proposed system has been compared to an alternative configuration for the adaptive filtering block. This approach is more intuitive and immediate with respect to the proposed one hence it is suitable for giving us a benchmark. In this case the soundfield is no longer represented and filtered in the spherical harmonic domain but in the space domain, i.e. sampled at the microphones positions. The signals involved in the adaptive filtering process are hence the one sensed by a spherical microphone array and they are denoted with $p_q(t)$ with $q = 1, .., Q$ and Q is the total number of microphones. A bank of Q adaptive filters are designed having as input the result of beamforming $z(t)$ and as desired signal each $p_q(t)$. The output of each filter $y_q(t)$ corresponds to the denoised signal for the q-th microphone. The adaptive filter choosed is still the FDAF and the initial configuration is the same used for our solution and displayed in Table 4.6. A scheme of one block of the filter bank is shown in Figure 4.2. If compared with Figure 3.9 the differences between the two methods are further evident.

Note that in this configuration the filtering process is repeated indipendently for all the Q signals. The space domain representation of the soundfield does not allow to avoid filtering some of the $p_q(t)$ as we have done in the spherical harmonic domain.

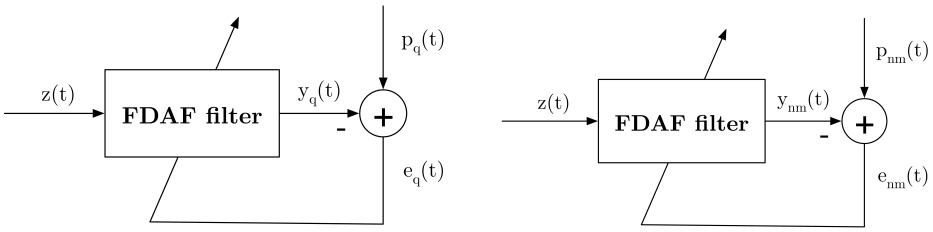

Figure 4.2: Block of adaptive filtering for a single microphone signal on the left, for a single spherical harmonic coefficient on the right

Table 4.7: SNR at microphones and T60 values used for the simulations

T60	0.5	1.0	1.5	2.0	2.5						
SNR_m	0.0	2.5	5.0	7.5	10.0	12.5	15.0	17.5	20.0	22.5	25

4.3.2 Directional noise case

As explained in the Section 4.1 a set of simulations have been synthetized having a directional desired source and a directional noise source. The parameters of each simulation are shown in Table 4.7, where SNR_m corresponds to the averaged SNR imposed at the microphones and T60 is the reverberation time.

Referring to Section 3.4, after the desired source signal has been extracted by the beamformer, the bank of adaptive filters restore the spatial informations by iteratively adapt the input, which is the beamformer output $z(t)$, to a desired signal, which is the noisy spherical harmonic coefficient $p_{nm}(t)$. One filter for each component of order n and degree m is designed up to a stop order $h \leq N$, where N is the maximum order of the SHT. In our case, as described in Section 4.1, the maximum order assumed is $N = 3$. The stop order h is an important parameter that need to be tuned for achieving both efficiency and good performances of the overall system. Therefore the algorithm has been tested stopping the adaptive filtering at all the orders from the minimum $h = 0$ up to the maximum one $h = N = 3$.

4.3.2.1 SNR

The SNR in general expresses the relationship between the desired source and the noise in terms of power. Since the goal of our application is to suppress the interfering noisy signal, we expect an increase of SNR after applying the proposed processing scheme to the acquired microphone signals.

In order to compute the SNR after the processing, we exploit the linearity of the proposed system. More specifically, we compute separately

the energy of the noise and of the source, before and after applying the denoising processing scheme. The approximation used for the computation of the SNR is the one proposed in (4.4). The SNR at the input for the s-th loudspeaker in the j-th window is denoted with $\text{SNR}_j^{i,s}$. Referring to (4.4), it is extracted using as desired source the loudspeaker signal $x_s(t)$ and as noise $n_s(t)$, obtained from the decoding of the representation of a soundfield where only source is present and a soundfield where only noise is present respectively. Then the $\text{SNR}_j^{i,s}$ are averaged over all the S loudspeakers obtaining SNR_j^i.

The SNR at the output for the s-th loudspeaker and for the j-th window is denoted with $\text{SNR}_j^{o,s}$. In this case isolating the desired source and noise components after the proposed denosing algorithm requires a further step. All the processing chain, beamformer and adaptive filtering, is applied to a soundfield where only the source or only the noise are present for all the stop orders $h = 0, 1, 2, 3$. The results are then decoded for the reproduction setup obtaining the after-processing desired source $\tilde{x}_s(t)$ and the after-processing noise $\tilde{n}_s(t)$. The SNR as defined in (4.4) and using $\tilde{x}_s(t)$ and $\tilde{n}_s(t)$ is computed for each s and then averaged over all the S loudspeakers. The result is denoted with SNR_j^o. This operation is repeated using the benchmark algorithm proposed in 4.3.1.

In Figure 4.3 the values of SNR_j are plotted against the window index j, hence their time behaviour is shown. The simulation parameters considered are SNR_m =15 dB and T60= 1.5 s. The curves plotted represent:

- SNR_j^i, indicated in the legend with the label 'input';
- SNR_j^o for the proposed solution stopping at order $h = 0$, $h = 1$, $h = 2$ and $h = 3$;
- SNR_j^o for the benchmark approach, indicated with the label 'sdf';

