POLITECNICO DI MILANO
Scuola di Ingegneria dell'Informazione

POLO TERRITORIALE DI COMO

Master of Science in
Computer Engineering

Evaluation of Artifacts in Sound Field Rendering Techniques: from an Objective to a Subjective Approach

Supervisor: Dr. Fabio Antonacci
Assistant Supervisor: Dr. Antonio Canclini

Master Graduation Thesis by: Lucio Bianchi
Student Id. number: 755135

Academic Year 2011/2012

POLITECNICO DI MILANO
Scuola di Ingegneria dell'Informazione

POLO TERRITORIALE DI COMO

Corso di Laurea Specialistica in
Ingegneria Informatica

La valutazione degli artefatti nelle tecniche di *rendering* di campi acustici: da un approccio oggettivo a un approccio soggettivo

Relatore: Dr. Fabio Antonacci
Correlatore: Dr. Antonio Canclini

Tesi di Laurea Specialistica di: Lucio Bianchi
Matricola: 755135

Anno Accademico 2011/2012

Sommario

In questa tesi proponiamo una metodologia basata su considerazioni psicoacustiche per valutare gli artefatti introdotti da sistemi di *rendering* di campi acustici. La realizzazione pratica di tecniche di *rendering* come Wave Field Synthesis (WFS) e Geometric Rendering (GR) provoca sempre delle approssimazioni che alterano il campo sonoro riprodotto. Queste approssimazioni causano fronti d'onda che impattano sull'area di ascolto prima (pre-echi) e dopo (post-echi) il fronte d'onda desiderato; considerato che questi fronti d'onda aggiuntivi sono concentrati in una finestra temporale corta, questi condizionano la percezione del timbro. Inoltre, le approssimazioni producono anche una distorsione della forma dei fronti d'onda, causando una localizzazione errata della sorgente sonora virtuale.

Noi proponiamo una classe di metriche per caratterizzare la distorsione timbrica basate sull'effetto psicoacoustico del mascheramento temporale e che usano una soglia di mascheramento mutuata dall'ambito della codifica percettiva dell'audio. Descriviamo anche una metodologia che utilizza una trasformata di Hough generalizzata per stimare la posizione di una sorgente virtuale data la curvatura dei fronti d'onda, permettendoci di valutare gli errori commessi da un ascoltatore nella localizzazione di una sorgente virtuale a causa degli artefatti spaziali introdotti dal sistema di *rendering*.

Dal momento che la nostra metodologia richiede la conoscenza della risposta all'impulso spazio-temporale, adottiamo una metodologia di misura ben conosciuta, basata su una schiera virtuale di microfoni; scomponendo il campo in armoniche circolari (Circular Harmonics Decomposition (CHD)) siamo in grado di estrapolare il campo acustico nell'intera area di ascolto.

Per ottenere una valutazione soggettiva degli artefatti abbiamo condotto dei test di ascolto formali. Il risultato dei test di ascolto ha una forte correlazione con i risultati della metodologia proposta in questa tesi, così possiamo concludere che la nostra metodologia di valutazione è abbastanza informativa da poter essere usata al posto di una più costosa valutazione soggettiva.

Abstract

In this thesis we propose a psychoacoustic-based methodology to evaluate the artifacts introduced by sound field rendering systems. We show that the practical realization of rendering techniques like Wave Field Synthesis (WFS) and GR always causes some approximations that alter the rendered sound field. These approximations produce acoustic wave fronts that impinge on the listening area before (pre-echoes) and after (post-echoes) the desired wave front; since these additional wave fronts are concentrated in a short time window, they affect the perception of the timbre. Furthermore, the approximations also produce a distortion of the shape of the rendered wave fronts, causing an erroneous localization of the virtual sound source.

We introduce a class of metrics aimed at characterizing the timbral distortion. In particular, our metrics are based on the psychoacoustic effect of masking in time domain and use on a masking threshold well known in the field of perceptual audio coding. We describe also a methodology that employs a generalized Hough Transform to estimate the position of a virtual source given the curvature of the wave fronts. This methodology allows us to evaluate the errors committed by a human listener in the localization of a virtual source, due to spatial artifacts introduced by the rendering system.

Since our methodology requires the description of the sound field in terms of the space-time impulse response, we adopt a well known measurement methodology, based on a virtual microphone array; this methodology adopts Circular Harmonics Decomposition (CHD) in order to extrapolate the sound field over the whole listening area.

In order to obtain a subjective evaluation of rendering artifacts we have conducted formal listening tests aimed at an assessment of timbral and spatial artifacts. The results of the listening tests have a strong correlation with the results of the psychoacoustic-based evaluation methodology, thus we can conclude that our evaluation methodology is informative enough to be used in place of a more costly subjective assessment.

Contents

List of Figures

List of Tables

List of Acronyms

WFS	Wave Field Synthesis
GR	Geometric Rendering
ILD	Interaural Level Difference
ITD	Interaural Time Difference
MLS	Maximum Length Sequence
CHD	Circular Harmonics Decomposition
FEM	Finite Element Method
SVD	Singular Value Decomposition
ITU-R	International Telecommunication Union - Radiocommunication Sector
BS	Broadcasting service (Sound)
EBU	European Broadcasting Union
SQAM	Sound Quality Assessment Material
FFT	Fast Fourier Transform
IFFT	Inverse Fast Fourier Transform

Nomenclature

p The sound pressure

c The sound speed

$\mathbf{n}$ The unit length vector

$\mathbf{x} = (x, y)^T$ Two-dimensional Cartesian coordinates

$\mathbf{r} = (\rho, \theta)^T$ Two-dimensional polar coordinates

$(\cdot)^T$ The operator that denotes vector transposition

$(\cdot)^H$ The operator that denotes Hermitian conjugate

PER^- Peak-to-(pre)-Echo Ratio

PER^+ Peak-to-(post)-Echo Ratio

DER^- Direct-to-(pre)-Echo Ratio

DER^+ Direct-to-(post)-Echo Ratio

Chapter 1

Introduction

This work of thesis is the result of an experimental research aimed at the evaluation of artifacts produced by sound field rendering techniques. In particular we have worked on objective and subjective methodologies able to provide informative results about the distortion introduced by rendering techniques. With objective evaluation we mean the acquisition of data and information from experimental or simulative results that are not mediated by human perception. On the other hand, the process of acquisition of data and information for a subjective evaluation involves judgments mediated by human perception.

The adoption of complex sound field rendering technologies is motivated by the fact that conventional sound reproduction methods, like stereophony and surround, suffer from serious problems. In particular, the spatial acoustic properties of the sound field reproduced by conventional methods can be perceived properly only in a small listening area, called *sweet spot* [1]. This limitation implies that only one listener at time can attend the acoustic performance in an optimal way and, moreover, the listener cannot move from the ideal position.

More complex sound field rendering techniques have been developed to overcome such limitations. We can classify these techniques into two groups, according to the technology adopted to reproduce the sound scene.

On the one hand we have rendering techniques based on binaural methods, which use headphones to reproduce the required sound pressure at the listener ears. This approach poses some practical limitations, since a tracking of the listener position is needed in order to deliver spatial acoustic cues coherent with its position in the virtual sound scene.

On the other hand, we have techniques that use loudspeaker arrays to

reproduce the sound field in a wide area, possibly enclosing many listeners; in this way the sweet spot limitation of conventional techniques is overcome. These are techniques in which the sound field produced by loudspeaker arrays appears to originate from any desired position in space; thus they create the illusion of a virtual sound source. This impression is delivered by producing wave fronts which are coherent with the position of the virtual source inside the sound scene. These techniques are usually referred to as *sound field synthesis*.

The aspect that differentiates sound field synthesis techniques is the principle behind the computation of loudspeaker driving signals. In our work we have employed WFS [2], which is based on the physics of wave propagation and GR [3], which is based on a geometrical model of sound propagation. If the requirements of their physical foundation are met, such techniques allow the exact reproduction of the desired sound field.

However, the practical realization of sound field rendering systems always imposes some constraints. In particular, the loudspeaker array only approximates a spatially continuous distribution of sound sources because it is limited in length and the spacing between loudspeakers is finite. Another constraint is that the radiance pattern of the loudspeakers is not purely omnidirectional at all frequencies in the audio range; thus, their behavior deviates from the one of an ideal acoustic monopole. Other technical constraints are imposed by the emission system (i.e. D/A converter, power amplifiers and loudspeakers) which is limited in bandwidth and whose frequency response is not perfectly flat.

In [4] the authors show that such approximations of the ideal system cause a distortion of the rendered sound field, which no more coincides with the desired one. In particular, the major distortions are the spatial aliasing due to the finite spacing between loudspeakers and truncation effects due to the limited aperture of the array. Moreover, in a real environment even reflections and reverberation contribute to alter the rendered sound field. It is shown in [5] that these deviations from the desired sound field cause artifacts that are perceivable by a human listener.

The literature presents two classes of approaches aimed at assessing the impact of sound field distortions on the human perception. On the one hand there are objective approaches, which prescribe the measurement of specific features of the rendered sound field in order to evaluate the differences from an ideal behavior. Methods based on objective measurements are important to provide an indication of the overall quality of a sound field rendering technique but they are not informative enough since they do not take human perception into account.

On the other hand there are subjective approaches, which are based on a formal listening test; in these tests a listener is asked to make judgments on specific features of the sound stimulus which is offered to him, thus their results are intrinsically mediated by human perception. The main drawback of subjective approaches consists of their high cost in terms of working time and people involved, due to the large number of trained listeners needed in order to obtain reliable results.

In our research activity we have proposed an alternative approach which takes the advantages of objective methods but it provides measures related to human perception. In particular, our approach provides a psychoacoustic obtained by metrics that are representative of human perceptive experience. In order to develop these metrics, a deeper insight on the nature of artifacts is needed.

We have considered the case of rendering an impulsive sound field; it is shown that, in such cases, the rendering systems exhibit a well defined direct wave front at all positions inside the listening area. In addition to this direct wave front, the loudspeaker array emits also secondary wave fronts immediately before and after the direct one: they are named *pre-echoes* and *post-echoes*, respectively [5]. The time interval between successive echoes is so short that such replicas are not perceived as separate acoustic events; however, it is shown in [6] that pre-echoes and post-echoes are perceived by a human listener as a timbral distortion of the rendered sound signal.

Within our research activity we have studied the impact of pre-echoes and post-echoes on the perception of timbral quality and we have introduced a class of metrics aimed at characterizing the timbral distortion of an impulsive sound signal delivered by a sound field rendering system. In particular, our metrics are based on the ratio between the power carried by the direct wave front and the power of pre-echoes and post-echoes. We have based our analysis on the psychoacoustic effect of masking in time domain in order to discriminate whether a specific replicated wave front affects the perception of timbral quality or not; the discrimination is based on the time interval between the considered echo and the direct one and, moreover, according to a masking threshold well known in the field of perceptual audio coding.

Furthermore, it is shown in [4] that the wave fronts produced by a rendering system are distorted by spatial aliasing and truncation effects, due to the finite aperture of the loudspeaker array. In particular, it is shown in [7] that a distortion of the shape of rendered wave fronts affects spatial acoustic cues like Interaural Level Difference (ILD) and Interaural Time Difference (ITD).

ILD is defined as the difference in level of the sound arriving at the two ears, mainly due to head shadowing effects. On the other hand, ITD is the difference in arrival time of the sound at the two ears. ILD and ITD are known to be the most important spatial acoustic cues that allows sound localization. Thus, if ILD and ITD are altered with respect to the desired situation, they may cause an erroneous localization of the sound source.

In order to better understand the effect of spatial artifacts, we can consider the case of rendering a distant sound source, whose acoustic contributions in the listening area would be plane waves. Mainly due to truncation effects caused by the finite aperture of the loudspeaker array, the wave fronts generated by the rendering system will not be plane, but their shape will be distorted. This causes the alteration of ILD [8] in such a way that the virtual source is perceived closer in space, with respect to the desired position.

In our work we have developed a methodology to retrieve the position of a virtual source given the curvature of the wave front produced. In this way we are able to evaluate the errors committed by a human listener in the localization of a virtual source, due to spatial artifacts introduced by the rendering system.

Our methodologies to evaluate the timbral artifacts and the localization require the knowledge of the room impulse response, which describes the sound field inside a listening area. In particular, the room impulse response is a function of space and time which characterizes what a listener placed in any point of the listening area would hear. The room impulse response visualizes explicitly the presence of pre-echoes and post-echoes, and moreover it allows to visualize the spatial distortion of the wave fronts.

A naive approach to measure the room impulse response would be to place a large number of microphones inside the listening area and recording the microphone signals generated by the emission of a suitable sound field. This approach appears to be very simple and straightforward but it has some obvious drawbacks. At first, it requires the use of a very large number of high quality microphones, pre-amplifiers and D/A converters in order to obtain a high resolution impulse response, thus its cost is prohibitive. Moreover, the presence of many microphones inside the listening area will cause scattering effects which will severely alter the recorded sound field.

In our work we have adopted the room impulse response measurement methodology described in [9] and [10]. This methodology employs only one microphone mounted on a rotating rig that allows to sample the sound field over a large number of positions on a circumference. The sound field acquired in this

way is then decomposed into *circular harmonics* and extrapolated on the whole listening area. For the purpose of measuring the room impulse response, the sound field rendering system emits a Maximum Length Sequence (MLS) which is white in a limited bandwidth [11]. This way the measurement cost is minimal since only one high quality acquisition chain is needed (i.e. one microphone, pre-amplifier and D/A converter) and the sound field is less severely altered by the presence of the measurement setup.

We have implemented a measurement setup based on this methodology in an acoustically controlled rendering room provided by the *Sound and Music Computing Lab* of Politecnico di Milano, Polo Regionale di Como. With this experimental setup we have measured the impulse response of sound field rendering systems reproducing a single virtual source located at different positions; thus we have been able to apply our methodologies to evaluate timbral and spatial artifacts on real data.

As the last step of our work, we have designed and conducted listening tests to evaluate the perceptive performance of two sound field rendering techniques: Wave Field Synthesis (WFS) and Geometric Rendering (GR). Our tests had the goal to separately assess the impact of timbral and spatial artifacts introduced by these rendering techniques on human listeners.

Finally, we have found a strong correlation between the results provided by our metrics and the results of the subjective tests. Thus we can state that our evaluation methodologies are informative enough to be used in place of a much more costly subjective assessment.

Outline

Now we present an outline of this manuscript. In Chapter 2 we describe state-of-the-art techniques devoted to sound field rendering. In particular, we present Wave Field Synthesis (WFS) and Geometric Rendering (GR) with emphasis on the principles behind them. Furthermore, we describe the nature of artifacts introduced by sound field rendering techniques and we present some contributions addressing the problem of an evaluation of these artifacts.

In Chapter 3 we present the measurement methodology that we have adopted to obtain a room impulse response. In particular, we introduce CHD to extrapolate the sound field over the whole listening area. Moreover, we show how to obtain a space-time representation and a space-frequency representation from the measured impulse response of the rendering system.

In Chapter 4 we propose a methodology for both objective and psycho-

acoustic-based evaluations of sound field rendering techniques in terms of the timbral quality and the localization of a rendered virtual source. We start by introducing the psychoacoustic effect known as masking in time domain; then we propose two psychoacoustic-based metrics to assess the impact of pre-echoes and post-echoes on the perception of timbral quality. Furthermore, we introduce a technique aimed at estimating the position of the virtual source given the curvature of the wave fronts in the listening area. This technique relies on a generalization of the Hough transform to analyze a snapshot of the wave fronts image and estimate the position of the virtual source that generated such wave fronts.

In Chapter 5 we describe the criteria leading to the design and planning of formal listening tests aimed at a subjective assessment of the impact of rendering artifacts on the perception of timbral quality and on localization of a virtual source. In particular, we show how we have employed standard recommendations from ITU-R, Broadcasting service (Sound) (BS). We have considered Recommendation ITU-R BS.1116 [12] and Recommendation ITU-R BS.1534 [13] and adapted them to our specific needs.

In Chapter 6 we present all simulations and experimental results in order to prove the effectiveness of our evaluation methodology. Moreover, we present a discussion of the results, aimed at highlighting the strong correlation between the results provided by our psychoacoustic-based metrics and the results of subjective tests.

Finally, in Chapter 7 we draw some conclusions and show possible future works.

Chapter 2

Background

Conventional sound reproduction techniques like stereophony and surround suffer from a sweet spot limitation, i.e. the desired sound field can be correctly reproduced inside a small listening area, thus limiting the number of listeners that can attend the acoustic performance in an optimal way. In this chapter we review state-of-the-art sound field rendering techniques like WFS and GR, which are aimed at a correct reproduction of the sound field in a larger listening area.

In its theoretical formulation WFS allows a correct reproduction of the desired sound field; however, a practical realization of a WFS system forbids the fulfillment of all its theoretical requirements. In particular, an array of loudspeakers is used to approximate a continuous distribution of sound sources; thus, some artifacts arise, produced by the spatial sampling of the wave fronts (due to the finite spacing between loudspeakers) and to truncation effect (due to the finite aperture of the array). Moreover, when working in a real environment WFS does not provide any compensation of the room acoustics, thus reflections will significantly alter the rendered sound field.

On the other hand, GR produces a sound field which is an approximation of the desired one in a least-squares sense over an arbitrary listening area, given the geometry of the environment, the reflective properties of the walls and the loudspeakers distribution. Thus, geometric rendering can compensate the room acoustics even in reverberant environments, but it introduces artifacts arising from the least-squares approximation of the desired sound field.

It becomes important to evaluate the artifacts introduced by rendering systems from a perceptual standpoint. However, a formal subjective assessment of such artifacts is costly, since it requires a large panel of trained listeners to produce reliable results. The alternative provided by the literature consists

only in objective approaches, which does not take into any account the human perception. A cost-effective methodology to assess the impact of artifacts on the perceptual performance of rendering systems is therefore in order.

The rest of this chapter is organized as follows. In Section 2.1 we present conventional sound reproduction technologies and highlight their limitations to motivate the introduction of more complex sound field rendering techniques. The mathematical and physical background necessary to represent sound field according to the physics of wave propagation is presented in Section 2.2. Then, in Section 2.3 we present WFS as a physically motivated sound field rendering technique. Furthermore, in Section 2.4 we motivate the adoption of a geometric approach to model sound propagation and we review GR as a sound field rendering technique based on this geometric model. Finally, in Section 2.5 we describe the artifacts introduced by sound field rendering systems and present the state-of-the-art approaches to evaluate the rendering quality.