As evident from the plot, both the proposed and the benchmark approach lead globally to an increment of the SNR at the output with respect to the input SNR. In particular the benchmark approach produces results very similar to the one obtained for stop order $h = 3$. Intuitively, by filtering all the components up to the order N of the SHT we are exploiting all the spatial informations provided by the spherical harmonic representation for the reconstruction of the noise-free sound field. This is the same situation achieved when the adaptive filtering is implemented in the space-time domain, hence using all the microphones signals. The difference in the SNR at the ouput will be due to numerical precision and in practice we can consider these two methods equivalent. For this reason from now on we will not show the benchmark approach results, but we will consider them equal to the results produced by the proposed algorithm stopping at order $h = 3$. As underlined in the figure, stopping the adaptive filtering stage at $h = 0$ produces a smaller boost of the SNR at the output with respect to stop orders $h > 0$. This increase is still significant and it is justified by the considerations we have done about

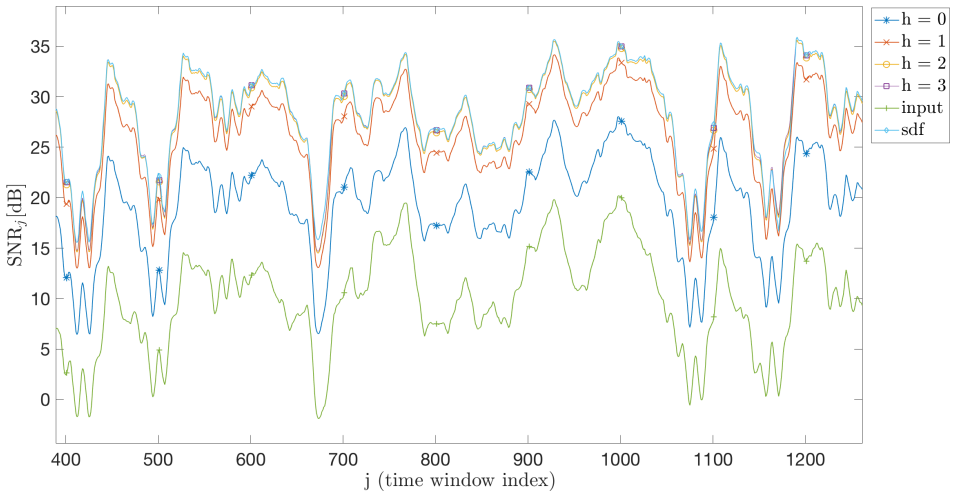

Figure 4.3: SNR$_j$ over the window index j in the directional noise case

the spatial properties of the directional sound field in Section 3.2. Since the energy of the noise is higher in the first components of the spherical harmonic representation, stopping at order 0 have produced an acceptable increment of the SNR with a minimum effort. In fact the adaptive filter in the spherical harmonic domain has been applied only one time, for order 0 and degree 0, while for higher orders at the ouput the coefficients of the noisy soundfield are forwarded at the ouput unchanged. This reasoning can be further extended to explain why the SNR$_j^o$ for orders $h = 1$, $h = 2$ and $h = 3$ are so similar. Increasing the number of components in the adaptive filtering does not lead to a substantial increase in the SNR because the noise is very small energetically in the higher order components.

In the next two plots, the values of SNR$_j$ are further averaged over the J windows. Then the increment of the SNR obtained through the denoising algorithm is extracted and it is indicated with $\Delta\text{SNR} = \frac{1}{J}\sum_{j=1}^{J} \text{SNR}_j^o - \text{SNR}_j^i$. This procedure has been repeated for all the values of T60 and SNR$_m$.

In Figure 4.4 the values of ΔSNR are plotted against the reverberation time T60 values with a fixed SNR at the microphones SNR$_m$ = 15dB. Each curve corresponds to a different stop order, which are $h = 0$, $h = 1$, $h = 2$ and $h = 3$. The boost in terms of SNR decreases just slightly for higher values of T60. This result is positive, since it shows that the system proposed is robust even if applied to signals acquired in enclosed spaces with long reverberation time.

In Figure 4.5 the values of ΔSNR are plotted against the SNR$_m$ values with a fixed T60 = 1.5 s. The curves presented are obtained from the output of the proposed algorithm stopping at order $h = 0$, $h = 1$, $h = 2$ and $h = 3$. The plot shows that the increment of the SNR obtained from the application of the denoising algorithm is further increasing with

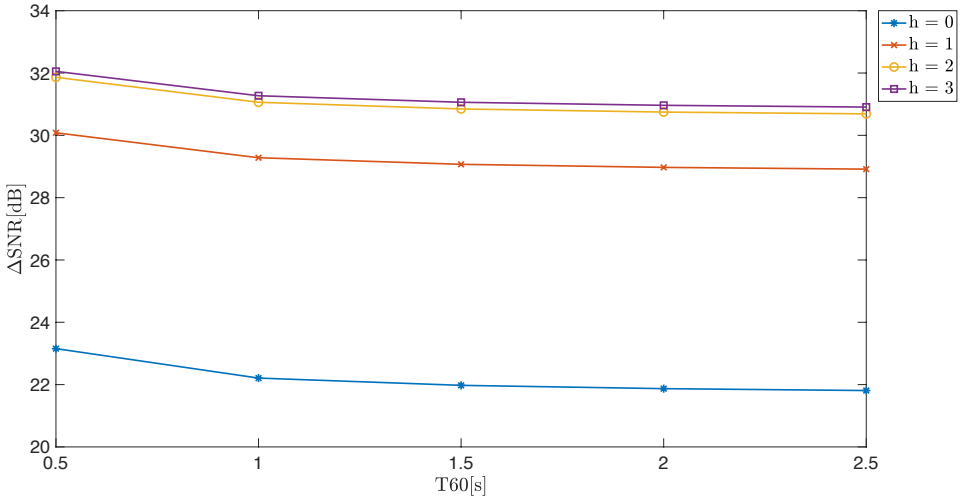

Figure 4.4: ΔSNR for T60 = [0.5,1,1.5,2,2.5] s and SNR_m = 15dB in the directional noise case

respect to the input SNR. Intuitively for higher values of the SNR at the input, the noise will be more easily suppressed through spatial filtering, improving the overall performances in terms of SNR of the proposed system.

4.3.2.2 Normalized MSE

In our application the NMSE values states how good is the approximation of the noise-free soundfield. Given that the goal of the proposed denoising algorithm is to reconstruct the soundfield as if only the source were present, the NMSE corresponds to the normalized approximation error of this estimation.

The NMSE values are computed for both the input and the ouput of the processing chain. The approximation used is the one defined in (4.2).

The NMSE at the input for the s-th loudspeaker and for the j-th window is denoted with $\text{NMSE}_j^{i,s}$. Referring to (4.2), the approximation signal $p_s(t)$ corresponds to the signal for the s-th loudspeaker after the decoding of the noisy soundfield representation. The approximated signal is $x_s(t)$ and it corresponds to the s-th loudspeaker signal after the decoding of a soundfield where only source is present. Then the $\text{NMSE}_j^{i,s}$ are averaged over all the S loudspeakers obtaining NMSE_j^i. Note that in practice the input NMSE is equal to the reciprocal of input SNR.

The NMSE at the output for the s-th loudspeaker and for the j-th window is denoted with $\text{NMSE}_j^{o,s}$. Referring to (4.2), the approximation signal $y_s(t)$ corresponds to the signal for the s-th loudspeaker produced by the decoding of the results of our proposed solutions. The approximated signal is again $x_s(t)$. $\text{NMSE}_j^{o,s}$ is computed for each s and then averaged over all the S loudspeakers.obtaining NMSE_j^o.

In Figure 4.6 the values of NMSE_j are plotted against the window

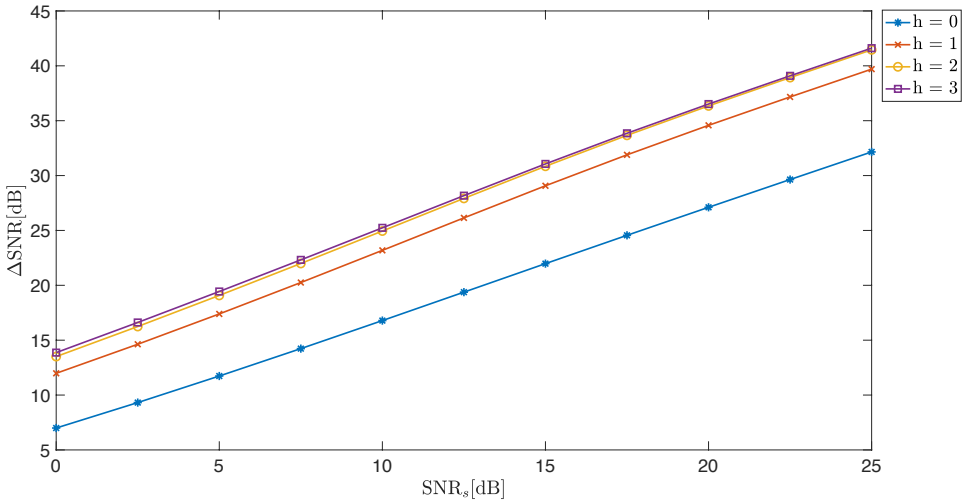

Figure 4.5: ΔSNR for T60=1.5 s and SNR$_m$=[0,2.5,5,7.5,10,12.5,15,17.5,20,22.5,25] dB in the directional noise case

index j, hence their time behaviour is shown. The simulations parameters considered are SNR$_m$ = 15dB and T60 = 1.5s. The curves plotted are the NMSE$_j^i$ indicated in the legend with the label 'input' and NMSE$_j^o$ for stop order $h = 0$,$h = 1$,$h = 2$ and $h = 3$; The plotted curves show that the NMSE after denoising process is in general lower than the one in the input, therefore we are actually approximating the noise-free soundfield. From this plot a rough analysis of the NMSE behaviour depending on the stop order h can be outlined. Again the NMSEs for order $h = 1$, $h = 2$ and $h = 3$ are very similar while the curve for stop order $h = 0$ has on average sligthly greater values of the NMSE with respect to the others h. Therefore in the NMSE sense, implementing the algorithm stopping at order $h = 0$ produce a less precise but still acceptable reconstruction.

In the next plots, the values of NMSE$_j$ are further averaged over the J windows. Then the averaged NMSE at the output is computed and denoted $\overline{\text{NMSE}}^o = \frac{1}{J}\sum_{j=1}^{J} \text{NMSE}_j^o$. Notice that the lower is $\overline{\text{NMSE}}^o$ the more precise is the estimation of the noise-free soundfield. This procedure has been repeated for all the considered values of T60 and SNR$_m$.

In Figure 4.7 the values of $\overline{\text{NMSE}}^o$ are plotted against the considered values of T60 having as SNR$_m$ = 15 dB. Each curve corresponds to a different stop order, more specifically $h = 0$, $h = 1$, $h = 2$ and $h = 3$. The variation is very small, $\overline{\text{NMSE}}^o$ is almost constant for each h hence we can conclude that, also in the NMSE sense, our algorithm is robust to variation of the reverberation time. Note that lower values of $\overline{\text{NMSE}}^o$ are obtained for stop order $h > 0$.