2.1 Sound Reproduction

In this section we present an historical overview of conventional sound reproduction techniques. Finally, we illustrate the limitation of such techniques and motivate the introduction of more complex sound reproduction systems.

Sound reproduction is the process of reproducing sound waves (such as voice, singing, instrumental sounds and other sound effects) by means of electronic and electro-mechanical devices. The development of sound reproduction techniques, starting from the early work of Blumlein in 1930's, follows the intent of delivering a realistic sound scene.

From the early days of phonograph in the late-19th century, monophonic sound reproduction was the rule for almost all audio production scenarios. Typically, monophonic sound reproduction systems consist of one single loudspeaker or, in situations where an increased sound pressure is needed, multiple loudspeakers fed by a single signal. Such a system does not allow the listener to identify the position of the virtual sound source. On the contrary, the listener always perceives the sound as coming from the loudspeaker itself.

In 1931, Alan Blumlein developed two-channel recording methods [14] in the attempt of creating an illusion of directionality and sound scene perspective. These methods prescribe the use of two loudspeakers fed by independent signals. Such techniques are referred to as *two-channel stereophony* and nowadays they are commonly employed in entertainment applications (e.g. FM radio and TV broadcasting, popular music production).

The term *stereophony* also refers to more complex systems like *surround* systems, which employ a set of loudspeakers surrounding the listeners. Nowadays cinema and soundtracks are the major applications of surround techniques. Ambisonics can be considered as a further development of surround systems. In its original formulation Ambisonics is based on quadraphonic techniques and pursues local wave front reconstruction [15].

With both stereophonic and surround systems the correct reproduction of the sound scene is restricted to a narrow listening area, usually named *sweet spot*. Outside this area timbral and spatial distortions occur, due to comb filter effects and limited spatial extent of the local wavefronts produced.

In the next sections we show how current sound field synthesis techniques like WFS and GR overcome this sweet-spot limitation and we introduce the necessary mathematical and physical background.

2.2 Wave-Based Representation

In this section we discuss some mathematical and physical facts and their application to sound field synthesis. In particular, we derive the physical foundation of sound field synthesis techniques starting from the acoustic wave equation.

In general, sound fields are represented depending on one time coordinate t and three spatial coordinates. Since the scenario considered in this work is that of sound field synthesis restricted to a plane, we consider only two spatial coordinates that in Cartesian and polar coordinate systems are denoted by

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} \text{ and } \mathbf{r} = \begin{pmatrix} \rho \\ \theta \end{pmatrix}. \tag{2.1}$$

The polar coordinates are related to Cartesian coordinates by the following relationships:

$$\mathbf{x} = \rho \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \text{ and } \mathbf{r} = \begin{pmatrix} \sqrt{x^2+y^2} \\ \arctan(y/x) \end{pmatrix}. \tag{2.2}$$

The basic two-dimensional representation of acoustic phenomena is the *wave equation*:

$$\nabla^2 p(\mathbf{x},t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p(\mathbf{x},t) = 0, \quad \mathbf{x} \in S, \tag{2.3}$$

whose solutions are called *sound fields*. The wave equation describes the spatial distribution of the sound pressure $p(\mathbf{x},t)$ on a surface S and its variation

over time. Considering the case of no sound sources inside the surface S, the right-hand-side of Equation (2.3) is equal to zero. A complete derivation of the acoustic wave equation starting from basic physical laws is reported in [16].

A simple solution of Equation (2.3) are the plane-waves:

$$p(\mathbf{x}, t) = g\left(t + \frac{1}{c}\mathbf{n}^T\mathbf{x}\right), \tag{2.4}$$

where $\mathbf{n}$ is a unit length vector and the superscript ^{T} indicates vector transposition. By substituting Equation (2.4) into Equation (2.3) it can be proved that Equation (2.4) is a solution of the wave equation, thus Equation (2.4) represents a sound field. The terminology "plane waves" is justified by the fact that the argument of the function g represents a plane in space with normal vector $\mathbf{n}$. As time t increases, the plane waves propagates through space with speed c.

General solutions of the Equation (2.3) can be obtained by superposition of plane waves of different frequencies and directions: this property can be easily verified considering that Equation (2.3) is linear.

An useful representation of a sound field describes the sound pressure as a function of frequency and spatial position. In order to obtain this space-frequency representation, the Fourier transform with respect to time t is applied to $p(\mathbf{x}, t)$:

$$P(\mathbf{x}, \omega) = \mathcal{F}_t\left\{p(\mathbf{x}, t)\right\} = \int_{-\infty}^{+\infty} p(\mathbf{x}, t)e^{-j\omega t}\mathrm{d}t. \tag{2.5}$$

Applying the differentiation property of Fourier the following result is acheived:

$$\mathcal{F}_t\left\{p(\mathbf{x}, t)\right\} = -\omega^2 P(\mathbf{x}, \omega) \tag{2.6}$$

that turns the wave equation (2.3) into the *Helmholtz equation*

$$\nabla^2 P(\mathbf{x}, \omega) + k^2 P(\mathbf{x}, \omega) = 0, \tag{2.7}$$

where $k = \omega/c$ is the wave number.

When a sound source (or a distribution of sources) $P_0(\mathbf{x}, \omega)$ is present in the acoustic scene, the Helmholtz equation is modified to

$$L\left\{P(\mathbf{x}, \omega)\right\} = \nabla^2 P(\mathbf{x}, \omega) + k^2 P(\mathbf{x}, \omega) = P_0(\mathbf{x}, \omega), \tag{2.8}$$

where the differential operator $L(\,\cdot\,) = \nabla^2(\,\cdot\,) + k^2(\,\cdot\,)$ has been introduced.

A distribution of sources $P_0(\mathbf{x}, \omega)$ produces a sound pressure at position

$\mathbf{x}$ that can be described by integrating the *Green's function* $G(\mathbf{x}|\xi,\omega)$ for all sources location $\xi \in S$:

$$P(\mathbf{x},\omega) = \int_S G(\mathbf{x}|\xi,\omega) P_0(\xi,\omega) \mathrm{d}\xi. \tag{2.9}$$

The integration is performed within the surface S that encloses all sources $P_0(\mathbf{x},\omega)$ and all locations where an evaluation of the sound pressure $P(\mathbf{x},\omega)$ is of interest; typically the surface S is determined by the boundaries of a room.

To determine the Green's function the operator L from Equation (2.8) is applied to Equation (2.9)

$$L\left\{P(\mathbf{x},\omega)\right\} = \int_S L\left\{G(\mathbf{x}|\xi,\omega)\right\} P_0(\xi,\omega) \mathrm{d}\xi = P_0(\mathbf{x},\omega). \tag{2.10}$$

For the right equality to be valid, the Green's function must satisfy the wave equation (2.3) with a spatial impulse as inhomogeneity, where the inhomogeneity term is represented by

$$L\left\{G(\mathbf{x}|\xi,\omega)\right\} = \nabla^2 G(\mathbf{x}|\xi,\omega) + k^2 G(\mathbf{x}|\xi,\omega) = \delta(\mathbf{x}-\xi). \tag{2.11}$$

The solution of Equation (2.11) depends on the propagation medium and the boundary conditions; for free field sound propagation [17] the Green's function is given as

$$G(\mathbf{x}|\xi,\omega) = \frac{1}{4\pi} \frac{e^{-jk|\mathbf{x}-\xi|}}{|\mathbf{x}-\xi|}. \tag{2.12}$$

The Green's function can be regarded as an extension of the impulse response for multi-dimensional systems.

The approach outlined above requires the knowledge of the source distribution. In the case of unknown sources, we are forced to adopt a different approach based on *Huygens' principle*, which states that the wavefront of a propagating wave can be reconstructed considering a superposition of spherical waves radiated from every point on the wavefront at a prior instant. Formally, Huygens' principle is formulated as a superposition of contributions emitted by a distribution of sources on a virtual surface: this formulation is known as *Kirchhoff-Helmholtz integral* [18]:

$$P(\mathbf{x},\omega) = \oint_{\partial S} \left(\frac{\partial}{\partial \mathbf{n}} G(\mathbf{x}|\xi,\omega) P(\xi,\omega) - G(\mathbf{x}|\xi,\omega) \frac{\partial}{\partial \mathbf{n}} P(\xi,\omega) \right) \mathrm{d}\xi. \tag{2.13}$$

An intuitive explanation of Kirchhoff-Helmholtz integral is the following: if

the listening area S is free of sound sources, then the sound pressure $P(\mathbf{x}, \omega)$ is determined by the values of the sound pressure and its gradient on the surface ∂S. Thus any physically meaningful sound field in the listening area S can be generated by loudspeakers placed on the boundary ∂S.

2.3 Wave Field Synthesis

In this section we introduce Wave Field Synthesis (WFS) as a physically motivated sound field synthesis technique. In particular we show that its theoretical derivation is based on the Kirchhoff-Helmholtz integral. Then we illustrate an approach to simplify the implementation of a WFS rendering system adopting only monopole secondary sources. Furthermore we introduce the concept of spatial sampling as one origin of the artifacts in the rendered sound field. In the conclusion of this section we propose a discussion of the approaches commonly adopted to determine loudspeakers driving signals.

Wave Field Synthesis (WFS) is a sound field synthesis technique that uses an array of loudspeakers to reproduce a sound field; it is formulated in terms of the acoustic wave equation and Green's functions [2]. The Kirchhoff-Helmholtz integral (Equation (2.13)) provides the theoretical basis for this sound reproduction technique.

The free-field Green's function, given by Equation (2.12) can be interpreted as the sound field of a monopole point-like source distribution on the boundary ∂S. But the Kirchhoff-Helmholtz integral (Equation (2.13)) also involves the directional gradient of the Green's function, which can be interpreted as the field of a dipole source lying in the direction of the normal vector $\mathbf{n}$. Thus, the interpretation of Kirchhoff-Helmholtz integral can be restated as: the acoustic pressure inside an area S can be controlled by distribution of monopole and dipole point-like sources on the boundary ∂S.

This interpretation allows to outline a technical system for sound reproduction consisting of appropriate loudspeakers approximating monopole and dipole sources. These loudspeakers are placed on the boundary of a listening area enclosing the possible listener positions and they are driven by appropriate signals to reproduce the desired sound field.

2.3.1 Monopole and Dipole Sources

According to Kirchhoff-Helmholtz integral (Equation (2.13)), the use of monopole and dipole sources allows a precise reproduction of the desired sound field,

which is recreated as $P(\mathbf{x}, \omega)$ for all positions inside S and it is zero outside. Usually this condition is not required: in a practical sound reproduction system an arbitrary sound field outside S can be tolerated, as long as its sound intensity is moderate and the reproduction inside S is not impaired. Thus, only one type of sound sources can be used and some sound pressure outside S is tolerated.

For a practical implementation, it is more convenient to discard dipole sources, since monopoles can be well approximated by small loudspeakers in closed cabinets; instead dipole speakers are constructed by mounting a loudspeaker on a flat panel, thus they are far less efficient (considering the same driver). According to [17, 19], one technique to derive a monopole-only version of the Kirchhoff-Helmholtz integral consists of a modification of the free-field Green's function.

In particular, the contribution of dipole sources can be discarded by adopting a modified Green's function that has to obey the following condition

$$\frac{\partial}{\partial \mathbf{n}} G_N(\mathbf{x}|\xi, \omega) = 0. \tag{2.14}$$

This modified Green's function is usually named *Neumann Green's function*, because Equation (2.14) formulates a homogeneous Neumann boundary condition imposed on ∂S [19]. To derive the desired Neumann Green's function, it is noticed that the condition imposed by Equation (2.14) is satisfied by the superposition of two free-field Green's functions:

$$G_N(\mathbf{x}|\xi, \omega) = G(\mathbf{x}|\xi, \omega) + G(\bar{\mathbf{x}}(\mathbf{x})|\xi, \omega), \tag{2.15}$$

where the position $\bar{\mathbf{x}}(\mathbf{x})$ is chosen as the mirror image of $\mathbf{x}$ with respect to the tangent plane in ξ on the surface ∂S [19]. Furthermore, on the boundaries we can write [1]

$$G_N(\mathbf{x}|\xi, \omega) = 2G(\mathbf{x}|\xi, \omega), \tag{2.16}$$

hence $G_N(\mathbf{x}|\xi, \omega)$ is equal to a point source with double strength.

Inserting $G_N(\mathbf{x}|\xi, \omega)$ from Equation (2.16) as Green's function into the Kirchhoff-Helmholtz integral (2.13) leads to

$$\begin{aligned} P(\mathbf{x}, \omega) &= -\oint_{\partial S} G_N(\mathbf{x}|\xi, \omega) \frac{\partial}{\partial \mathbf{n}} P(\xi, \omega) \mathrm{d}\xi \\ &= -\oint_{\partial V} 2G(\mathbf{x}|\xi, \omega) \frac{\partial}{\partial \mathbf{n}} P(\xi, \omega) \mathrm{d}\xi \end{aligned} \tag{2.17}$$

for $\mathbf{x} \in S$, while outside S the sound field consists of a mirrored version of

the sound field inside S. The result of Equation (2.17) is known as the type-I Raleigh integral [20] and it describes the sound field $P(\mathbf{x}, \omega)$ inside an area S generated by a distribution of monopole sources placed on the boundary ∂S.

The two-dimensional free-field Green's function is given by [17]

$$G(\mathbf{x}|\xi, \omega) = \tilde{G}(\rho, \omega) = -\frac{j}{4} H_0^{(2)} \left(\frac{\omega}{c}\rho\right) \tag{2.18}$$

with the Hankel function of the second kind and order zero [21]

$$H_0^{(2)}(u) = J_0(u) - jN_0(u), \tag{2.19}$$

where $J_0(u)$ and $N_0(u)$ denote, respectively, the Bessel and Neumann functions of first kind and order zero [22] and

$$\rho = \|\mathbf{x} - \xi\| = \sqrt{(x - \xi_x)^2 + (y - \xi_y)^2} \tag{2.20}$$

is the distance between the listener position $\mathbf{x}$ and the source position ξ.

Due to circular symmetry, $G(\mathbf{x}|\xi, \omega)$ depends only on the distance ρ, thus it can be replaced by $\tilde{G}(\rho, \omega)$ (Equation (2.18)). For large values of ρ (i.e. $(\omega/c)\rho \gg 1$) the far-field approximation of the Hankel function [22] is adopted

$$H_0^{(2)} \left(\frac{\omega}{c}\rho\right) \approx \sqrt{\frac{2j}{\pi \left(\frac{\omega}{c}\rho\right)}} e^{-j\left(\frac{\omega}{c}\rho\right)}. \tag{2.21}$$

The far-field approximation of the Hankel function is used to approximate the two-dimensional Green's function as follows

$$\tilde{G}(\rho, \omega) = H(\omega) A(\rho) \frac{1}{4\pi} \frac{e^{-j\left(\frac{\omega}{c}\rho\right)}}{\rho}, \tag{2.22}$$

where the two terms

$$H(\omega) = \sqrt{\frac{c}{j\omega}} \quad \text{and} \quad A(\rho) = \sqrt{2\pi\rho} \tag{2.23}$$

are, respectively, the frequency-dependent term (i.e. it causes spectral modification) and a space-dependent term that causes amplitude modification depending on the distance.

The substitution of Equation (2.12) into the Raleigh-I integral (2.17) leads

to

$$P(\mathbf{x},\omega) = -\oint_{\partial S} G_0(\mathbf{x}|\xi,\omega) D(\mathbf{x}|\xi,\omega) \mathrm{d}\xi \tag{2.24}$$

for $\mathbf{x} \in S$, where

$$D(\mathbf{x}|\xi,\omega) = 2A(\rho)H(\omega)\frac{\partial}{\partial \mathbf{n}} P(\xi,\omega). \tag{2.25}$$

Equation (2.25) describes the wave propagation in a three-dimensional space with listener locations $\mathbf{x}$ assumed to lie in the $x-y$-plane. $D(\mathbf{x}|\xi,\omega)$ denotes the monopole source signals.

2.3.2 Spatial Sampling

For a practical implementation, the spatially continuous source distribution of Equation (2.24) has to be replaced by an arrangement of a finite number of loudspeakers with a monopole-like directivity. The resulting sound field is obtained replacing the integral in Equation (2.24) by a sum over the loudspeaker positions ξ_n

$$P(\mathbf{x},\omega) \approx -\sum_n G_0(\mathbf{x}|\xi_n) D(\mathbf{x}|\xi_n,\omega) \Delta\xi_n, \tag{2.26}$$

where $\Delta\xi_n$ is the distance between loudspeakers (which are not required to be equidistant). The effect of finite spacing of the loudspeakers can arise as artifacts due to spatial sampling. An anti-aliasing condition has been derived in [4] as

$$f \leq \frac{c}{\Delta\xi_n(1+|\cos\alpha_{\mathrm{pw}}|)}, \tag{2.27}$$

where $0 \leq \alpha_{\mathrm{pw}} < \pi$ is the incidence angle of the virtual plane waves.

2.3.3 Determination of the Loudspeakers Driving Signals

The driving signals for the loudspeakers at positions ξ_n are obtained from Equation (2.25) by inverse Fourier transformation with respect to ω. These signals depend on the position of the listener with the term $A(\rho)$. In practical situations, this amplitude modulation term is set to a fixed position inside the area S: in this way, the loudspeaker signals are independent of the listener's position. It is remarked that the frequency compensation $H(\omega)$ and $\frac{\partial}{\partial \mathbf{n}} P(\xi_n,\omega)$ do not depend on any information about the listener, so their calculation can be performed correctly for all positions inside S.

A crucial point in the determination of the loudspeaker signals is the value of the gradient $\frac{\partial}{\partial \mathbf{n}} P(\xi_n,\omega)$ at the loudspeaker positions. A naive solution is to measure the original sound field with properly positioned and oriented

second-order microphones. However, this approach is very limiting. More versatile approaches are:

- *model-based approach:* the pressure gradient is determined from a model of the acoustic scene (e.g. free-field propagation or more elaborate models that take into account the room acoustics) [23];
- *data-based approach:* the spatial characteristics of the room acoustics in the recording environment are determined by microphone measurements, in particular, a set of room impulse responses is derived for use in Wave Field Synthesis reproduction [24].