In Figure 4.8 the values of $\overline{\text{NMSE}}^o$ are plotted against the SNR$_m$ values with a fixed T60 = 1.5 s. The curves presented are obtained from the output of the proposed algorithm stopping at order $h = 0$, $h = 1$,

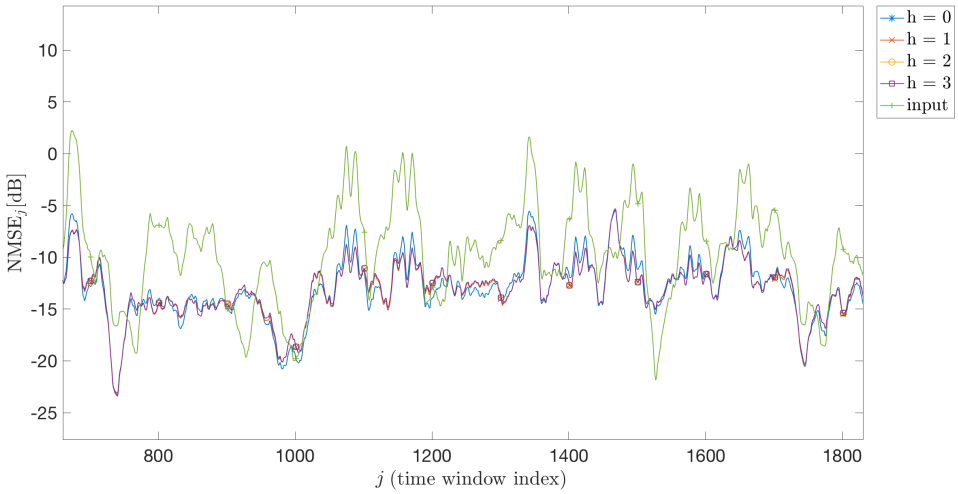

Figure 4.6: NMSE$_j$ over the window index j in the directional noise case

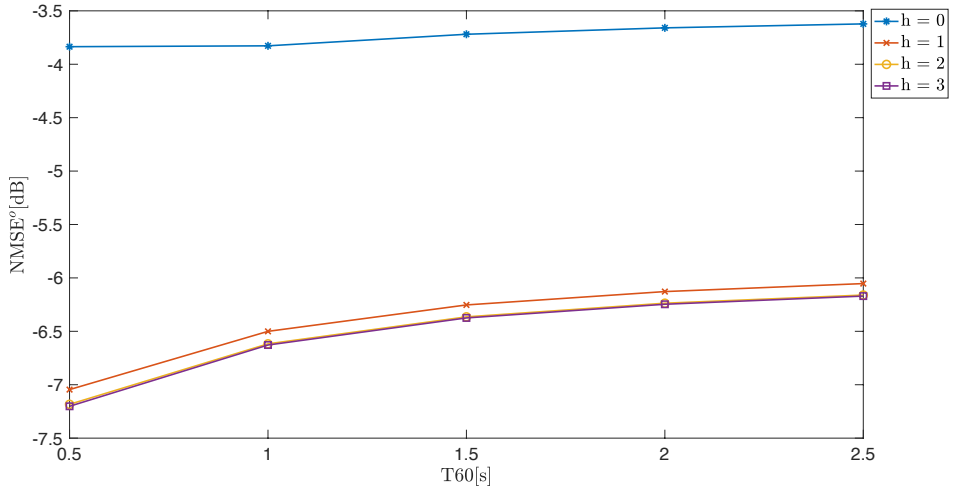

Figure 4.7: $\overline{\text{NMSE}}^o$ for T60 = [0.5,1,1.5,2,2.5] s in the directional noise case

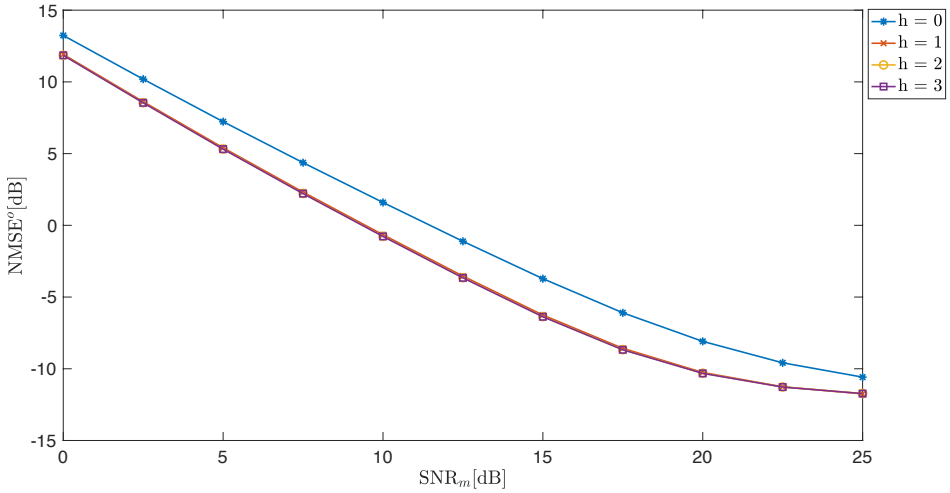

Figure 4.8: $\overline{\text{NMSE}}^o$ for T60 = 1.5 s and SNR$_m$= [0,2.5,5,7.5,10,12.5,15,17.5,20,22.5,25] dB in the directional noise case

$h = 2$ and $h = 3$. The plot shows that for lower values of the SNR at the microphones the approximation of the noise-free soundfield is less precise, in fact NMSE values are higher. This behaviour is produced by the fact that if the noise is energetically comparable to the desired source the beamformer result can include some noise components. By designing a more complex spatial filter that achieve a more precise extraction of the desired source components this problem can be overcame. By the way with the proposed solution for values of SNR$_m > 7dB$ an acceptable estimation of the desired source is achieved. Also in this case stopping at orders higher than 0 an even better NMSE is obtained.

4.3.3 Diffuse noise case

A set of simulations have been synthetized having a directional desired source and a spatially diffuse noise. The parameters of each simulation are shown in Table 4.7, where SNR$_m$ corresponds to the SNR imposed in average at the microphones and T60 is the reverberation time for the directional source only. As explained above in Section 4.3.2, a further parameter of the simulation is the stop order h.

4.3.3.1 SNR

The definition of the signals involved and the formulation for SNR is the one outlined in Section 4.3.2.1.

In Figure 4.9 the values of SNR$_j$ are plotted against the window index j. The simulations parameters considered are SNR$_m$ = 15dB and T60 = $1.5s$ for the directional desired source. The curves plotted represent SNR$_j^i$, indicated in the legend with the label 'input' and SNR$_j^o$ for the proposed solution stopping at orders $h = 0$, $h = 1$, $h = 2$ and $h =$

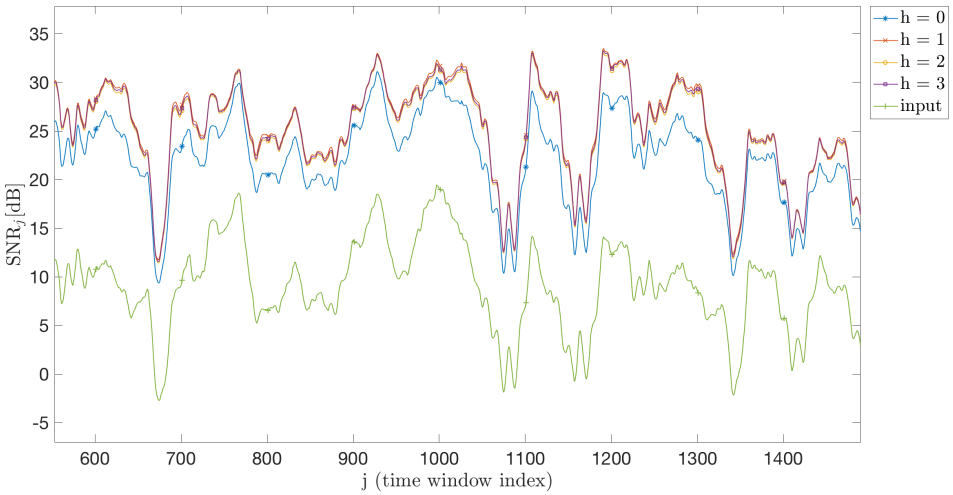

Figure 4.9: SNR_j over the window index j in the diffuse noise case

3. With respect to the directional case, SNR_j for stop order $h = 0$ is higher, hence implementing adaptive filtering up to maximum order 0 yields even better performances. This behaviour is expected, given the considerations done about the spatial properties of a diffuse sound field in Section 3.2. Since the diffuse noise components are concentrated mostly in the first spherical harmonic coefficient, noise suppression is achieved even using a single adaptive filter. This result is particularly positive since it shows that in case of diffuse noise good results are achieved with a minimum effort. The values of SNR_j for stop order $h = 1$, $h = 2$ and $h = 3$ are are equivalent in practice.

As we have done for the directional noise case, the variation of SNR between the input and the ouput is computed and denoted with ΔSNR.

In Figure 4.10 the values of ΔSNR are plotted against the reverberation time T60 values for the desired directional source having fixed SNR at the microphones $\mathrm{SNR}_m = 15$dB. Each curve represents the application of the algorithm with a different stop order, which are $h = 0$, $h = 1$, $h = 2$ and $h = 3$. As observed in the directional noise case, the boost in terms of SNR decreases just slightly for higher values of T60. This is a remarkable results, since it shows that the system proposed is robust also as for reverberating rooms.

In Figure 4.11 the values of ΔSNR are plotted against the SNR_m values with a fixed T60 = 1.5 s for the directional desired source. The curves presented are obtained from the output of the proposed algorithm stopping at order $h = 0$, $h = 1$, $h = 2$ and $h = 3$. The plot shows that the increment of the SNR obtained from the application of the denoising algorithm is further increasing with respect to the input SNR. The same reasoning we have done for the directional case can be applied here. For higher values of the SNR at the input, the noise will be better suppressed through spatial filtering, hence the overall performances exhibits

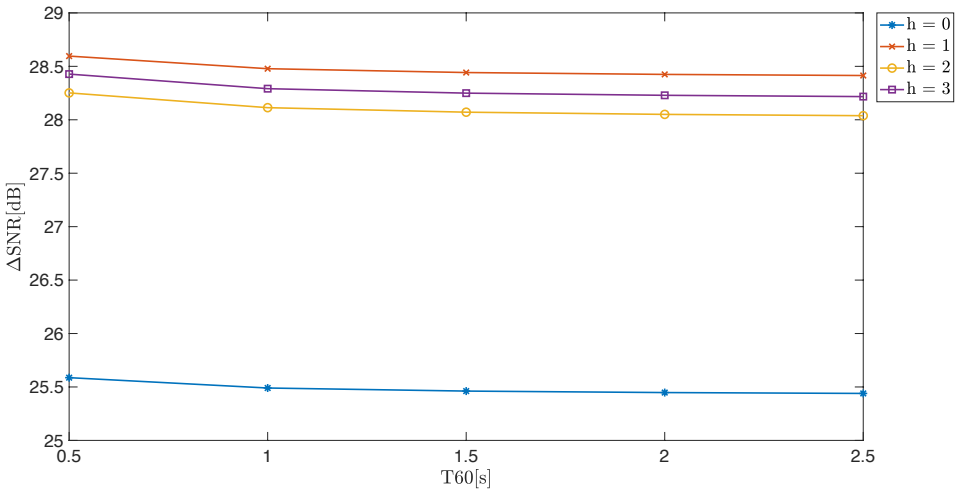

Figure 4.10: ΔSNR for T60 = [0.5,1,1.5,2,2.5] s and SNR_m = 15dB in the diffuse noise case

a noticeable improvement.

4.3.3.2 Normalized MSE

The description of the signals involved and the expression for NMSE is the one outlined in Section 4.3.2.2.

In Figure 4.6 the values of NMSE_j are plotted against the window index j, hence their time behaviour is shown. The simulations parameters considered are $\text{SNR}_m = 15\text{dB}$ and $\text{T60} = 1.5s$ for the directional desired source. The curves plotted are the NMSE_j^i indicated in the legend with the label 'input' and NMSE_j^o for stop order $h = 0$,$h = 1$,$h = 2$ and $h = 3$. Compared to the directional noise case, now the curve for $h = 0$ shows that in terms of NMSE stopping at order 0 allows to achieve very good results.