2.4 Geometric Representation

In this section we motivate the adoption of a geometric-based representation for the description of a sound field by means of the *Eikonal equation* [25]. Then we present Geometric Rendering (GR) as a sound field rendering system based on the geometrical modeling of sound propagation.

In the following we adopt the symbol S to denote the Fourier transform of the sound source signal. The context will make clear if the actual occurrence of the symbol S denotes the source signal or the listening area.

According to [26] the eikonal equation is derived starting from the Fourier transform of a solution of acoustic wave equation (2.3)

$$P(\mathbf{x}, \omega) = S(\omega)A(\mathbf{x}, \omega)e^{j\omega T(\mathbf{x},\omega)}, \tag{2.28}$$

where $S(\omega)$ is the Fourier transform of the source signal, $A(\mathbf{x}, \omega)$ is an amplitude term depending on position and frequency and the phase $T(\mathbf{x}, \omega)$ depends on position and frequency. The phase term is usually approximated as $T(\mathbf{x})$ (i.e. neglecting the frequency dependence) since, this way, the existence of a wavefront is made implicit. The phase function $T(\mathbf{x})$ is called *eikonal*.

The amplitude term $A(\mathbf{x}, \omega)$ can be approximated by separating the dependency from position and frequency

$$A(\mathbf{x}, \omega) = A_0(\mathbf{x}) + \frac{A_1(\mathbf{x})}{j\omega} + \frac{A_2(\mathbf{x})}{(j\omega)^2} + \dots. \tag{2.29}$$

At high frequencies, all terms successive to $A_0(\mathbf{x})$ (which are inversely proportional to frequency) can be neglected; thus Equation (2.28) becomes

$$P(\mathbf{x}\omega) = S(\omega)A_0(\mathbf{x})e^{j\omega T(\mathbf{x})}, \tag{2.30}$$

which in space-time domain becomes

$$p(\mathbf{x}, t) = A_0(\mathbf{x}) s(t - T(\mathbf{x})). \tag{2.31}$$

Equation (2.31) states that the sound field $p(\mathbf{x}, t)$ is constituted by replicas of the signal $s(t)$ delayed by travel time $T(\mathbf{x})$, propagated without distortions.

The substitution of Equation (2.30) into the Helmholtz equation (2.7) results in

$$\nabla P = S \nabla A_0 e^{j\omega T} + S A_0 j\omega \nabla T e^{j\omega T}, \tag{2.32}$$

and

$$\begin{aligned} \nabla^2 P = {} & S \nabla^2 A_0 e^{j\omega T} + S \nabla A_0 j\omega \nabla T e^{j\omega T} + j\omega S \nabla A_0 \nabla T e^{j\omega T} + \\ & + j\omega S A_0 \nabla^2 T e^{j\omega T} - \omega^2 S A_0 (\nabla T)^2 e^{j\omega T}. \end{aligned} \tag{2.33}$$

After eliminating common terms and eliminating S, the following equations are obtained:

$$(\nabla T(\mathbf{x}))^2 - \frac{1}{c^2(\mathbf{x})} = 0; \tag{2.34}$$

$$2A_0(\mathbf{x}) \cdot \nabla A_0(\mathbf{x}) \cdot \nabla T(\mathbf{x}) + A_0^2(\mathbf{x}) \cdot \nabla^2 T(\mathbf{x}) = 0; \tag{2.35}$$

$$\nabla^2 A_0(\mathbf{x}) = 0. \tag{2.36}$$

Recalling that the phase T depends both on position and frequency, we notice that Equation 2.34 depends on ω^2, while Equation 2.35 depends on ω and Equation 2.36 does not depend on frequency. Thus, for high frequencies the linear and constant terms (Equation 2.35 and 2.36) can be neglected. Equation (2.34) governs the propagation: it is called *Eikonal equation* and it is a particular approximation of the Helmholtz's equation valid only at high frequencies. Its solutions are called *rays*.

Geometric methods aimed at solving the wave equation 2.3 are based on the high frequency approximation provided by the Eikonal equation. A geometric representation of sound propagation presents several advantages over a wave-based representation. In particular, to obtain the sound field generated by a virtual source in an enclosure by means of the wave-based representation we are forced to find approximate solutions of the acoustic wave equation (2.3) with Finite Element Method (FEM), which computational complexity is considerable. On the other hand, a geometric representation allows an efficient modeling of sound propagation employing techniques like ray tracing [27], radiosity [28], image source method [29] and beam tracing [30].

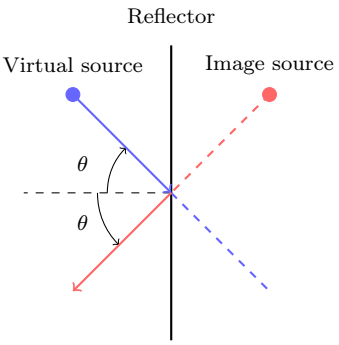


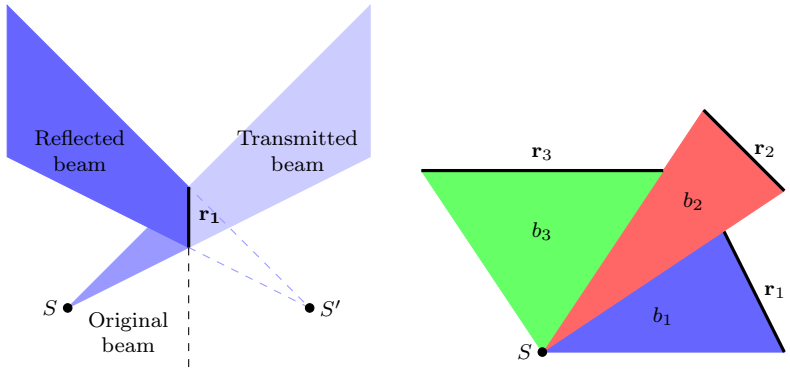
Figure 2.1: Generation of a reflective ray explained according to the image source theory.

2.4.1 Representation of the Sound Field as Superposition of Beams

In the previous paragraph we have discovered that the circular wave front can be locally represented by a ray, which is, as a matter of fact, the vector orthogonal to the wave front. For plane waves, this representation holds for all the points in space. On the other hand for circular wave fronts a complete description of the wave front requires that rays are defined for all the points in space. As a consequence, the rays can not define a compact representation of the sound field, because the representation of a single wave front requires the use of an infinite number of rays, one for each looking direction. For this reason in the literature other more compact representations of wave fields have been developed, which leverage on rays. Among the others, in this work we focus on a representation based on *acoustic beams*, which is going to be discussed in the next few paragraphs.

When rays encounter an obstacle along their propagation, their energy is transferred to transmitted and reflected rays. We neglect the presence of transmitted rays. Under the hypothesis of planar obstacles, which is assumed to be valid throughout this thesis, the reflected ray travels on a direction which is specular with respect to the incoming ray (i.e. Snell's law). The generation of the reflective ray can also be explained using the *image source* theory, which explains the arise of the reflective ray as a ray generated from a source whose position is mirrored with respect to the wall, as it is depicted in Figure 2.1.

As a matter of fact when multiple reflectors are present in the environment, there is a plurality of rays that bounce multiple times over the walls. In order to compactly represent the wave field, the concept of acoustic beam has been introduced. An acoustic beam is a bundle of rays starting from the same source position and illuminating the same reflector, as depicted in Figure 2.2a. In this way, the acoustic path (i.e. ray) that links the source and the receiver



(a) *Reflected and transmitted beams generated by a point-like source S* and a reflector $\mathbf{r}_1$.

(b) *Acoustic beams traced from the source location and illuminating reflectors $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$.*

Figure 2.2: Acoustic beams: illumination of a reflector (Figure 2.2a) and separation of beams (Figure2.2b).

is a subset of the beam. In fact, the beam can be conceived as the visibility of a region from a point (i.e. the source position). When a beam, during its propagation, meets an obstacle it is reflected. The position of the image source can be easily predicted using the image source theory introduced above. The bundle of rays coming from the reflection of the acoustic beam, on the other hand, required a further subdivision into beams, each characterized by illuminating a single reflector, as shown in Figure 2.2b.

This reflection and branching procedure is iterated until the beams die out. It has been developed a model of the wave field that takes inspiration from the beams reflection procedure. The whole wave field, in fact, is conceived as a superposition of acoustic beams, each characterized by a virtual source, an orientation and an aperture.

2.4.2 Beam Tracing

Beam tracing is an efficient geometric solution to the modeling of sound propagation based on acoustic beams. This method was originally developed in [31] for image rendering applications and later it has been extended in [30] to audio rendering.

In the last paragraph we have described the process of splitting/branching of beams: the beam tracing method organizes and encodes this process into a specialized data structure called *beam tree*. The construction of the beam tree is an iterative process based on visibility evaluation.

When the receiver is specified, the paths linking source and receiver can be determined exploiting only the informations stored in the beam tree, through

a lookup of the data structure. Thus, the beam tracing approach enables a real-time rendering of sounds in complex environments even when receivers are moving.

In [32] the authors show a method for constructing the beam tree through a lookup on a precomputed data structure called *global visibility function*, which describes the visibility of a region as a function of the viewing angle and the source location. A *visibility diagram* is used, which is a re-mapping of geometric primitives and functional elements (rays, beams, reflectors, sources, receivers, etc...) onto the *ray space*. This change of parametrization allows to reduce the cost of the beam splitting operations from $\mathcal{O}(n^3)$ (for traditional beam tracing) to $\mathcal{O}(n)$, where n is the number of reflectors. In this latter case, the costly operation is the determination of the global visibility but, since at this stage the representation is independent from the source and receiver locations, it can be accomplished in off-line mode.

2.4.3 Geometric Rendering Engine

Consider the problem of rendering the acoustics of a virtual environment that, as we have seen in the last paragraphs, can be modeled as a superposition of beams. These acoustic beams are synthesized through a proper spatial filtering of the signal fed to the loudspeakers.

Exploiting the whole information gathered from the geometry of the environment (i.e. the position and orientation of the reflectors), in [3] the authors show a method to render sources in both near and far field. They consider the case of rendering an acoustic beam through a distribution of loudspeakers. The beam has origin in $\mathbf{s}$, angular aperture ϕ and orientation θ. The sound field is rendered inside an area enclosing a set of N control points $\mathbf{a}_1, \ldots, \mathbf{a}_N$, while the M loudspeakers are located t $\mathbf{p}_1, \ldots, \mathbf{p}_M$. A radiance beam pattern $\Theta(\theta, \phi)$ is associated to the source and it is a function of the beam orientation and aperture.

The contribution of the m-th loudspeaker to the sound field in the control point $\mathbf{a}_n$ is

$$\Psi_{m,n} = H_m g(\mathbf{p}_m, \mathbf{a}_n), \tag{2.37}$$

where H_m is the coefficient applied to the signal emitted from the m-th loudspeaker and $g(\mathbf{p}_m, \mathbf{a}_n)$ is the Green's function (2.12) from loudspeaker m to the control point $\mathbf{a}_n$. The sound field in $\mathbf{a}_n$ is the sum of all signals from

the loudspeakers

$$\Psi_n = \sum_{m=1}^{M} \Psi_{m,n} = \sum_{m=1}^{M} H_m g(\mathbf{p}_m, \mathbf{a}_n). \tag{2.38}$$

The author's goal is to render the acoustic beam emitted by a virtual source placed in $\mathbf{s}$, oriented toward the direction θ and with an angular aperture ϕ. Thus, the desired response in the point $\mathbf{a}_n$ can be written as

$$\bar{\Psi}_n = g(\mathbf{s}, \mathbf{a}_n)\Theta(\theta, \phi, \alpha_n), \tag{2.39}$$

where $\Theta(\theta, \phi, \alpha_n)$ is the value of the radiation pattern of the virtual source at point $\mathbf{a}_n$ (α_n is the angle under which the n-th listening point is seen from $\mathbf{s}$).

The sound field is obtained imposing that the spatial response of the array approximates the spatial response of the virtual source (i.e. $\Psi_n = \bar{\Psi}_n$). In particular:

$$\mathbf{g}_n^T \mathbf{h} = g(\mathbf{s}, \mathbf{a}_n)\Theta(\theta, \phi, \alpha_n), \tag{2.40}$$

where $\mathbf{h} = [H_1, H_2, \ldots, H_M]^T$ is the coefficient vector and $\mathbf{g}_n = [g(\mathbf{p}_1, \mathbf{a}_n), g(\mathbf{p}_2, \mathbf{a}_n), \ldots, g(\mathbf{p}_M, \mathbf{a}_n)]^T$ is the juxtaposition of the Green's functions from the m-th loudspeaker to the control point $\mathbf{a}_n$. Considering all control points at once, the following matrix formulation is obtained

$$\mathbf{G}\mathbf{h} = \mathbf{r_d}, \tag{2.41}$$

where $\mathbf{r_d} = [g(\mathbf{s}, \mathbf{a}_1)\Theta(\theta, \phi, \alpha_1), \ldots, g(\mathbf{s}, \mathbf{a}_N)\Theta(\theta, \phi, \alpha_N)]^T$ is the desired response and $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_N]^T$ is the $N \times M$ propagation matrix from each loudspeaker to each control point. In order to obtain a smooth beam pattern $N \gg M$ is used.

The system of Equation (2.41) is over-determined and it does not admit an exact solution. An estimation $\hat{\mathbf{h}}$ of the vector $\mathbf{h}$ can be calculated by introducing the pseudo-inverse operation on the matrix $\mathbf{G}$

$$\mathbf{G}^\dagger = (\mathbf{G}^H \mathbf{G})^{-1} \mathbf{G}^H, \tag{2.42}$$

where $\mathbf{G}^H$ denotes the Hermitian conjugate of the matrix $\mathbf{G}$. The loudspeaker coefficients are approximated by

$$\hat{\mathbf{h}} = \mathbf{G}^\dagger \mathbf{r_d} = (\mathbf{G}^H \mathbf{G})^{-1} \mathbf{G}^H \mathbf{r_d}. \tag{2.43}$$

In general $\mathbf{G}\hat{\mathbf{h}} \neq \mathbf{r_d}$, but $\hat{\mathbf{h}}$ represents the best solution of the problem in a least squares sense.

In order to avoid instability issues due to the possible bad conditioning of $(\mathbf{G}^H\mathbf{G})$, a reconditioning through a Singular Value Decomposition (SVD) is needed

$$\mathbf{G}^H\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{2.44}$$

where $\mathbf{U}$ and $\mathbf{V}$ are, respectively, the left and right singular vectors and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_M)$ is the singular value diagonal matrix. The greatest index k that guarantees $\sigma_k/\sigma_1 \geq 0.01$ is identified and the first k columns and rows of matrices $\mathbf{U}$,$\mathbf{V}$ and $\mathbf{\Sigma}$ are retained. Therefore, the approximate inverse matrix is

$$(\mathbf{G}^H\mathbf{G})^{-1} \approx \mathbf{V}_k\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^H. \tag{2.45}$$

The inversion of $\mathbf{G}^H\mathbf{G}$ is a costly operation but it can be precomputed off-line once the positions of loudspeakers and control points are known.

In [3] the authors show how this approach can be extended to wide-band signals and to multiple beams, preserving the spatial Nyquist criterion, which means that the maximum operating frequency is limited by the loudspeakers reciprocal distance:

$$f_{\max} < \frac{c}{2d}, \tag{2.46}$$

where d is the distance between emitters.

2.5 Evaluation of Rendering Quality

In this section we describe the artifacts arising in sound field rendering systems, separating the timbral artifacts (i.e. the artifacts that produce a distortion of the timbre) from spatial artifacts (i.e. artifacts that produce an erroneous localization of the virtual source). Then, we review the state-of-the-art approaches for the evaluation of the audio quality of a rendered sound field. We identify two kinds of approaches in the literature, i.e. objective and subjective approaches.

The practical realization of sound field rendering systems like WFS (Section 2.3) and GR (Section 2.4) always imposes some constraints. In particular, the loudspeaker array only approximates a spatially continuous distribution of sound sources because it is limited in length and the spacing between loudspeakers is finite. Another constraint is that the radiance pattern of the loudspeakers is not purely omnidirectional at all frequencies in the audio range;

thus, the behavior of the loudspeakers deviates from the one of an ideal acoustic monopole. Other technical constraints are imposed by the emission system (i.e. D/A converter, power amplifiers and loudspeakers), which is limited in bandwidth and whose frequency response is not perfectly flat.

In [4] the authors show that such approximations of the ideal system cause a deviation of the rendered sound field from the desired one. In particular, the spatial aliasing and the truncation effect cause artifacts that are perceivable by a human listener.

In particular, considering the case of rendering an impulsive sound field we notice that the rendering systems exhibit a well defined direct wave front at all positions inside the listening area. In addition to this direct wave front, the loudspeaker array emits also secondary wave fronts immediately before and after the direct one, which are named *pre-echoes* and *post-echoes*, respectively [5]. The time interval between successive echoes is short, so these replicas are not perceived as separate acoustic events but they are perceived by a human listener as a timbral distortion of the rendered sound signal.

Furthermore, it is shown in [4] that the wave fronts produced by a rendering system are distorted in such a way that they impact on spatial acoustic cues (ILD and ITD), causing an erroneous localization of the sound source.

The literature presents two classes of approaches aimed at assessing the impact of sound field distortions on the human perception. On the one hand we have objective approaches, which prescribe the measurement of specific features of the rendered sound field in order to evaluate the differences from an ideal behavior. On the other hand we have subjective approaches, which are based on formal listening tests in which a listener is asked to make judgments on specific features of the sound stimuli which are offered to him.

2.5.1 Objective Evaluation

An objective methodology to evaluate the accuracy of sound field rendering techniques is presented in [10]. This evaluation methodology distinguishes between the target sound field, which is the desired outcome of the reproduction; the theoretical sound field, which is an approximation of the target sound field with a discrete loudspeaker array driven by a specific rendering technique; the measured sound field which results from acoustic measurements in the room where the rendering system is operating.

The work in [10] introduces two evaluation metrics based on root means square error. Such metrics can be used to directly compare results related to

different rendering techniques and different distributions of virtual sources, since a normalization of the sound field is performed. Moreover, the normalization makes the results of the metrics independent of the loudspeakers volume and of the gain of the microphone used for acoustic measurements.