As we have done for the directional noise case, the output NMSE_j^o is averaged along the J windows, obtaining $\overline{\text{NMSE}}^o$.

In Figure 4.13 the values of $\overline{\text{NMSE}}^o$ are plotted against the considered values of T60 for the directional desired source having as $\text{SNR}_m = 15$ dB. Each curve corresponds to a different stop order, i.e. $h = 0$, $h = 1$, $h = 2$ and $h = 3$. Like in the directional case, the variation of $\overline{\text{NMSE}}^o$ is very small for all the stop orders, hence our algorithm is robust to variation of the reverberation time.

In Figure 4.8 the values of $\overline{\text{NMSE}}^o$ are plotted against the SNR_m values with a fixed T60 = 1.5 s for the directional desired source. The curves presented are obtained from the output of the proposed algorithm stopping at order $h = 0$, $h = 1$, $h = 2$ and $h = 3$. Also in this case if the noise is energetically comparable to the desired source the approximation of the noise-free sound field achieve a lower precision. As already mentioned, this behaviour can be overcame by designing a more complex

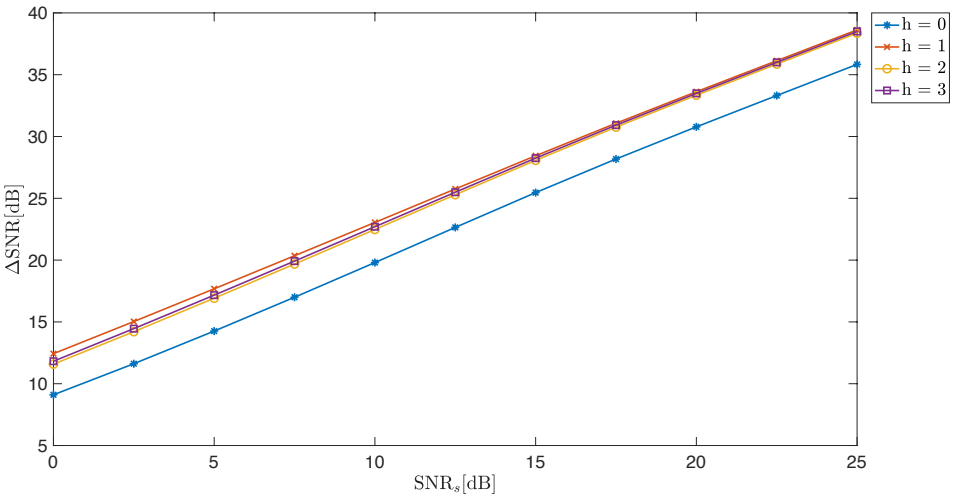

Figure 4.11: ΔSNR for T60=1.5 s and SNR$_m$=[0,2.5,5,7.5,10,12.5,15,17.5,20,22.5,25] dB in the diffuse noise case

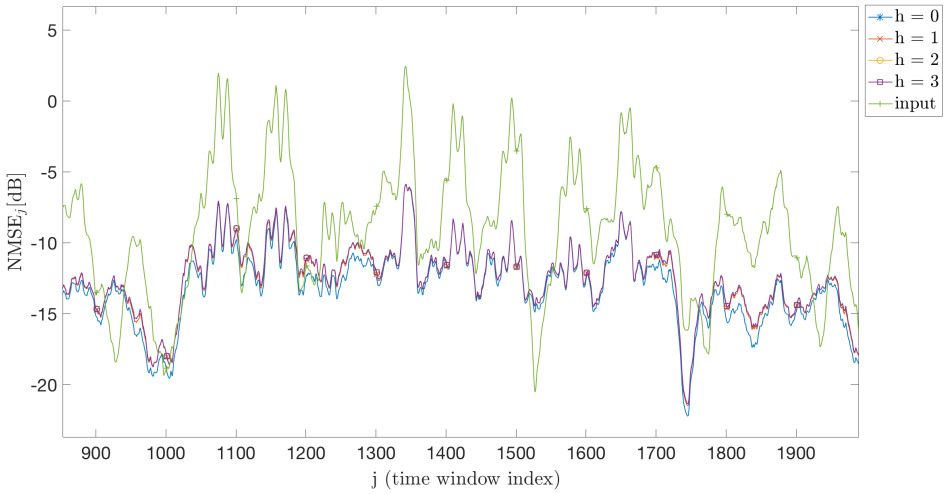

Figure 4.12: NMSE$_j$ over the window index j in the diffuse noise case

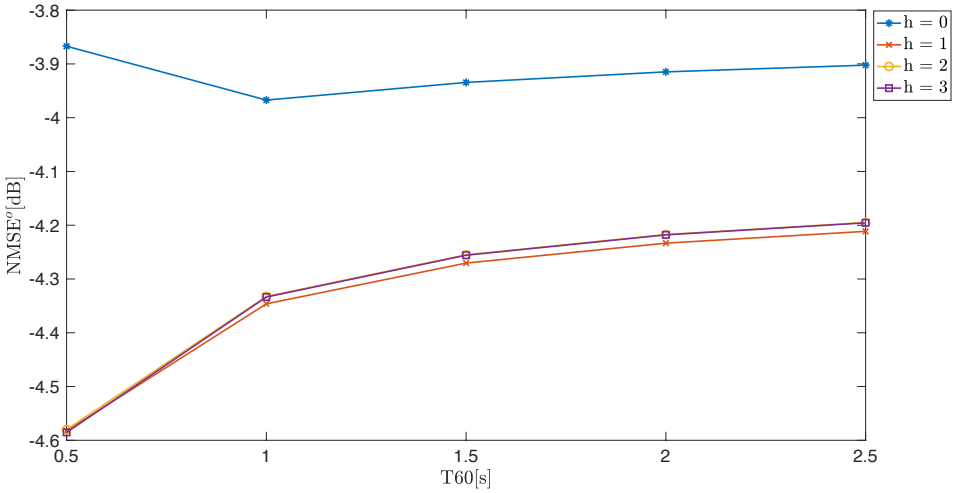

Figure 4.13: $\overline{\text{NMSE}}^o$ for T60 = [0.5 s,1 s,1.5 s,2 s,2.5 s] in the diffuse noise case

beamformer for extracting more precisely the desired source components. Moreover it can be pointed out that in average the NMSE at the output for stop order $h = 0$ is equal to NMSE for higher stop orders. This result further validates the fact that a very good result can be achieved in general with a minimum effort.

4.3.4 NMSE in the frequency domain

A final consideration on the NMSE is addressed in the frequency domain. The reconstruction of the noise-free soundfield achieved by the system is limited by the effects of spatial aliasing. As illustrated in Section 1.3, the maximum frequency for which perfect spatial reconstruction can be achieved by applying SHT and inverse SHT in chain is given by the parameters of the spherical microphone array and from the maximum order used for SHT N. For our configuration the spatial aliasing frequency corresponds to $f = \frac{Nc}{2\pi r} = 3900$Hz. We have computed the NMSE of the result of the proposed solution with respect to noise-free sound field in the frequency domain for $h = 3$, T60 = 0.5 s and $\text{SNR}_m = 35$ dB.

The results are shown in Figure 4.15 for a frequency range from 100 Hz to 15 kHz. Note that the desired signal is extracted from a violin performance. Consequently the NMSE at low frequencies is high because the desired source energy at low frequencies is negligible. Starting from approximately 100 Hz up to the spatial aliasing frequency $f_a = 3900Hz$ the NMSE is globally small, decreasing up to -15 dB. For $f > f_a$ NMSE gets higher and higher due to the effects of spatial aliasing.

4.3.5 Conclusive remarks

In this chapter we have shown how the proposed solution has been validated. In particular the results have focused on the relationship between

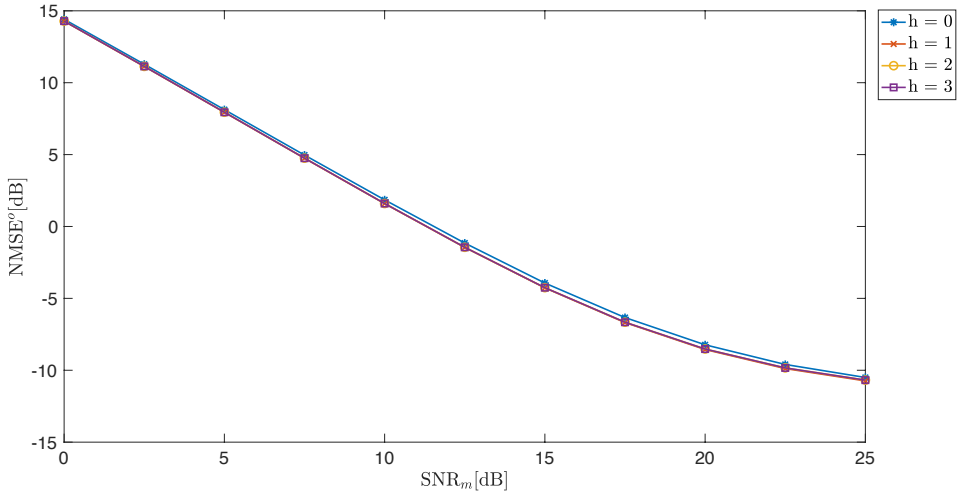

Figure 4.14: $\overline{\text{NMSE}}^o$ for T60 = 1.5 s and SNR$_m$= [0,2.5,5,7.5,10,12.5,15,17.5,20,22.5,25] dB in the diffuse noise case

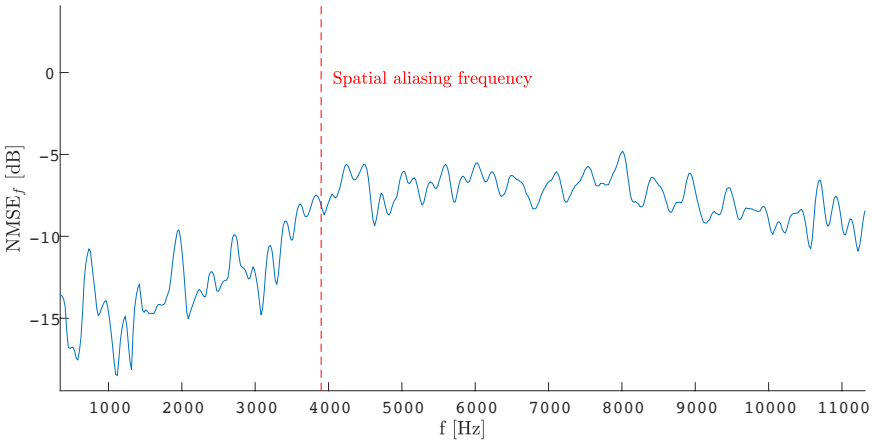

Figure 4.15: NMSE$_f$ plotted against frequency

the stop order for the adaptive filtering block and the metrics proposed. It can be concluded that our proposed method can be tuned to be particularly efficient in computation and at the same time ensuring an effective noise cancellation.