An alternative metric is presented in [33]; it is based on root mean square error of the Fourier series angular coefficients of the sound field. This contribution presents a generalized RMSE metric for the difference between the target, the theoretical and the measured sound field. The authors also provide a generalization of this metric that allows to restrict the evaluation to a specific range of frequencies of interest.

Since they make no assumptions on the specific rendering technique adopted, the methodologies presented in [10] and [33] are general enough to be employed with many rendering techniques. However, the main drawback of these two contributions is that they do not relate the presented metrics with perceptual considerations.

One attempt to relate the sound field distortions to psychoacoustic effects is presented in [34], where the authors evaluate the perception of distance of a virtual source in a sound scene rendered with a WFS system. In particular, ILD is exploited as a distance cue ([8]), relying on the ability of ILD to reproduce the curvature of a wave front. The authors also describe experiments and simulations to compare the ILD at the listeners ears with a real source and a virtual source. In particular, the authors concentrate on the case of *focused sources*, which are virtual sources closer to the listener than the loudspeakers.

2.5.2 Subjective Evaluation

Subjective evaluation of sound field rendering techniques is based on the analysis of results coming from formal listening tests.

In some cases, a subjective evaluation have been presented to validate the results of an objective evaluation. This is the case of the psychoacoustic analysis presented in the last paragraph ([34] and its companion paper [7]), where authors describe formal listening tests to evaluate the impact of sound field distortions on spatial acoustic cues like ILD.

A more relevant contribution can be found in [7], which is representative of how to conduct listening tests for sound field rendering evaluation in real-world conditions. In particular, the authors evaluate the impact of several parameters on the perceptual performance of WFS: cardinality and spacing of the loudspeaker array, position of the virtual source, position of the listener

relative to the array.

The authors present two different test methods to evaluate the overall audio quality and the spatial impression. In order to evaluate the overall audio quality, the test method is derived from Recommendation ITU-R BS.1534 [13] and listeners are asked to rate the sound quality according to ITU-R scale. On the other hand, the spatial impression is evaluated by asking the listeners to indicate the perceived position of the virtual sound source.

In this contribution the high cost of the listening tests is mitigated by introducing some relaxation of the original prescriptions. In particular, the test method to evaluate the overall audio quality prescribes no reference nor anchor, in contrast with the Recommendation ITU-R BS.1534 [13] from which it has been derived. Another issue that highly influences the cost of subjective assessment is the size of the test panel (i.e. the number of test subjects), but authors do not present any information regarding the number of listeners involved.

2.6 Conclusions

In this chapter we have shown that sound field rendering techniques like WFS and GR can not reproduce a free-of-artifacts sound field in real-world conditions. In WFS artifacts arise from the practical implementation of the rendering system, which generates spatial aliasing and truncation effects and, moreover, WFS can not compensate the acoustics of the environment in which it is operating. On the other hand, the soundfield reproduced by GR comes as an approximation of the desired sound field taking room acoustics into account, but the least-squares approximation introduces some artifacts.

We have shown that such artifacts impair the perceptual performance of a sound field rendering system. In particular, they introduce a timbral distortion (due to the presence of secondary wave fronts emitted by the rendering system) and an alteration of the spatial impression (due to the distortion of the shape of the wave fronts).

We have presented two classes of approaches aimed at assessing the impact of sound field distortions on the human perception. Objective approaches are important to provide an indication of the overall quality of a sound field rendering technique but they are not informative enough since they do not take human perception into account. On the other hand, subjective approaches offer results that are intrinsically mediated by human perception, but their high cost in terms of working time and people involved prevents the adoption

of such approaches on a large scale.

The lack of a cost-effective methodology to assess the impact of artifacts on the perceptual performance of rendering systems motivates the introduction of our psychoacoustic-based metrics, presented in Chapter 4.

Chapter 3

Sound Field Measurement

The evaluation of timbral artifacts and spatial impression requires the knowledge of the room impulse response, which describes the sound field inside a listening area. The room impulse response is required because it visualizes explicitly the presence of pre-echoes and post-echoes and it allows to visualize the spatial distortion of the wave fronts.

A naive approach to measure the room impulse response would be to place a large number of microphones inside the listening area and recording the microphone signals generated by the emission of a suitable sound field. This approach has some obvious drawbacks: it requires the use of a very large number of D/A converters, preamplifiers and microphones, whose positioning inside the listening area will cause scattering effects.

The measurement methodology presented in this chapter allows to measure the sound field with only one microphone mounted on a rotating rig. This microphone is used to sample the sound field over a large number of positions on a circumference, without altering the sound field with the immersion of many microphones. The sound field acquired this way is then decomposed into *circular harmonics* and extrapolated on the whole listening area.

If the measurements of the sound pressure were ideal, it could be possible to get an exact extrapolation of the sound field. However, due to non-idealities of the measurements (temporal and spatial sampling, sensor noise, positioning error), only an approximation of the sound field quantities can be obtained. The methodology presented in this chapter is not immune to these non-idealities but it is robust against the propagation of error introduced by these non-idealities.

The rest of this chapter is organized as follows. In Section 3.1 we present the decomposition of a sound field into circular harmonics and describe how this decomposition can be exploited to perform an accurate measurement of a sound

field. Moreover, we discuss the drawbacks of a direct implementation of the CHD and present a practical solution. The principles behind the employment of a virtual microphone array to sample the sound field are presented in Section 3.2. In Section 3.3 we describe a whole methodology to measure the space-time impulse response of a sound field rendering system and we introduce an alternative representation of the sound field in the space-frequency domain.

3.1 Circular Harmonics Decomposition

In this section we describe the decomposition of a sound field into circular harmonics and exploit this representation to extrapolate the sound field in spatial positions different from the locations of the sensors. Moreover, we discuss a numerical issue arising from a direct implementation of such principles and present a practical solution which allows to solve this issue.

In order to obtain a modal representation of the sound field, we need to adopt a polar coordinate system. Such representation can then be expanded into Fourier series to obtain a modal representation by *circular harmonics* [9].

We can observe that the function which are solutions of the acoustic wave equation 2.3 are restricted to a specific class of functions. Taking into account the special nature of acoustic signals, it has been shown in [9] that the sound field in a whole listening area can be reconstructed starting from a relatively small number of sampling points on a circumference.

In order to obtain a modal representation, we need to adopt a polar coordinate system (radius ρ and angle ϕ). In this representation the sound pressure and its Fourier transform are described by

$$p(t, \rho, \phi) \rightarrow P(\omega, \rho, \phi). \tag{3.1}$$

Thanks to the periodicity of $P(\omega, \rho, \phi)$ with respect to the angle ϕ ([9]), it can be written as a Fourier series with angular coefficients $\mathring{P}_\mu(\omega.\rho)$

$$\mathcal{S}_\phi \left\{ P(\omega, \rho, \phi) \right\} = \mathring{P}_\mu(\omega, \rho) = \frac{1}{2\pi} \int_0^{2\pi} P(\omega, \rho, \phi) e^{-j\mu\phi} \mathrm{d}\phi, \tag{3.2}$$

and conversely

$$\mathcal{S}_\mu^{-1} \left\{ \mathring{P}_\mu(\omega, \phi) \right\} = P(\omega, \rho, \phi) = \sum_{\mu=-\infty}^{\infty} \mathring{P}_\mu(\omega, \rho) e^{j\mu\phi}. \tag{3.3}$$

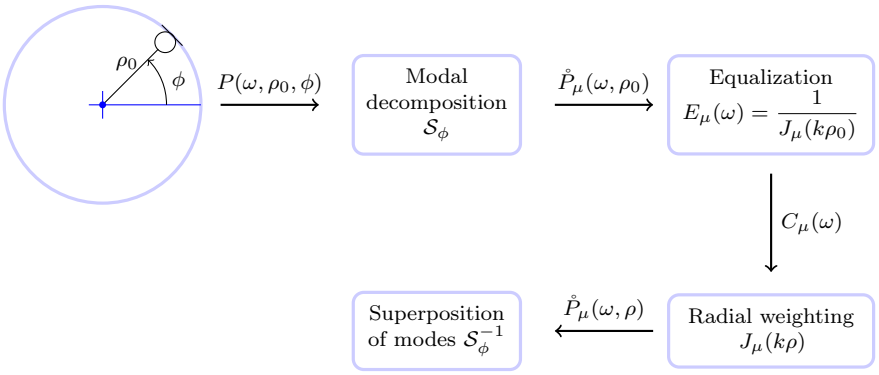


Figure 3.1: Principle of 2D decomposition and extrapolation of sound fields from measurements on a circular aperture [9]. Notice that the measurements have to be performed only for one radius ρ_0, while the extrapolation is possible for all $\rho > 0$ as long as the extrapolation area is free of sources. This picture has been adapted from [9].

If we consider a spatial region free of sources, we can write the angular coefficients as

$$\mathring{P}_\mu(\omega, \rho) = C_\mu(\omega) J_\mu(k\rho), \tag{3.4}$$

where $C_\mu(\omega)$ is the μ-th circular harmonic at frequency ω, k is the wavenumber and $J_\mu(\,\cdot\,)$ is the Bessel function of first kind and order μ.

In order to compute the coefficients of the circular harmonics, a straightforward way is to implement Equation (3.4), considering that circular harmonic coefficients do not depend on the radius ρ. Starting from the measurements of the sound pressure on a circle with radius ρ_0 (Figure 3.1 top left), the circular harmonic coefficients are obtained as

$$C_\mu(\omega) = \frac{1}{J_\mu(k\rho_0)} \mathring{P}_\mu(\omega, \rho_0) = E_\mu(\omega) \mathring{P}_\mu(\omega, \rho_0) \tag{3.5}$$

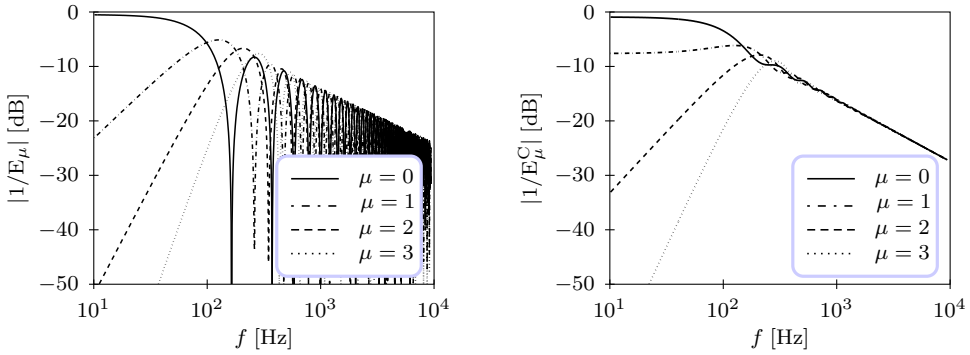
where we have introduced the equalization function

$$E_\mu(\omega) = \frac{1}{J_\mu(k\rho_0)} \tag{3.6}$$

as depicted in the top row of Figure 3.1.

In theory the circular harmonics coefficients $C_\mu(\omega)$ can be derived exactly from Equation (3.5) and they contain the full information about the two-dimensional sound field under analysis. Thus, they could be used to calculate the sound pressure at any position (ρ, ϕ) (second row of Figure 3.1).

However, a direct implementation of Equation (3.5) is not possible due to the non-idealities enumerated at the beginning of this chapter. Furthermore,



(a) $\alpha = 0$, *omnidirectional microphone.* **(b)** $\alpha = 0.5$, *cardioid microphone.*

Figure 3.2: Absolute value of the denominator of the equalization function (3.6) for different microphone directivities and $\rho_0 = 0.78$ m. The figures is adapted from [9].

the noise introduced in the measurement process is amplified by the equalization function Equation (3.6) when $|J_\mu(k\rho_0)| < 1$. Unfortunately, zeros of the Bessel function turn out to limit the usable frequency range to narrow bands between the peaks of $E_\mu(\omega)$, as illustrated in Figure 3.2a. This constraint is not acceptable for systems which are designed to deliver broadband audio signals [9].

A cost-effective solution to the problem of the zeros of the Bessel function is to use a single cardioid microphone [35]: we will adopt this solution in the rest of this work. However, in order to separate the contributions coming from pressure and pressure gradient, we need to consider a cardioid microphone as a superposition of an omnidirectional and a pressure gradient microphone (i.e. a figure-8 microphone). In this way we can write [9]

$$C(\omega, \rho_0, \phi) = \alpha P(\omega, \rho_0, \phi) - \alpha \varrho c V_\rho(\omega, \rho_0, \phi), \tag{3.7}$$

where α is a coefficient that controls the directivity, $P(\omega, \rho_0, \phi)$ is the pressure signal (obtained, in principle, from the omnidirectional microphone), ϱ is the density of the air and $V_\rho(\omega, \rho_0, \phi)$ is the radial component of the particle velocity (as recorded by a fictitious pressure gradient microphone). To obtain an ideal cardioid directivity, the parameter α is set to 0.5.

Since the Fourier series is linear, the coefficients can be superimposed

$$\begin{aligned} \mathcal{S}_\phi \left\{ C(\omega, \rho_0, \phi) \right\} &= \alpha \mathcal{S}_\phi \left\{ P(\omega, \rho_0, \phi) \right\} - \alpha \varrho c \mathcal{S}_\phi \left\{ V_\rho(\omega, \rho_0, \phi) \right\} \\ &= C_\mu(\omega) \left[\alpha J_\mu(k\rho_0) - \alpha j \mathrm{sgn}(\omega) J'_\mu(k\rho_0) \right], \end{aligned} \tag{3.8}$$

where $J'_\mu(k\rho_0)$ represents the derivative of $J_\mu(k\rho_0)$; these two functions do not

have common zeros [22]. Thus, the equalization function

$$E_\mu^C(\omega) = \frac{1}{\alpha J_\mu(k\rho_0) - \alpha j \mathrm{sgn}(\omega) J'_\mu(k\rho_0)} \tag{3.9}$$

is free of peaks, resulting in the least noise amplification over audio frequency range [9], as illustrated in Figure 3.2b.

3.2 Virtual Microphone Array

In this section we introduce the concept of virtual microphone array as a tool to perform sequential measurements on a circumference in an automatic fashion. Then, we discuss the advantages and disadvantages of sequential measurements compared to static parallel setups.

The circular geometry is the most convenient to automatically measure a sound field at a large number of positions. These measures can be performed by mounting a a cardioid microphone mounted on a rotating rig, which is driven by a stepper motor. With this simple setup we are able to place accurately the microphone at a suitable number of positions on a circumference, sampling the sound field in each position; at each acquisition step, the same excitation signal is produced in the room. A *virtual array* like this provides the same data than a multichannel measurement setup but with greatly reduced costs. In particular, it allows to reduce the number of microphones to be employed and also the number of preamplifiers and A/D converters, compared with a parallel (i.e. multichannel) measurement setup.

This fact holds only if the same excitation signals are reproducible for each step of the sequential measurement procedure. This possibility applies for room impulse response measurement [36]. Provided that room acoustics is constant over the whole measurement time, the signals at measurement positions are perfectly reproducible.

The advantages of sequential measurements over parallel measurements can be easily identified:

- we can employ a reduced number of the measurement microphones;
- we can neglect the additional cost of the high precision rotating device, considering that the number of microphone preamplifiers and A/D converters is reduced with the number of microphones;
- when a very dense spatial sampling is needed, a sequential setup results

in less scattering at the single microphone;

- for measurements over a large area only a larger stand is needed while the number of microphones is not increased; this results in a more portable setup.

However, we can devise also some disadvantages associated with a sequential measurement setup in comparison with parallel setups. In particular, the measurement time is increased: this turns out to be a relevant problem if the location where measurements have to be performed is accessible only for a limited period; moreover, for large measurement tasks the room acoustics may not be sufficiently constant over the whole measurement time (e.g. temperature may change) and environmental noises could become an issue if the measurement lasts several hours. As a consequence, the application of virtual arrays is limited to the task of room impulse response measurement (i.e. the task considered in our work), for which the preceding conditions hold.

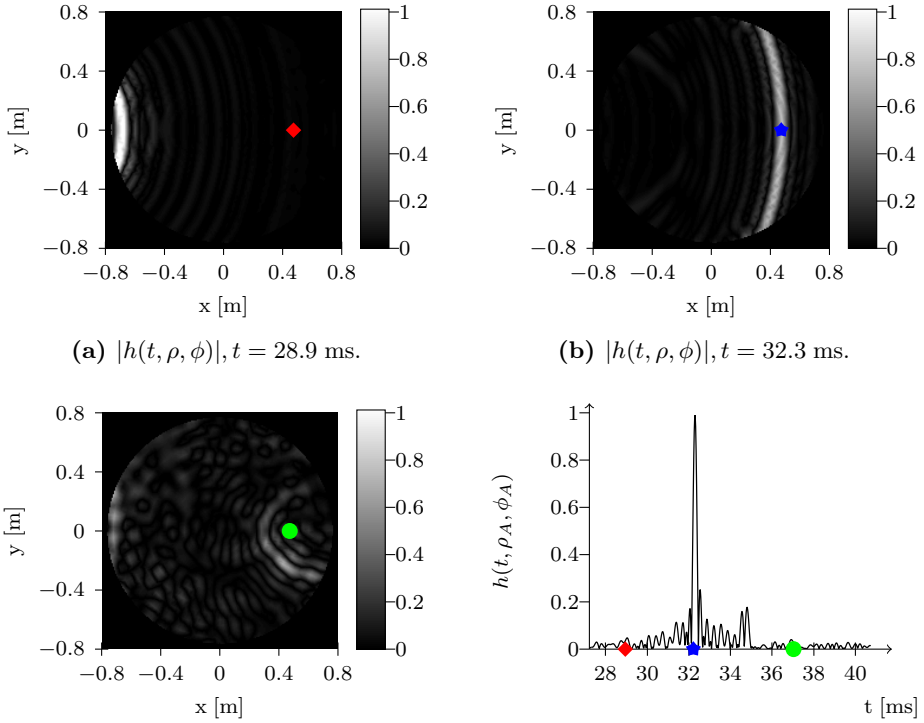
3.3 Space-Time Impulse Response and Space-Frequency Representations

In this section we describe a methodology for room impulse response measurements which is based on the CHD (Section 3.1) and adopts the sequential measurement setup described in Section 3.2.

The classical procedure to measure the impulse response of a system consists of an excitation of the system with a deterministic signal. Then the reaction of the system is recorded and deconvolved with the known excitation signal. In our case, the system under evaluation is the room where a sound field rendering system is operating and a recording of the system reaction is provided by microphone signals.

The most straightforward excitation signal would be a band-limited Dirac impulse, which would allow to skip the deconvolution step. However this procedure is not practical since it is not robust in terms of signal-to-noise ratio. As a matter of fact, impulses can be generated by practical systems only with a finite amplitude so that less excitation energy can be fed into the system than with a less impulsive excitation signal. Thus, signals with a lower crest factor are preferred since they allow an higher average power of the measurement signal.

In the following of this work we will adopt a pseudo-random Maximum Length Sequence (MLS) as excitation signal. MLS are binary signals with a



(a) $|h(t, \rho, \phi)|, t = 28.9$ ms.

(b) $|h(t, \rho, \phi)|, t = 32.3$ ms.

(c) $|h(t, \rho, \phi)|, t = 37$ ms.

(d) $|h(t, \rho, \phi)|$ *at the marked point.*

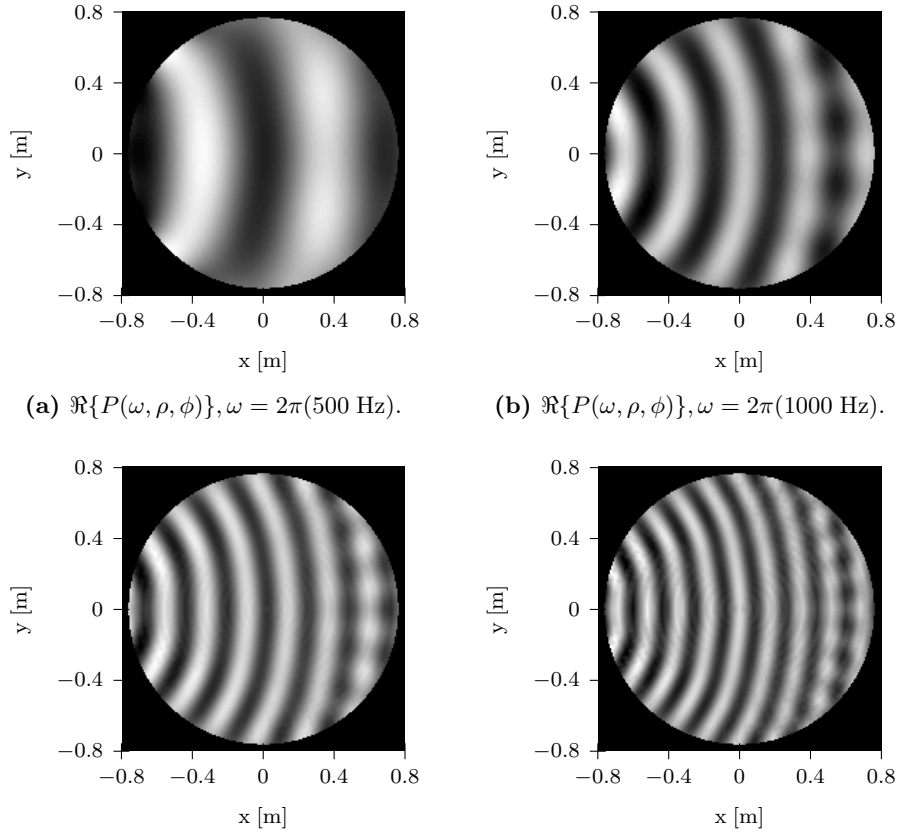
Figure 3.3: Space-time impulse response of a sound field rendering system controlled to reproduce a virtual omnidirectional source at point $(-5\text{ m}, 0\text{ m})$. The absolute value of the sound field pressure is shown for three time instants (3.3a, 3.3b, 3.3c) and at a given point (3.3d).

crest factor of 0 dB, so that they allow a wide exploitation of the dynamics of the D/A converters and, consequently, a high signal-to-noise ratio.

An illustrative example of a measured space-time impulse response is depicted in Figure 3.3. The rendering system adopted for this measurement is composed of $M = 32$ loudspeakers disposed on a linear array of length 2.035 m. For this experiment, the system was installed in a rectangular dry room, the reverberation time T60 being approximately equal to 50 ms. The rendering system was simulating the presence of an omnidirectional virtual source located 2.5 m behind the array.

In particular, Figures 3.3a-3.3c show three snapshots of the normalized absolute value of the propagating wavefronts for time instants t in the range [28.9 ms $\sim$ 37 ms], the time $t = 0$ s being the time at which the wavefront is emitted by the loudspeaker array.

Figure 3.3d shows the relationship between the impulse response $|h(t, \rho, \phi)|$, taken at time instant and the impulse response $|h(t, \rho_x, \phi_x)|$ considered in the



(a) $\Re\{P(\omega, \rho, \phi)\}, \omega = 2\pi(500 \text{ Hz})$. **(b)** $\Re\{P(\omega, \rho, \phi)\}, \omega = 2\pi(1000 \text{ Hz})$.

(c) $\Re\{P(\omega, \rho, \phi)\}, \omega = 2\pi(1500 \text{ Hz})$. **(d)** $\Re\{P(\omega, \rho, \phi)\}, \omega = 2\pi(2000 \text{ Hz})$.

Figure 3.4: Space-frequency representation of a sound field generated by a rendering system controlled to reproduce a virtual omnidirectional source at point $(-5 \text{ m}, 0 \text{ m})$. The real part of the sound field pressure is shown for four frequencies between 500 Hz and 2000 Hz.

marked point at coordinates $\rho_x = 0.5$ m, $\phi_x = 0$ rad.

Another useful way to represent a sound field is to describe the sound pressure as a function of the spatial position and frequency. In order to obtain a space-frequency representation starting from the space-time representation, we can apply the Fourier transform with respect to time t to the space-time impulse response

$$P(\omega, \rho, \phi) = \mathcal{F}_t \left\{ p(t, \rho, \phi) \right\} = \int_{-\infty}^{\infty} p(t, \rho, \phi) e^{-j\omega t} \mathrm{d}t. \tag{3.10}$$

An illustrative example of space-frequency representation derived from the space-time impulse response of Figure 3.3 is depicted in Figure 3.4.

In particular, Figures 3.4a-3.4d show four images of the real part of the sound field $P(\omega, \rho, \phi)$ for frequencies f in the range [500 Hz $\sim$ 2 kHz].

3.4 Conclusions

In this chapter we have presented a methodology to measure the impulse response of a sound field rendering system. This methodology relies on the decomposition of a sound field into circular harmonics and on a virtual microphone array that makes sequential measurements feasible. In particular, we have discussed a practical implementation of the CHD procedure which employs a single cardioid microphone to avoid the amplification of noise due to numerical issues.

Moreover, we have described a complete procedure to obtain the space-time impulse response of a sound field rendering system. We have shown that, starting from the knowledge of the sound pressure on finite positions on a circumference we can obtain the space-time impulse response of the system in the whole listening area. For this purpose, the rendering system emits a MLS, which is white in a limited bandwidth and, by means of the virtual microphone array, we acquire the sound pressure at the desired positions.

This measurement procedure makes no assumption on the specific technique adopted for the rendering, thus it is general enough to be employed with most of the rendering techniques presented in the literature.

Chapter 4

Psychoacoustics-Based Evaluation of Rendering Artifacts

In Chapter 2 we have shown that the practical realization of sound field rendering systems always imposes some constraints which cause a distortion of the rendered sound field. It is shown in [5] that these distortions causes artifacts that are perceivable by a human listener.

In order to better characterize the nature of such artifacts, we consider the case of rendering an impulsive sound field. In this case the rendering systems exhibit a well defined direct wave front at all positions inside the listening area, immediately preceded and followed by secondary wave fronts, named *pre-echoes* and *post-echoes*. It is shown in [6] that pre-echoes and post-echoes are perceived by a human listener as a timbral distortion of the rendered sound signal.

In this chapter we introduce a class of metrics aimed at characterizing the timbral distortion of an impulsive sound signal delivered by a sound field rendering system. In particular, our metrics are based on the ratio between the power carried by the direct wave front and the power of pre-echoes and post-echoes. We have based our analysis on the psychoacoustic effect of masking in time domain. This effect allows us to discriminate whether a specific replicated wave front affects the perception of timbral quality or not; the discrimination is based on the time interval between the considered echo and the direct one and, moreover, according to a masking threshold. This approach is widely adopted in the literature of perceptual audio coding.

Then, we consider that the shape of the wave fronts produced by a rendering system is distorted by spatial aliasing and truncation effects. Thus, such artifacts may cause an erroneous localization of the sound source.

In this chapter we present a methodology to retrieve the position of a virtual source given the curvature of the wave fronts. In this way we are able to evaluate the errors committed by a human listener in the localization of a virtual source, due to spatial artifacts introduced by the rendering system. This methodology relies on a generalized Hough transform to retrieve the position of the virtual source. The localization results obtained with the generalized Hough transform are then compared to subjective listening tests to prove the accuracy of the proposed localization method.

The rest of this chapter is organized as follows. In Section 4.1 we analyze the impact of pre-echoes and post-echoes on the perception of timbral quality. Then we introduce the effect of masking in time domain and present our psychoacoustic-based metrics. In Section 4.2 we present a methodology to retrieve the position of a virtual sound source given the space-frequency representation of the sound field.

4.1 Timbral Artifacts

In this section we address the problem of evaluating timbral artifacts introduced by sound field rendering systems. Timbral artifacts are generated by undesired wavefronts produced by sound field rendering systems due to their intrinsic non-ideality. We will show that such artifacts will arise as pre-echoes and post-echoes in the impulse response of a rendering system. Then, we introduce objective metrics to quantify the presence of such artifacts in an impulse response. Since these metrics have no relation with human perception, we introduce another class of metrics which are based on psychoacoustic considerations. In particular, we review the masking effect in time domain and we adopt a masking curve, well-known in the literature of perceptual audio coding, to develop our metrics.

The space-time impulse response obtained with the measurement methodology presented in Chapter 3 allows to consider a sound field in terms of propagating wavefronts. The ideal space-time impulse response of a sound field rendering system should consist of one single narrow wavefront which impinge on the listening area at a time instant coherent with the location of the virtual source.

However, every practical sound field rendering system is subjected to non-idealities which deviate the behavior of the actual system with respect to the ideal case. The first, macroscopic, difference between a practical system and the theoretical formulation is the spatial sampling due to the use of a discrete distribution of loudspeakers instead of a continuous one. Furthermore,

truncation effects arise due to finite size of the sources arrangement. In [5] the authors show that such non-idealities produce acoustic wavefronts that impinge on the listening area before or after the desired wavefront. These additional wavefronts are named *pre-echoes* and *post-echoes*, respectively.

Moreover, practical realizations of sound field rendering systems are subjected to other non-idealities. In particular, we notice that the sound reproduction system has a frequency dependent response, due to the presence of non-ideal D/A converters, amplification stages and non-ideal mechanical behavior of the loudspeaker itself. In addition, loudspeakers present a non-omnidirectional radiance pattern (i.e. they do not act like ideal monopole sources, cfr. Section 2.3.1). In the end, another non-negligible deviation from the ideal behavior is the reflective behavior of the walls of the room in which the rendering system is operating. Indeed, the case of system operation in an anechoic environment is not of practical interest and, however, in the following paragraphs we will show that post-echoes due to wall reflection arise even in an acoustically controlled environment, if no room compensation is applied.

Informal listening experiments presented in [6] show that the presence of pre-echoes and post-echoes affect the perception of quality of a rendered sound field in terms of timbral coloration, alteration of the timbral character, loss of definition of the transients and other timbral artifacts.

4.1.1 Pre-Echoes and Post-Echoes

Time-domain artifacts can have a great influence on the perception of quality of a sound field rendering technique. These artifacts can be seen as a noticeable amount of signal energy coming before (*pre-echo*) or after (*post-echo*) the desired virtual source sound [5].

Post-echoes are a common phenomenon in everyday experience: consider the scenario in which a sound source is operating in a real-life environment enclosed by reflective walls. The presence of echoes as delayed versions of the sound source signals along with the sense of spaciousness given by reverberations are an expected issue in such non-anechoic environments. However, post-echoes produced as artifacts by a sound field rendering system are not related to any sense of improved spaciousness or impression of the acoustic environment. Since they are uncorrelated with respect to source position inside the virtual environment, they cause comb-filter effects that are perceived as a timbral degradation of the original sound source.

On the other hand, pre-echoes are a kind of artifacts arising in digital audio

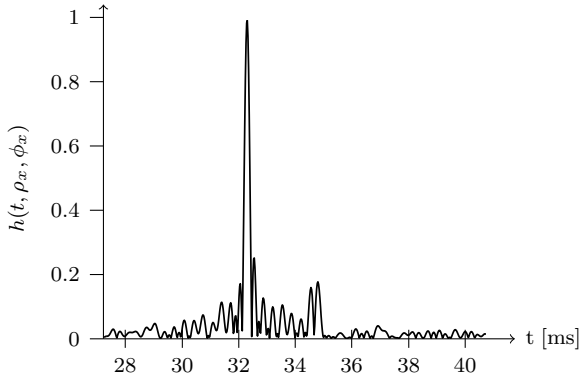


Figure 4.1: Impulse response of a sound field rendering system. Notice the presence of pre-echoes and post-echoes arising as a series of uncorrelated peaks mixed with a noisy behavior.

processing, thus they are not so common as post-echoes in everyday experience. In the field of perceptual audio coding, pre-echoes are generated by blocking artifacts in transform coders [37]. In the context of sound field rendering, some pre-echoes contributions arise from the fact that loudspeaker filter coefficients generated by a sound field synthesis technique like WFS or GR are non-causal [38].

We address the problem of an evaluation of pre-echoes and post-echoes restricted to sound field rendering applications. In this context pre-echo and post-echo phenomena arise as a time-domain distortion of the impulse response produced by a sound field rendering technique. An illustrative example of an impulse response distorted by pre-echoes and post-echoes is showed in Figure 4.1, where a series of peaks mixed with a noisy behavior can be noticed before and after the main peak of $|h(t, \rho_x, \phi_x)|$. The notation $h(t, \rho_x, \phi_x)$ indicates a temporal impulse response taken at point (ρ_x, ϕ_x).

In order to highlight the relationship existing between additional wavefronts introduced by a rendering system and pre-echoes and post-echoes, in Figure 3.3 on page 33 we have shown a space-time representation of the sound field. In particular, we have represented the absolute value of the spatial impulse response of the rendering system at different time-instants. Notice the presence of secondary wavefronts immediately before and after the main one in Figures 3.3a and 3.3b, which contribute to produce pre-echoes and post-echoes effects.

The reflective behavior of the walls of the room where the rendering system is operating must be considered as an additional source of post-echoes artifacts. Figure 3.3c shows the space-time impulse response of a sound field rendering system operating in an acoustically controlled environment (with a reverbera-

tion time T60 of approximately 50 ms) at the time instant in which the main wavefront has passed through the listening area and has been reflected by close wall.

In the next paragraph we present an objective methodology to quantify the amount of pre-echoes and post-echoes starting from the impulse response of a sound field rendering system.

4.1.2 Objective Evaluation

Consider a temporal impulse response $h(t, \rho_x, \phi_x)$ obtained by sampling the space-time impulse response $h(t, \rho, \phi)$ at point $x = (\rho_x, \phi_x)$. This impulse response is constituted by main peak (representing the main wave front propagating in the environment) and a series of secondary peaks coming before (pre-echoes) and after it (post-echoes). These secondary peaks represent undesired wave fronts in the sound field. Our goal is to quantify the amount of such secondary peaks in relation with the main peak of the impulse response $h(t, \rho_x, \phi_x)$.

The key idea, on which the proposed metrics rely, is to evaluate the power ratio between the main peak of the impulse response and the tails before and after it. For this objective evaluation we adopt a simple model of the impulse response. In particular, we consider the impulse response of a sound field rendering system as composed by three segments:

- an early part (referred to as *pre-echo*) from the beginning of the impulse response to the main peak;
- a central part (referred to as *direct peak*) which includes only the main peak;
- a late part (referred to as *post-echo*) from the main peak to the end of the impulse response.

The metric that we are going to propose is analogous to the *Direct-to-Reverberant Ratio* introduced in [39] as a criterion to evaluate the impulse response of a reverberant room. We introduce *Peak-to-(pre)-Echo Ratio* PER^{-}

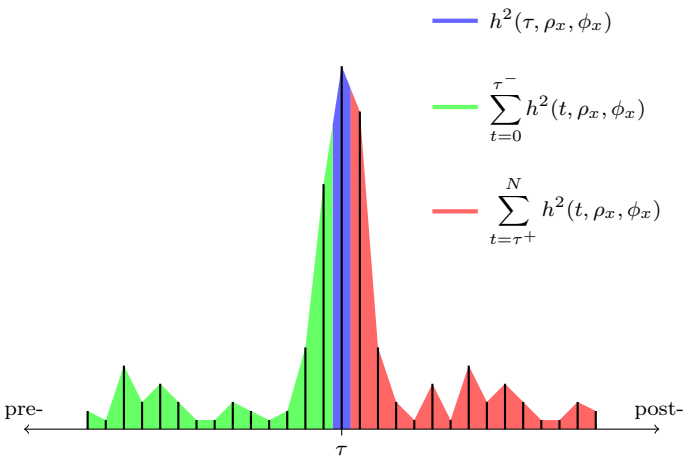


Figure 4.2: Contributions in the computation of PER^+ and PER^-.

and *Peak-to-(post)-Echo Ratio* PER^+ as:

$$\mathrm{PER}^- = 10\log_{10}\left[\frac{\frac{1}{\Delta T}h^2(\tau, \rho_x, \phi_x)}{\frac{1}{\tau^-}\sum_{t=0}^{\tau^-} h^2(t, \rho_x, \phi_x)}\right], \tag{4.1}$$

$$\mathrm{PER}^+ = 10\log_{10}\left[\frac{\frac{1}{\Delta T}h^2(\tau, \rho_x, \phi_x)}{\frac{1}{N-\tau^+}\sum_{t=\tau^+}^{N} h^2(t, \rho_x, \phi_x)}\right], \tag{4.2}$$

where ΔT is the sampling period, τ is the measured time of arrival of the main echo; τ^- and τ^+ are the time instants immediately preceding the main peak; $h(t, \rho_x, \phi_x)$ is the impulse response at the considered analysis point $x = (\rho_x, \phi_x)$ and N is the length of the impulse response.