Chapter 5

Conclusions and Future Works

In this work a denoising algorithm was developed, for 3D acoustic scenes captured by compact spherical microphone arrays. Denoising is accomplished in the spherical harmonic domain, with the goal of suppressing the noisy interference while preserving the spatial acoustic cues. We have assumed that the sound scene composed of a desired sound source, which can move in space, and a noise, which can be directional or spatially diffuse, has been captured from a spherical distribution of microphones mounted on a rigid baffle. The microphones signals are then projected by means of a linear transformation in the spherical harmonic domain. This operaton provides a meaningful representation of the spatial properties of the sound field and several analysis and beamforming techniques can be formulated efficiently in the spherical harmonic domain. Noise suppression is achieved by means of two steps. In a first analysis stage the DOA of the desired source is extracted by applying the well-known MUSIC algorithm suitable for the chosen domain. In the second processing stage first the design of a fixed-weights spatial filter is engaged. More precisely a Maximum Directivity beamformer is steered toward the estimated DOA of the desired source. The spherical harmonic coefficients are linearly combined with the beamforming weigths and an estimate of desired source signal is obtained. However this operation reduces the dimensionality of the input to a single channel, hence a second step is necessary for recovering the spatial information. A bank of frequency domain adaptive filters, one for each order and degree of the spherical harmonic representation, is designed having as input the source signal estimated by the beamformer and as desired signal each spherical harmonic coefficient of the acquired soundfield. The ouput of each adaptive filter is an estimation of the correspondent spherical harmonic coefficient of the noise-free soundfield. The adopted representation of the soundfield allows to improve the efficiency of this operation. In fact a further investigation has shown that the energy of both the directional and spatially diffuse noise fields are not equally distributed along the components of the spherical harmonic decomposition, hence a smaller number of adap-

tive filters have to be implemented in practice. This result allows to add to the proposed algorithm a tuning parameter, the stop order, i.e. the maximum order up to adaptive filtering is performed. By choosing a suitable value for the stop order it is possible to regulate the trade off between the computation efficiency and the accuracy in the estimation of the noise-free soundfield.

A number of simulations have been implemented, in which our method has been applied and compared to a benchmark approach. We observed that, in terms of both SNR and NMSE, the outlined denoising algorithm is effective and in a specific configuration equivalent to the benchmark algorithm. Moreover the results reveal that the overall system is robust to non-ideal conditions of the room in terms of reverberation and its behaviour is coherent in case of highly disturbed sound fields. This analysis has been repeated for all the values of the tuning parameter and has been shown that very good results can be obtained even with a minimum computational effort.

The proposed denoising approach produces promising results and can be tuned to be very efficient thanks to the properties of the spherical harmonic domain. Moreover this problematic is rarely addressed in the literature and this work is a functional tool that can be used in practical situations. In fact the processing technique designed can be successfully applied in post production for tridimensional recordings and the results are indipendent from any potential reproduction system.

The designed system can be further improved or extended in future works. For example sound scenes with more than one desired source can be considered. In this case both DOA tracking and beamforming can be extended to multiple source case. Moreover a more complex analysis stage can be engaged for extracting specific properties of the soundfield, like the diffuseness of the noise, and consequently optimizing the overall processing block. Alternatively the beamforming phase can be refined for precise source extraction, adopting for example the LCMV spatial filtering technique.

Bibliography

[1] D. Jarrett, E. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. 08 2016.

[2] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[3] E. G. Williams, *Fourier acoustics*. 1999.

[4] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8. 2015.

[5] MH Acoustics LLC, "Eigenmike microphone array." https://mhacoustics.com/products#eigenmike1.

[6] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 2182–2196, Dec 2014.

[7] D. P. Jarrett, M. Taseska, E. A. P. Habets, and P. A. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and narrowband doa estimates," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 967–978, May 2014.

[8] O. Thiergart and E. A. P. Habets, "An informed lcmv filter based on multiple instantaneous direction-of-arrival estimates," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 659–663, May 2013.

[9] A. Unsöld, "Beiträge zur quantenmechanik der atome," *Annalen der Physik*, vol. 387, no. 3, pp. 355–393.

[10] I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication: Challenges and Perspectives*. Springer Publishing Company, Incorporated, 2010.

[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, pp. 1408–1418, Aug 1969.

[12] Y. Peled and B. Rafaely, "Linearly constrained minimum variance method for spherical microphone arrays in a coherent environment," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 86–91, May 2011.

[13] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2828–2840, 2012.

[14] S. Tervo, "Direction estimation based on sound intensity vectors," in *Signal Processing Conference, 2009 17th European*, pp. 700–704, IEEE, 2009.

[15] P. S. Diniz, *Adaptive Filtering: Algorithms and Practical Implementation*. Kluwer Academic Publishers, 2002.

[16] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, vol. 9, pp. 14–37, Jan 1992.

[17] F. Ortolani and A. Uncini, "A new approach to acoustic beamforming from virtual microphones based on ambisonics for adaptive noise cancelling," in *2016 IEEE 36th International Conference on Electronics and Nanotechnology (ELNANO)*, pp. 337–342, April 2016.

[18] P. Berner, R. Toms, K. Trott, F. Mamaghani, D. Shen, C. Rollins, and E. Powell, "Technical concepts: Orientation, rotation, velocity and acceleration, and the srm," 2008.

[19] E. Habets, "Smir generator." https://www.audiolabs-erlangen.de/fau/professor/habets/software/smir-generator.

[20] D. Jarrett, E. Habets, M. Thomas, and P. A Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," vol. 132, pp. 1462–72, 09 2012.

[21] H. Kuttruff, *Room Acoustics*. Abingdon, U.K.: Spon, 5 ed., 2009.

[22] A. Politis, "Acoustical spherical array processing library." https://github.com/polarch/Spherical-Array-Processing.

[23] "Matlab dsp system toolbox," 2017.

[24] S. Moreau, J. Daniel, and S. Bertet, "3d sound field recording with higher order ambisonics – objective measurements and validation of a 4th order spherical microphone," vol. 1, 01 2006.

[25] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of the audio engineering society*, vol. 60, no. 10, pp. 807–820, 2012.

[26] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 130, no. 1, pp. 11–16, 1983.

[27] B. Rafaely and S. Member, "Analysis and Design of Spherical Microphone Arrays," vol. 13, no. 1, pp. 135–143, 2005.

[28] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," vol. 12, pp. 713 – 716, 11 2005.