In order to clarify how the computation of PER^- and PER^+ is carried on, in Figure 4.2 we show the contributions concurring in the computation of the presented metrics. In particular, we show an illustrative impulse response, highlighting its segmentation into early, central and late parts. We remark that high values of PER^- and PER^+ are expected when pre-echoes and post-echoes are, respectively, not relevant.

The application of the metrics PER^- and PER^+ allows us to quantify pre-echoes and post-echoes artifacts from a purely objective point of view. However, such metrics do not provide any relation between the resulting

numerical quantities and the perception of quality in a rendered sound field. Hence, such results can only provide poor information about the actual impact of pre-echoes and post-echoes on the human perception.

With the aim of providing some evaluation metrics more closely related to human perception, in the next paragraph we present a class of psychoacoustics-based metrics based on the masking effect.

4.1.3 Psychoacoustics-Based Evaluation

We can notice from simulations presented in [5] that, for typical rendering systems installed in small auditoria, pre-echoes and post-echoes are mainly concentrated in a short time window. This time window is centered around the main peak and has a duration of about 10 ms. From this observation we can gather that pre-echoes and post-echoes are so close to the main peak that they are not perceived as separated echoes. Hence, an analysis based on the well known *precedence effect* does not provide useful information.

Contrarily, we base our analysis on the *masking effect*, according to which we are able to explain the different perceptual impact of pre-echoes and post-echoes. More specifically, we draw on the masking curve presented in [37] and widely adopted in the literature of perceptual coding.

In the next paragraph we present a review of the masking effect in time domain.

Masking in Time Domain

Masking is a perceptive effect by which a sound stimulus, the *maskee*, becomes inaudible due to the presence of a louder stimulus, the *masker*. Masking can occur when the two stimuli are simultaneous (which is the case of *masking in frequency domain*), but it also occurs when the two stimuli are shifted in time. The presence of a louder stimulus (such as the main peak of our temporal impulse response) makes inaudible secondary stimuli coming after it (*post-masking*) or before it (*pre-masking*). While the effect of post-masking is more or less expected (it corresponds to a decay of the presence of the masker [40]), the effect of pre-masking appears to be non-causal.

This effect is interpreted by psychoacoustic considerations. It is known that the perception of a sound does not occur instantaneously, but requires a build-up time. Thus the weaker stimulus cannot be perceived because of the arrival of a louder stimulus which is processed "faster" by the hearing system [40]. However, only events occurring in a time window shorter with respect to

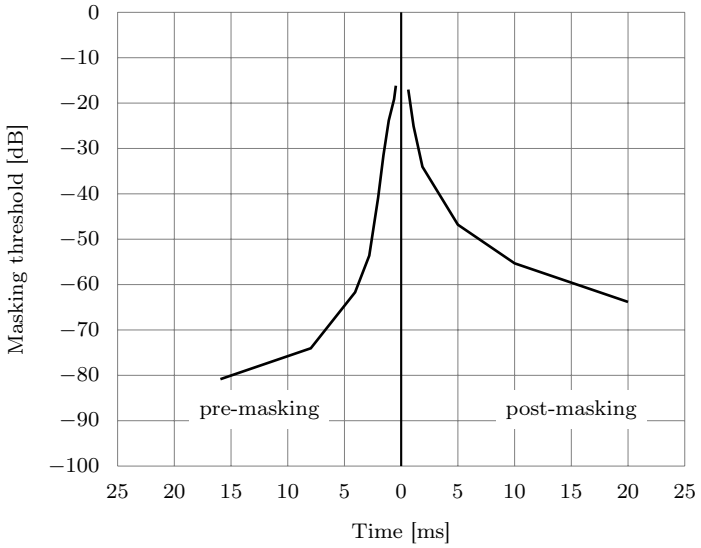


Figure 4.3: Masking experiment, reprinted from [37]. The figure shows masking threshold in dB due to a Gaussian-shaped impulse occurring at time $t = 0$ ms as a function of the relative time position between the masker and a noisy maskee.

post-echoes are masked. Moreover, the pre-masking effect presents a relevant variation among different subjects. Nonetheless, pre-masking is exploited in perceptual audio coding systems to hide the presence of pre-echoes generated by blocking artifacts in transform coders [37].

To quantitatively describe pre-masking and post-masking, we rely on experimental data presented in [37] and reported in Figure 4.3. This figure shows the masking threshold in dB due to a Gaussian-shaped impulse occurring at time $t = 0$ ms as a function of the relative time position between the masker and a noisy maskee. We notice that the curves relative to pre-masking and post-masking are not symmetrical. Indeed, the effect of pre-masking is shorter with respect to post-masking; this fact justifies the empirical evidence that pre-echoes are generally more annoying than post-echoes, which are masked in a stronger way. This masking curve has been widely adopted in the literature of perceptual coding and thus provides a reliable basis for our psychoacoustics-based analysis.

Pre-echoes and post-echoes evaluation metrics

In this paragraph we present a class of metrics more related to human perception than Peak-to-(pre/post)-Echo Ratio PER^{-} and PER^{+}. In particular, we base our metrics on the temporal masking effect, exploiting the experimental masking curve presented in [37] and reprinted in Figure 4.3. Although this masking

curve has always been exploited in a different research area (i.e. perceptual audio coding), we claim that its application will result in benefits also for the field of spatial audio systems.

In analogy with Peak-to-(pre/post)-Echo Ratio PER^- and PER^+ metrics, also the presented class of metrics relies on the idea to evaluate the average power ratio between the main peak of the temporal impulse response and the tails preceding and following the main peak. The point that qualifies the new class of metrics with respect to objective ones is that, in computing the metrics, we consider only those peaks that exceed the masking threshold at the considered time instant (relative to the time position of the main peak). We introduce *Direct-to-(pre)-Echo Ratio* DER^- and *Direct-to-(post)-Echo Ratio* DER^+ as:

$$\mathrm{DER}^- = 10\log_{10}\left[\frac{\dfrac{1}{\Delta\tau^+ + \Delta\tau^-}\displaystyle\sum_{t=\tau-\Delta\tau^-}^{\tau+\Delta\tau^+} h^2(t,\rho_x,\phi_x)}{\dfrac{1}{\tau-\Delta\tau^-}\displaystyle\sum_{t=0}^{\tau-\Delta\tau^-} \overline{h}^2(t,\rho_x,\phi_x)}\right], \tag{4.3}$$

$$\mathrm{DER}^+ = 10\log_{10}\left[\frac{\dfrac{1}{\Delta\tau^+ + \Delta\tau^-}\displaystyle\sum_{t=\tau-\Delta\tau^-}^{\tau+\Delta\tau^+} h^2(t,\rho_x,\phi_x)}{\dfrac{1}{\tau+\Delta\tau^+}\displaystyle\sum_{t=\tau+\Delta\tau^+}^{N} \overline{h}^2(t,\rho_x,\phi_x)}\right], \tag{4.4}$$

where τ is the measured time of arrival of the main echo; $\Delta\tau^-$ and $\Delta\tau^+$ are the time instants (obtained from the results reported in Figure 4.3) at which the threshold of audibility of the maskee is at -20 dB; $h(t,\rho_x,\phi_x)$ is the impulse response at the considered analysis point $x = (\rho_x,\phi_x)$; N is the length of the impulse response and finally $\overline{h}(t,\rho_x,\phi_x)$ is the part of the impulse response whose values exceed the threshold of audibility.

In order to clarify the procedure of computing DER^- and DER^+ metrics, in Figure 4.4 we show which contributions concur to the computation of the metrics and which others are neglected because of the masking effect. In particular, we show an illustrative impulse response segmented in an early, central and late part highlighting the masking threshold. We remark that high values of DER^- and DER^+ are expected when pre-echoes and post-echoes are not relevant, respectively.

Another important fact to be noticed is that the adoption of the masking curve of Figure 4.3 allows a more perceptually-related definition of the direct

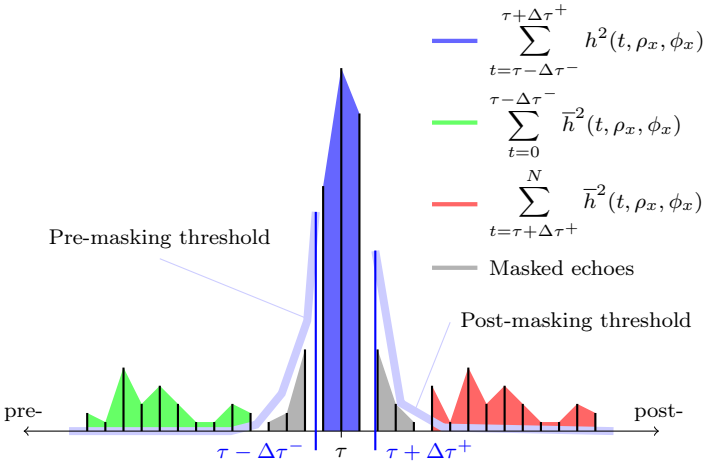


Figure 4.4: Contributions in the computation of DER^+ and DER^-. The filled areas show the energy contribution of pre-echoes, post-echoes and main peak in the computation of the presented metrics.

sound, with respect to a purely objective evaluation. Indeed, objective metrics such as Peak-to-(pre/post)-Echo Ratio consider as direct sound only the main peak of the impulse response. However, we know that human perception is subject to an integration time, thus perception is not instantaneous. This fact means that a human listener is not able to discriminate between sounds which are very close in time. As a result, we can enlarge the direct sound segment of the impulse response like it is clearly illustrated in Figure 4.4.

Notice that the principle underlying the computation of DER^- and DER^+ is analogous to PER^- and PER^+. What greatly differentiates the two classes of metrics is that PER^- and PER^+ are purely objective, since they do not take into account any psychoacoustic effects as it is depicted in Figure 4.2. On the contrary, the computation of DER^- and DER^+ takes into account the masking curve introduced in Paragraph 4.1.3 to neglect contributions that are lower than the curve, as depicted in Figure 4.4.

In this section we have introduced objective and psychoacoustic-based metrics to evaluate the impact of pre-echoes and post-echoes in sound field rendering applications. Objective metrics are based on a simple model which divides the impulse response into three segments and they evaluate the average power ratio between the central segment (which includes only the main peak of the impulse response) and the tails preceding and following it. On the other hand, subjective metrics are based on the psychoacoustic effect known as masking in time domain.

In the next section we address the problem of the localization of an acoustic

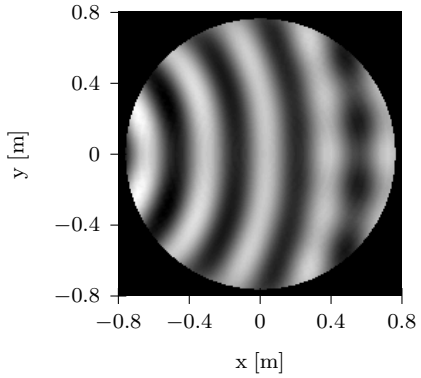


Figure 4.5: Space-frequency representation of the sound field produced by a single virtual source. $\omega = 2\pi \cdot 1000$ Hz.

virtual source, which constitutes another important requirement for a realistic sound field rendering application.

4.2 Localization of acoustic virtual sources

Along with the timbral quality, another attribute which makes a sound scene perceived as realistic by a human listener is the spatial impression. With this expression we indicate the ability of the sound field rendering system to deliver plausible acoustic cues that allow the listener to infer the geometry of the sound scene. In particular, the listener must be able to clearly identify the position of the virtual source and the geometry of the virtual environment in which the virtual source is operating.

In this section we describe a methodology to evaluate the spatial impression delivered by a sound field rendering technique, focusing on the identification of the location of a virtual source with an objective approach. Then, in Chapter 5 we will present listening tests aimed at assessing the accuracy of this objective evaluation methodology.

Now we focus on the identification of the position of the virtual source. We consider the space-frequency representation of a sound field generated by a single virtual source, as illustrated in Figure 4.5. We notice that such an image contains a set of concentric circles, each centered in the position of the virtual source and with different radii. Thus, the space-frequency representation of a sound field can be exploited to retrieve the location of the virtual source which generated such a sound field. This task can be accomplished with a generalized Hough transform technique.

In order to improve the accuracy of this localization methodology, a pre-

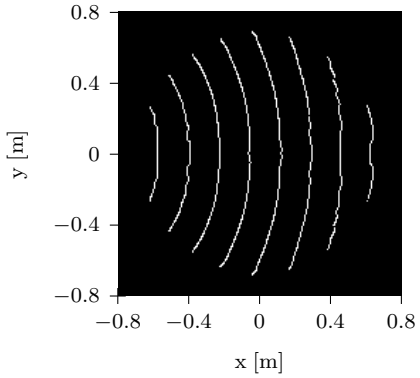


Figure 4.6: Image of the sound field after the pre-processing stage.

processing stage is needed. In particular, the space-frequency image of Figure 4.5 is used to feed an edge detection algorithm to obtain the image points which lay on the concentric circles. As edge detector algorithm we can adopt the *Canny Edge Detector* [41] which looks for the intensity gradient of the image. In Figure 4.6 we show the sound field image after the pre-processing stage. We notice that sound field image after the pre-processing stage can be stored into a matrix full of zeros, with ones in the points corresponding to the maximum intensity of the gradient in the original image.

In the next paragraph we introduce the Hough transform technique to extract circles from an image.

4.2.1 Hough Transform

The *Hough Transform* is a feature extraction technique used in a variety of application domains where image analysis is needed (e.g. computer vision, digital image processing, and others). Its original formulation [42] adopts a slope-intercept parametrization for straight lines. Today Hough transform is generally used in the form proposed in [43] and called *generalized Hough transform* and it turns the problem of the search of a curve into the search of maxima.

The purpose of the Hough transform technique is to find non-ideal occurrences of a specific object by a voting procedure, which is carried out in a *parameter space*. The candidates objects are obtained as local maxima in the *accumulator space*, which is explicitly computed by the algorithm which implements the Hough transform.

A practical implementation of the Hough transform employs an array as accumulator space to detect the candidate object. The dimension of the

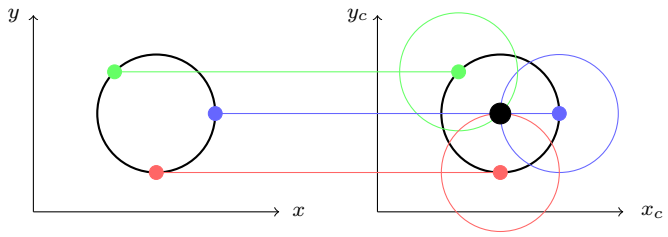


Figure 4.7: Correspondence between the image space and the parameter space for a circle detection problem.

accumulator is equal to the number of unknown parameters in the specific Hough transform problem. In the case of the detection of simple circle, the accumulator space is three-dimensional: two coordinates denote the position of the center in Cartesian coordinates, while the third coordinate denotes the radii.

In order to illustrate how the generalized Hough transform works for simple circles, we recall the equation of a circle in a plane:

$$(x - x_c)^2 + (y - y_c)^2 = r^2, \tag{4.5}$$

where x_c and y_c are the coordinates of the center and r is the radius. Thus, a single circle with center (x_c, y_c) and radius r can be parametrized as

$$x = x_c + r\cos(\theta) \tag{4.6}$$

$$y = y_c + r\sin(\theta). \tag{4.7}$$

When the angle θ sweeps from 0° to 360°, the points (x, y) trace the perimeter of the circle.

If an image contains points which lay on a circumference, the goal of an algorithm implementing the Hough transform is to find parameters triplets (x_c, y_c, r) which describe each circle. Thus, the parameter space for circle detection belongs to $\mathbb{R}^3$.

Considering the case of the detection of circles with known radius r_0, the parameter space is reduced to two dimensions and the objective is to find the coordinates (x_c, y_c) of the center. In Figure 4.7 we show the correspondence between the image space (on the left) and the parameter space (on the right). In particular, we notice that each circle in the image space generates a circle in the parameter space. The circles in the parameter space intersect in the point (x_c, y_c) which is the center of the circle in the image space. This center can be found with a two-dimensional Hough accumulator array.

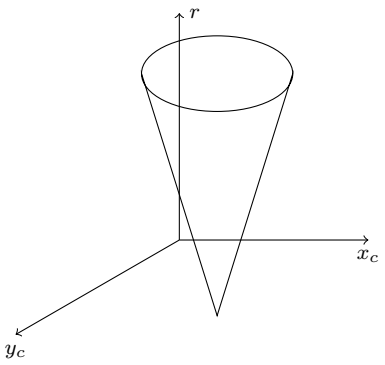


Figure 4.8: Three-dimensional parameter space for a circle detection problem. A conical surface is generated for each point (x, y) in the image space.

However, if the radius in the image space is not known, then the parameter space is three-dimensional. In this case, the locus of points in the parameter space will fall on the surface of a cone. The triplet (x_c, y_c, r) which identifies the searched circle will correspond to the accumulation cell where the largest number of cone surfaces intersect.

In Figure 4.8 we illustrate the generation of a conical surface in the parameter space for one single (x, y) point in the image space. A circle with a different radius is constructed at each "slice" r. The search of circles with unknown radii can be conducted using a three-dimensional Hough accumulator array.

4.2.2 Detection of Concentric Circles

The simple case of the detection of a circle allows us to introduce a further generalization of Hough transform to detect the center of a set of concentric circles. The surfaces generated in the parameter space are, in this case, a set of cones with vertex aligned on a line perpendicular to the (x, y)-plane. To populate the three-dimensional Hough accumulator we adopt the algorithm presented here.

1. We consider a three-dimensional accumulator with the first 2 dimensions specifying the coordinates of the circles centers and the third specifying radii. Since we need to consider circles whose centers are out of the image (i.e. sources external to our listening area), the first two dimensions of the accumulator are extended by 2 times the maximum radius we want to detect.

2. We build a radii map by computing the distances between each point and the center, then clear out-of-range radii.

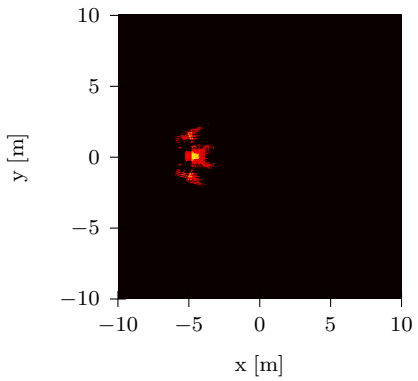


Figure 4.9: Two-dimensional Hough accumulator generated by the concentric circumferences of Figure 4.6.

3. For each pixel on the image we overlap the radii map and increment the corresponding position of the possible center in the accumulator matrix.

By following this algorithm we obtain a three-dimensional Hough accumulator space which we can inspect to find the actual center of all concentric circles.

We need a specific voting process in order to identify the tuple $((x_c, y_c)$ which represents the center of all the concentric circles in the image space. This specific voting process is described by the following steps.

1. We compute the maximum value $m_{\max}$ in the accumulator, i.e. we look for the rate of the highest rated cell in the accumulator.

2. Considering only one "slice" of the accumulator at time (i.e. considering a plane of constant radius r), we filter the accumulator by setting to zero all rates below a given threshold. Simulations have shown that a threshold fixed from 40% to 80% of $m_{\max}$ will provide good and reliable results.

3. We sum up the rates of all cells along the r dimension.

In this way we obtain a two-dimensional accumulator where the parameters are only x_c and y_c. Since all circles in the original image space were concentric, we expect to obtain a cluster of points in the accumulator in the neighborhood of the real center. In Figure 4.9 we show an illustrative two-dimensional accumulator with a distribution of points around the actual center of the circumference, located at $(-5, 0)$ m. This accumulator has been generated by the concentric circles of Figure 4.6.

Once a two-dimensional array has been obtained, we need to introduce a suitable policy to identify a single point inside the clustered distribution.

This point should be the most probable location of the center of the original concentric circles in the image space.

A naive approach would be to select the highest rated cell in the accumulator. However, such policy does not ensure reliable results because error contributions may sum up to corrupt the rate of a single cell.

In order to identify this point in a more robust way, we look for the gravity center of the clustered data, considering as weight the rates corresponding to each cell of the accumulator. The gravity center is found by means of the following formula:

$$\mathbf{x}_c = (x_c, y_c) = \frac{\sum_{i=1}^{N} \mathbf{p}_i m_i}{\sum_{i=1}^{N} m_i}, \tag{4.8}$$

where N is the total number of cells, $\mathbf{p}_i$ and m_i are, respectively, the vector representing each point in the two-dimensional accumulator and the corresponding rate.

4.3 Conclusions

In this chapter we have presented specific methodologies to evaluate the timbral quality and the spatial impression of a rendered sound field. In particular, we have introduced original psychoacoustic-based metrics to evaluate the impact of pre-echoes and post-echoes on the perceived timbral quality. Such metrics rely on the psychoacoustic effect of masking in time domain and are based on a masking curve which quantitatively describe the masking effect.

On the other hand, we have introduced a methodology to evaluate the position of a virtual sound source given the space frequency representation of a sound field. This methodology employs a generalized Hough transform in order to detect the center of a set of concentric circles which constitute the sound field image.

Chapter 5

Subjective Tests

In Chapter 4 we have introduced psychoacoustic-based methodologies to evaluate both timbral and localization artifacts introduced by a sound field rendering system. In this chapter we propose formal listening tests with the aim of providing a subjective validation of the previously introduced methodologies. In particular, we will discuss in detail the design phase of listening experiments, presenting the criteria that most suite the need of our work.

We have shown in Chapter 4 that the non idealities of the rendering system produce both time-domain artifacts (i.e. pre-echoes and post-echoes) and spatial artifacts (i.e. distortion of the shape of the wavefronts). From a perceptual standpoint, such artifacts produce, respectively, timbral modifications and erroneous localization.

In order to provide a subjective evaluation of such effects we need to rely on results gathered from listening tests. In the next section we present a formal methodology for devising a set of suitable listening tests, with the aim of producing reliable results.

We analyze separately the cases of a subjective assessment of timbral artifacts and localization of an acoustic virtual source. This choice is due to the different perceptual nature of the effects of these artifacts. On the one hand, timbral artifacts can be considered as small impairments of the timbral quality of a sound stimulus [6]. Thus they require a specific and strongly controlled test procedure in order to be assessed in a reliable way [12]. On the other hand, the localization of a virtual source is a macroscopic feature delivered by a rendering system, so it does not require such a controlled test method.

5.1 Test conditions

In this section we present the choice of general test parameters, which are valid for both the timbral artifacts and the localization tests. In particular, we discuss the selection of the test panel (i.e. the characterization of subjects involved in the listening tests) and the choice of the sound material adopted in the tests.

5.1.1 Selection of the Test Panel

All subjects involved in our listening tests are experienced in listening to sound in a critical way. In particular, the test subjects are staff members of the *Sound and Music Computing Lab* and students coming from the Masters' Degree in *Sound and Music Engineering* offered by Politecnico di Milano (Polo Regionale di Como).

Despite of this level of expertise of the listeners, we adopt a rejection technique successive to the actual test: after gathering all test results, we perform a *post-screening* operation in order to reject results from unreliable subjects. A straightforward policy to perform post-screening is based on inconsistencies of the test results produced by one subject compared with the mean result. This policy is not justifiable, since it would introduce severe biasing of the results. Instead, we performed post-screening adopting the statistical methodology presented in Recommendation ITU-R BS.1116 [12], which is aimed at identifying subjects that are not able to perform a correct identification of the hidden references (according to the test method presented in Section 5.2.2).

5.1.2 Test Material

In order to study the behavior of the rendering system in reproducing different sound material, we perform all the listening tests with two sound samples. In particular, we adopt Suzanne's Vega *Tom's Diner* [44] and castanets from European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) CD [45]. Such sound samples come from easily accessible sources and have already been adopted for listening tests in the context of sound field rendering evaluation [46].

We choose a vocal signal because human voice is considered critical to evaluate the sense of audio quality and it is known that localization is most sensitive with speech or singing [46]. On the other hand, the castanets sample

has been chosen to allow the evaluation of a more impulsivet signal, where the perception of the attack is critical. Both test samples are first converted into monophonic signals, and then they are rendered as virtual sources according to the test method explained in the next sections. Since the room in which listening tests are performed is dry, no room compensation is applied.

5.2 Timbral artifacts

In this section we describe the test method adopted for a subjective evaluation of timbral artifacts. Considered the nature of such artifacts as small impairments, we adopt a formal and heavily controlled test procedure. In particular, we modify the test method given in Recommendation ITU-R BS.1116 [12] to include the assessment of three stimuli with respect to two reference signals.

At the end of this section we discuss the statistical analysis procedure adopted in order to identify the average behavior of the system and the reliability of the results.

5.2.1 Experimental Design

As mentioned before, the perceptual evaluation of timbral artifacts requires a formal test method, as well as strong control over experimental conditions [12]. In particular, in the design phase of such experiments we need to ensure that uncontrolled factors will not deviate the results of the listening tests. As an example, we propose to each subject the same stimuli but in a different and random order. In this way we can ensure that the judgments made by the subjects are independent from the actual sequence of stimuli.

Another important factor that we have considered during the design phase of our listening test is that the listeners must not be overloaded, in order to prevent a loss of accuracy in the judgments.

As Recommendation ITU-R BS.1116 [12] suggests, we include control conditions in our listening tests. In particular, we introduce unimpaired stimuli among the test samples in a way that is unpredictable to the listeners. The differences between the judgments of these control stimuli and the actually impaired ones increases our ability to discriminate between reliable and not reliable subjects.

Table 5.1: ITU-R five-grade impairment scale given in Recommendation ITU-R BS.1284 [47].

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

5.2.2 Test Method

For our listening tests we modify the *double-blind triple-stimulus with hidden reference* method presented in Recommendation ITU-R BS.1116 [12]. In particular, we present five stimuli to the subject: three of them are blind while two are the references.

One subject at time is involved and five stimuli ("R1", "R2", "A", "B", "C") are presented to him: the subject has the possibility to swap from one stimulus to another at his discretion. The listener is aware that the unimpaired reference is always available as stimulus "R1" and that an highly impaired reference is available as stimulus "R2". The hidden unimpaired reference and two object stimuli are randomly assigned to stimuli "A", "B" and "C".

The subject is asked to rate the impairment on "A", "B" and "C" with respect to "R1" and "R2", according to the five-grade scale reported in Table 5.1, which is derived from the Recommendation ITU-R BS.1284 [47]. The stimuli can be repeated until the subject have made an assessment. Any perceivable timbral difference should be interpreted as an impairment, so the listeners are instructed to look for timbral artifacts such as coloration, distortion of the timbral character, smearing of transients, etc...

We remark that one of the stimuli "A", "B" and "C" should be indiscernible from "R1". This control stimulus is exploited to perform the post-screening rejection phase.

Before beginning the actual grading phase, we deliver a set of neutral instructions to the subject. These instructions include a brief introduction to the scope of the test and a description of the technique of presentation of the stimuli.

5.2.3 Statistical Analysis

A statistical analysis of the results is needed in order to identify the average performance of the system and the reliability of the results. For our experiments,

we resorted to the analysis introduced in [12], which is summarized in the following.

At first, the individual assessment for each test condition are normalized according to mean and standard deviation. This step is essential in order to compare results coming from different sessions. The normalization is achieved via the formula

$$z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s \tag{5.1}$$

where z_i is the normalized result of subject i, x_i is the score of subject i, x_{si} is the mean score for subject i in session s, x_s is the mean score of all subjects in session s, s_s is the standard deviation for all subjects in session s, s_{si} is the standard deviation for subject i in session s. Then, the mean score for each of the presentations is computed as

$$\bar{z} = \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{5.2}$$

where z_i is the score of the subject i and N is the number of subjects.

Post-Screening of Subjects

The test method described in the last paragraph provided three grades for each session and makes it possible to compare results on a subject-by-subject basis. In particular, for each trial we can take the algebraic difference between the grade for hidden reference and the grades for the objects.

The average of all the difference grades from the subject in the whole listening test is used to discriminate whether that subject is able to make correct identifications or not. In particular, if the subject was not successful at identifying the hidden reference with respect to the objects, the average would be close to zero since positive and negative grades will, on average, tend to balance. On the other hand, if the subject was able to to make correct identifications, the average of the differences will deviate from zero in the positive direction.

In order to practically discriminate reliable subjects, we set a threshold for the average of differences. If a subject produced grades whose average of differences is below the threshold, all the results from that subject will not be considered.

Table 5.2: Value of Student's t distribution for two-sided confidence interval of 95%.

Number of scores	t **value**	**Number of scores**	t **value**
1	12.70	12	2.179
2	4.303	15	2.131
3	3.182	20	2.086
4	2.776	25	2.060
5	2.571	30	2.042
6	2.447	40	2.021
7	2.365	50	2.009
8	2.306	100	1.984
9	2.262	$1000 \sim \infty$	1.962
10	2.228		

Presentation of the results

We decide to present the test results with an associated confidence interval, in order to provide an explicit indication of the reliability of the results. The confidence interval is derived from the standard deviation and the size of the listening panel. In particular, we use a 95% confidence interval which is given by

$$[\bar{z} - \delta, \bar{z} + \delta]. \tag{5.3}$$

In this case δ is determined by

$$\delta = t_{0.05} \frac{s_s}{\sqrt{N}}, \tag{5.4}$$

where $t_{0.05}$ is the t-value [48], i.e. the value of the Student's distribution for a two-sided confidence interval level of 95%, as reported in Table 5.2. Notice that the number of scores (i.e. the number of degrees of freedom) involved in the retrieval of the t-value is the number of scores resulting for a particular test condition after post-screening stage.

5.3 Localization of Acoustic Virtual Sources

In this section we describe the test method adopted for a subjective evaluation of the localization of a virtual sound source. In particular, we restrict to the case of a single virtual source rendered at the same distance from the listener but with different angles of incidence.

At the end of this section we discuss a statistical analysis procedure adopted

in order to identify the average behavior of the system and the reliability of the results.

5.3.1 Experimental Design

The spatial impression, here restricted to the localization of a virtual source, is a macroscopic feature delivered by a sound field rendering system. Thus, in order to gather reliable information from a subjective evaluation of the source position we can adopt a less costly test method, with respect to the method introduced to evaluate timbral artifacts. Thus, we do not introduce strong control conditions.

However, we need to take into account some aspects design phase of such experiments. In particular, as we made for the evaluation of timbral artifacts, we propose to each subject the same stimuli but in a different and random order. In this way we can ensure that the judgments made by the subjects are independent from the actual sequence of stimuli.

5.3.2 Test Method

In order to get a subjective evaluation of the localization of an acoustic virtual source we adopt a test method much simpler than the one introduced to evaluate timbral artifacts. In particular, we present three stimuli without any reference.

One subject at time is involved in the listening test and three stimuli ("A", "B", "C") are presented to him: the subject has the possibility to swap from one stimulus to another at his discretion. The listener is aware that sound fields rendered with different positions of the virtual source are randomly assigned to stimuli "A", "B" and "C".

The subject is asked to indicate the perceived angle of incidence of the virtual sources assigned to stimuli "A", "B" and "C". We provide an angular scale on the top of the linear loudspeaker array used for this test. Subjects can make a judgment on the perceived angle of incidence using this scale as an hint. In particular, the angular scale ranges from $-20°$ (left edge of the array) to $20°$ (right edge of the array) with a resolution of $10°$.

Listeners are aware that this scale is meant to provide only an hint and no assumptions are made on the relative positions of virtual sources and anchors of the scale.

Before beginning the actual grading phase, we deliver a set of neutral instructions to the subject. These instructions include a brief introduction to

the scope of the test and a description of the technique of presentation of the stimuli.

5.3.3 Statistical Analysis

The statistical analysis procedure adopted for localization tests is almost identical to that presented in Section 5.2.3 in the case of timbral evaluation. The only difference is that here no post-screening is performed, since the test aims at identifying a macroscopic feature of the sound stimuli and the requirement of expertise of the listeners is sufficient in order to obtain reliable results.

5.4 Conclusions

In this chapter we have proposed a subjective methodology to evaluate a rendered sound field under the aspects of timbral quality and spatial impression. In particular we discussed the issues of the selection of an adequate test panel and the choice of the sound material to be used through the tests.

We have presented a formal test method to assess the timbral artifacts introduced by a sound field rendering system. Furthermore, we have described a statistical analysis procedure meant to identify in a reliable way the average impact of timbral artifacts and the reliability of the test results.

Finally, we have presented the test method adopted to evaluate the localization of a virtual acoustic source. Such evaluation is less critical with respect to the evaluation of timbral artifacts, so we adopted a more relaxed test procedure. At the end, we reviewed the statistical analysis procedure presented in the case of a timbral evaluation with adaptations to the case of localization tests.

Chapter 6

Evaluation

In this chapter we show how we implement a measurement setup based on the methodology presented in Chapter 3 to obtain both space-time and space-frequency representations of a practical rendering system. Then, we show how the evaluation methodology is applied on real data and we validate the results through the subjective evaluation methodology described in Chapter 5. Since the evaluation methodology presented in Chapter 4 is based on the analysis of the space-time impulse response of the rendering system and on the space-frequency representation of the rendered sound field, we are interested in simulating and measuring the space-time impulse response and the space-frequency representation.

In particular, we introduce two simulative scenarios: in one case we simulate a loudspeaker array driven by a rendering technique to obtain its theoretical space-time impulse response; this impulse response is obtained by considering the rendering solution for filter coefficients and the propagation matrix. In the other case, we simulate the measurement procedure considering an ideal propagation (in an anechoic environment) and ideal omnidirectional loudspeakers. The comparison between the theoretical results and the results coming from a simulation of the measurement procedure allow us to characterize the error introduced by the measurement methodology.

Then, we apply the evaluation methodology to a real-world setup, which has been implemented in a semi-anechoic room. In particular, we measure the sound field and perform the psychoacoustic-based evaluation introduced in Chapter 4 on these real data. With an analysis of these results we are able to evaluate the artifacts introduced by the practical implementation of the rendering system. Furthermore, we show the results of a subjective evaluation conducted in the same environment adopted for the measurements.

The comparison between simulative results, experimental results and listening tests allows us to validate the psychoacoustic-based evaluation methodology.

This Chapter is structured as follows. In Section 6.1 we describe the experimental setup. Then, in Section 6.4.1 we discuss the errors introduced in the rendering of a sound field, considering its space-frequency representation obtained with both simulative and experimental procedures. In Sections 6.4.2 and 6.4.3 we show the results of the objective and psychoacoustic-based evaluation of timbral and localization artifacts, respectively. In order to validate our objective and psychoacoustic-based results, in Section 6.4.4 we present the results of formal listening tests aimed at assessing the impact of artifacts on the perceptual performance of a sound field rendering system. At the end, in Section 6.5 we draw some conclusions.

6.1 Setup

In this section we describe the experimental setup that we have adopted to obtain a measurement of the space-time impulse response and a space-frequency representation of the rendering system. Also the listening tests have been conducted using the same setup and in the same environment.

6.1.1 Acoustic Environment

The experiments have been performed in a semi-anechoic room with sizes of 4.3×4 m^2 and height 2.6 m. The room is depicted in Figure 6.1. With the expression "semi-anechoic room" we mean a room in which the walls, ceiling and floor are covered with sound absorbing material. This treatment is aimed at minimizing reverberation phenomena and allows to get propagation conditions closer to free-field. This room has a reverberation time T60 = 50 ms and it has been made available for our research by the *Sound and Music Computing Lab* of Politecnico di Milano, Polo Regionale di Como.

As it is depicted in Figure 6.1, in this room we have installed a sound field rendering system composed of $M = 32$ loudspeakers disposed on a linear array of aperture $l = 2.035$ m. The loudspeakers composing the array are *Empire M2* speakers whose specifications are reported in Table A.1. The mid-point of the array is placed 2.5 m far from the center of a circular listening area of radius $r = 1$ m. The geometry of the room with the installed loudspeaker array and the listening area is represented in Figure 6.2. This setup leads to a spatial Nyquist frequency $f_{\max} = c/2d \approx 2.7$ kHz, where c is the speed of

Figure 6.1: Semi-anechoic room in which the experiments have been performed. Notice the presence of the loudspeaker array.

sound and $d = 6.5$ cm is the relative distance between adjacent loudspeakers.

For all the simulations and the measurements considered in this chapter, we are interested in the sound field inside a circular listening area of radius $r = 0.8$ m and centered in the origin of the coordinate system (with reference to Figure 6.2). This area is then sampled on a grid of 201×201 points.

The sampling frequency adopted for these simulations is $f_s = 44.1$ kHz and the frequency axis is sampled in 1024 points between 0 Hz and $f_s/2 = 22.05$ kHz. The corresponding wave numbers k are defined as

$$k = \frac{2\pi f}{c}, \tag{6.1}$$

$c = 340$ m/s being the sound speed in air.

6.1.2 Measurement System

In the rendering room described in the previous paragraph we have installed a sound field measurement system that implements the measurement methodology described in Chapter 3. In particular, the measurement system is composed of the following devices:

- a measurement computer that plays the double role of controlling the rendering system and the acquisition device,
- a virtual circular microphone array composed of a cardioid microphone

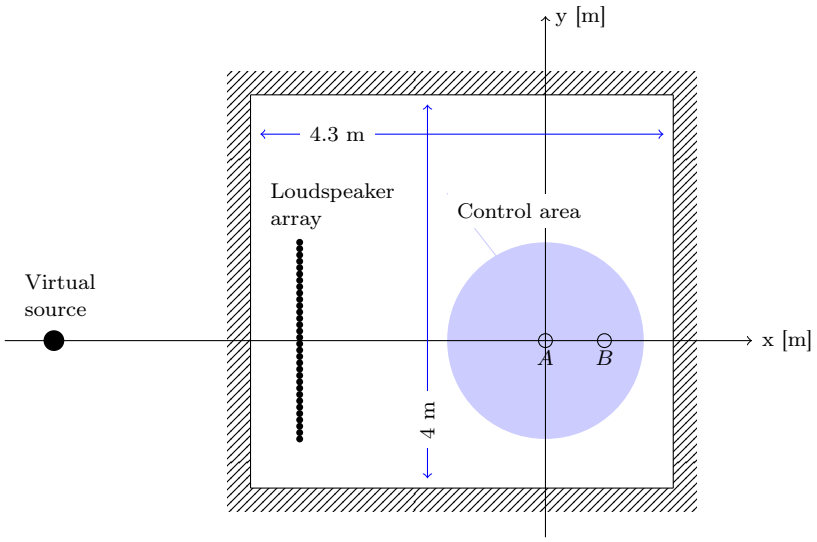


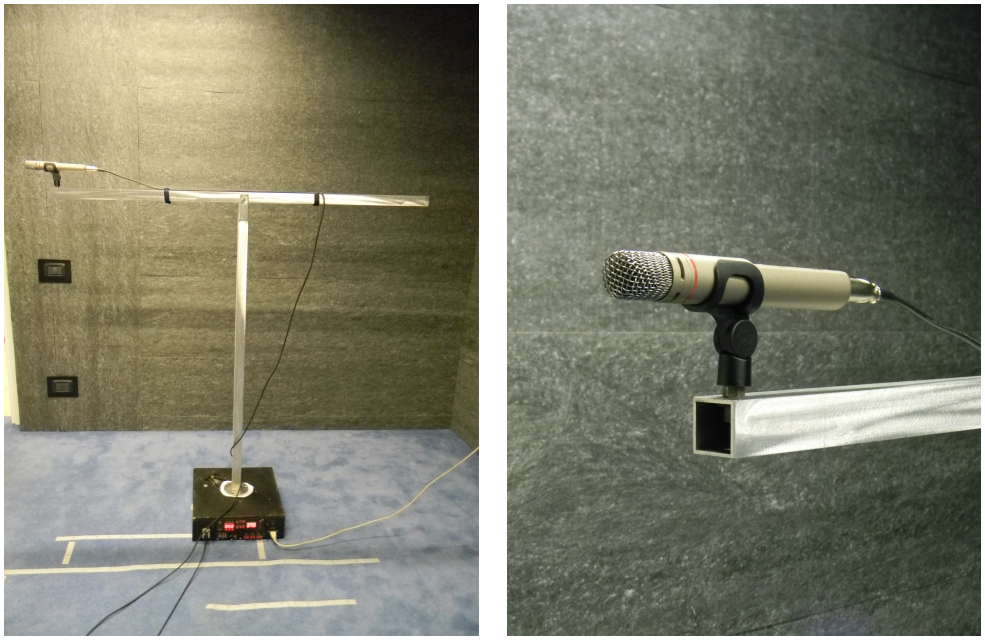
Figure 6.2: Geometry of the room and positioning of the installed sound reproduction system.

mounted on a rotating rig driven by a stepper motor,

- a microphone preamplifier, which amplifies the signal coming from the microphone and provides the necessary electrical power to the microphone (*phantom power*),
- an A/D and D/A converter which outputs the signals driving the loudspeaker array and acquires the signal coming from the microphone,
- an audio interface that allows the communication between the A/D and D/A converter and the measurement computer.

The core component of the system is the measurement computer that generates the signals to drive the loudspeaker array and simultaneously captures the microphone signal. Further, it computes the space-time impulse response and the space-frequency representation of the sound field in an off-line fashion. All these tasks are performed by a MATLAB-based environment comprehending a custom-made software driver for the stepper motor and the *Playrec* [49] utility to access the audio interface via *PortAudio* [50].

The virtual circular microphone array is shown in Figure 6.3a. It is realized with two aluminium rods soldered in a "T"-like structure that is driven by a stepper motor. This motor has a precision of 1° and its is controlled by a custom-made controller interfaced with the measurement computer via a serial interface.



(a) *Rotating device implementing the virtual circular microphone array.*

(b) *AKG C1000s microphone mounted at one end of the rotor.*

Figure 6.3: Virtual circular microphone array.

At one end of the horizontal aluminum rod we have mounted a cardioid microphone (Figure 6.3b), resulting in a measurement radius $\rho_0 = 0.85$ m. In order to balance the weight of the microphone, we have placed a counterweight on the opposite end of the rod. The employed microphone is an *AKG C1000s* condenser microphone, whose specifications are reported in Table A.2. The use of a single cardioid microphone oriented towards the direction orthogonal to the virtual circumference traced by the microphone is prescribed by the measurement methodology presented in Chapter 3.

The signal coming from the microphone is then amplified by a *Focusrite Octopre LE* microphone preamplifier, whose specifications (regarding only the preamplifier module) are reported in Table A.3.

The A/D and D/A converters employed to output the signals driving the loudspeaker array and to acquire the microphone signal are two *Aurora Lynx 16* that make available 32 input and 32 output channels. The specifications of the A/D and D/A converter *Aurora Lynx 16* are reported in Tables A.4-A.9. The audio interface used to enable the communication between the A/D and D/A converters and the measurement computer is an *RME HDSPE AES/EBU*.

Simulative Setup

The experimental setup described in the previous paragraph is replicated also for the simulations. However, in the simulative scenario we are interested in evaluating the performance of the system in an ideal scenario so we make some assumptions. First, we assume that sound propagation occurs in an ideal environment; thus, we have not modeled the reverberation time of the real room. Furthermore, we modeled the loudspeakers composing the array as point-like sources with an omnidirectional directivity pattern.

Considering the measurement system, we modeled the real cardioid microphone as an ideal cardioid resulting from the superposition of a coincident pair of microphones, i.e. a pressure microphone (omnidirectional) and a pressure-gradient microphone (figure-of-8). The acquisition chain (preamplifier and A/D converter) have been assumed to be ideal.

6.2 Computation of the Theoretical Space-Time Impulse Response

In this Section we describe the setup adopted in order to simulate a theoretical sound field, rendered by a technique like Wave Field Synthesis (WFS) or Geometric Rendering (GR). To do so, we first compute the vector of loudspeaker coefficients $h(\omega)$ for reproducing a target sound field in a predefined listening area. This is accomplished by employing Equation 2.17 for WFS, or Equation 2.41 for GR. The theoretical sound field $\mathbf{P}(\omega)$ is then obtained applying the resulting vector of loudspeaker coefficients $\mathbf{h}(\omega)$ to the propagation matrix $\mathbf{G}(\omega)$, i.e.

$$\mathbf{P}(\omega) = \mathbf{G}(\omega)\mathbf{h}(\omega). \tag{6.2}$$

In particular, $\mathbf{G}(\omega)$ contains the free-field Green's functions (Equation 2.12) from the loudspeakers to the points in the listening area. The computation of $\mathbf{h}(\omega)$ is performed on a per-frequency basis, sampling the frequency axis in 1024 points between 0 Hz and $f_s/2$, where $f_s = 44.1$ kHz is the sampling frequency. In our simulations and experiments we consider a circular listening area (see Figure 6.2) with radius $r = 0.8$ m, centered at the origin of the coordinate system, and sampled on a grid of 201×201 points.

The matrix $\mathbf{P}(\omega)$ is a space-frequency representation of the soundfield. We are also interested in obtaining a space-time representation (i.e., the space-time impulse response), which can be approximated through an Inverse Fast Fourier

Transform (IFFT) of $\mathbf{P}(\omega)$, obtaining $h(t, \rho, \phi)$. Notice that, in order to obtain a real impulse response, we need each row of $\mathbf{P}$ to be hermitian symmetric.

6.3 Measurement of the Space-Time Impulse Response

In this paragraph we describe the implementation of the measurement methodology presented in Chapter 3. The parameters of the implementation and the environmental variables correspond to those of the setup described in Sections 6.1.1 and 6.1.2.

Both simulations and experiments have been performed considering a scenario in which a single omnidirectional virtual source is rendered by the sound field rendering system with no room compensation. Referring to the geometry presented in Figure 6.2, we placed the omnidirectional virtual source at point $(-5, 0)$ m.

As highlighted in [38], the rendering engines that we have adopted define a solution $H_m(\omega)$ for loudspeaker filter coefficients in frequency domain. Hence, in order to use these filters in practical systems, we need to turn them into time-domain discrete filters $h_m(t)$. To implement GR, we sampled the frequency axis at 135 points between 100 Hz and 3 kHz, which results in a frequency resolution $\Delta f \approx 21.5$ Hz at a sampling frequency of $f_s = 44.1$ kHz.

The excitation signal adopted in our procedure is a pseudo-random *Maximum Length Sequence (MLS)* of order 17, which results in a signal $s(t)$ of length 131071 samples = 2.9721 s. The signal $s(t)$ is then filtered through the rendering filter $h_m(t)$ and the resulting signals are sent to the loudspeaker array.

The sound field produced by the loudspeaker array is sampled with the virtual circular microphone array previously described. The angular resolution is fixed to 2° which leads to 180 acquisition points on the circumference. The signals $p(t, \rho_0, \phi_i)$ captured by the microphone at each position are acquired by the measurement computer for the off-line processing stage aimed to reconstruct the sound field in the entire listening area, according to the methodology presented in Chapter 3.

In particular, the signals $p(t, \rho_0, \phi_i)$ are first converted into impulse responses $h(t, \rho_0, \phi_i)$ using the *Hadamard Transform*, as described in [11]. Since the CHD-based interpolation and extrapolation procedure presented in Chapter 3 works in the space-frequency domain, we need to transform the impulse responses $h(t, \rho_0, \phi_i)$ with a Fast Fourier Transform (FFT), thus obtaining $h(\omega, \rho_0, \phi_i)$. The CHD-based interpolation and extrapolation procedure allows us to obtain

the space-time impulse response $h(\omega, \rho, \phi)$ in the entire listening area, with the frequency axis sampled at 4096 frequencies between 0 Hz and $f_s/2 = 22.05$ kHz.

Finally, by performing an IFFT on the sound field $h(\omega, \rho, \phi)$ we can obtain the simulative space-time impulse response of the sound field rendering system $h_s(t, \rho, \phi)$ and the measured one $h_m(t, \rho, \phi)$.

6.4 Results

6.4.1 Rendered Sound Fields

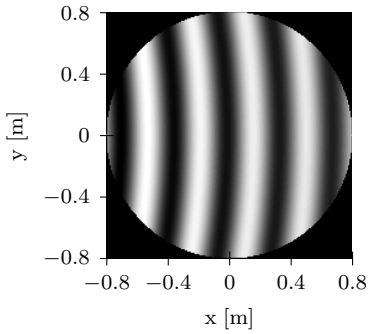
In this paragraph we show the simulated sound fields rendered by WFS and GR, driven to emulate the presence of a single omnidirectional virtual source placed at point $(-5, 0)$ m, according to the geometry presented in Figure 6.2. In particular, we show the space-frequency representation of these sound fields at frequency $f_1 = 1$ kHz and we compare them with the target sound field.

Figure 6.4a shows the target sound field, i.e. the sound field obtained by modeling the wave propagation in an ideal fashion. This is the sound field that we want to reproduce with a sound field rendering system.

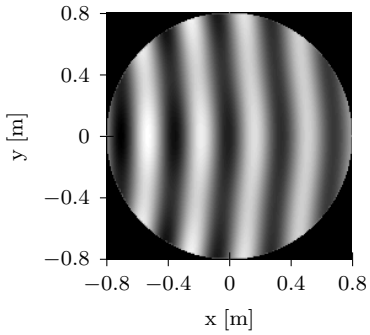
As we notice from Figures 6.4b and 6.4c even the ideal simulations of WFS and GR techniques present some deviation with respect to the target sound field. In particular, Figure 6.4b shows the simulated sound field produced by a rendering system implementing WFS with the setup described in Section 6.1. We notice that the wave fronts are not perfectly reproduced and even their shape deviates from the target case. On the other hand, Figure 6.4c shows the simulated sound field produced by GR. In this case the shape of the wave fronts is exactly reproduced, but we notice a progressive lack of definition while increasing the distance from the loudspeaker array.

The sound fields shown in Figure 6.4 give a visual proof of the artifacts introduced by sound field rendering techniques even in an ideal scenario. Indeed, the only non-idealities considered here are those regarding the finite aperture of the loudspeaker array and the fact that it is not a continuous distribution of sources.

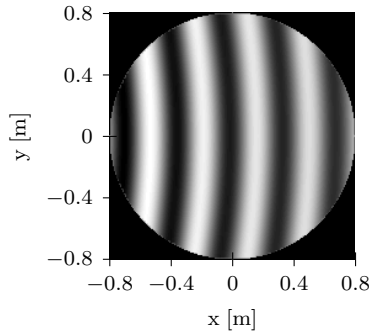
In the next paragraphs we will evaluate these artifacts employing the methodology introduced in Chapter 4.



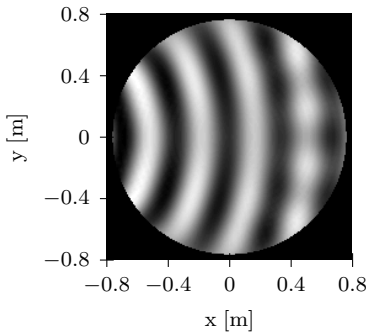
(a) *Target sound field.*



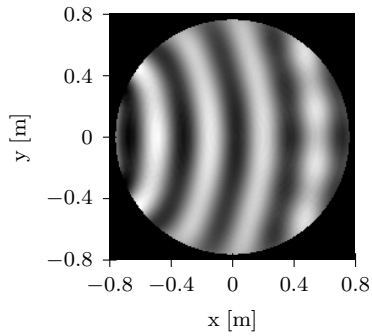
(b) *Theoretical sound field (WFS).*



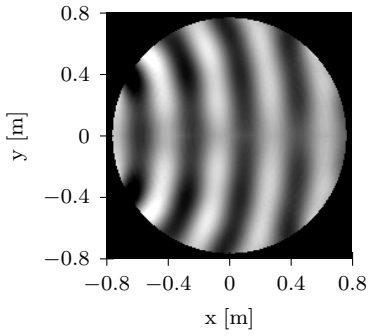
(c) *Theoretical sound field (GR).*



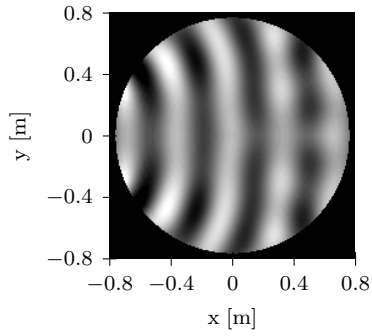
(d) *Simulative sound field (WFS).*



(e) *Simulative sound field (GR).*

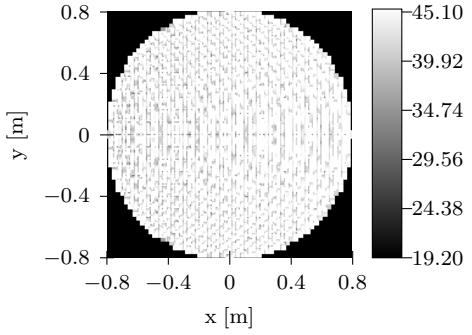


(f) *Measured sound field (WFS).*

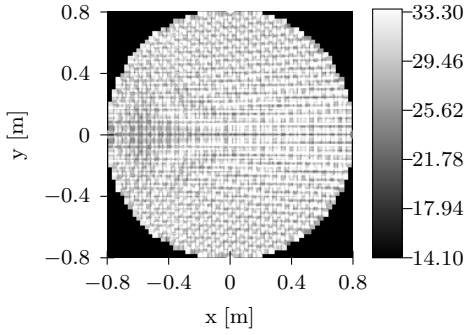


(g) *Measured sound field (GR).*

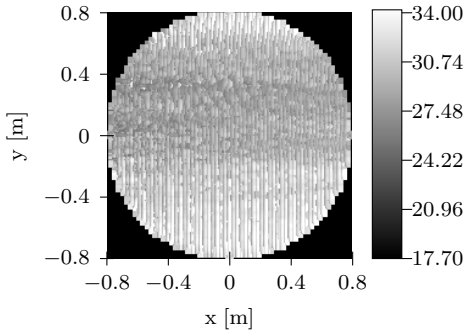
Figure 6.4: Comparison of target and simulated theoretical sound fields



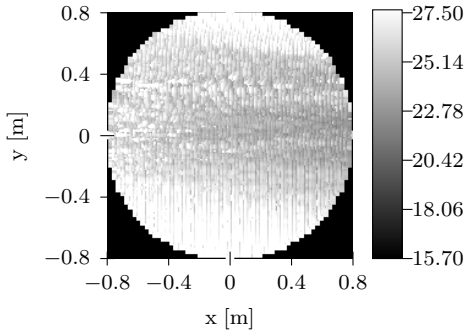
(a) PER$^-$, *theoretical sound field (WFS).*



(b) PER$^+$, *theoretical sound field (WFS).*



(c) PER$^-$, *theoretical sound field (GR).*



(d) PER$^+$, *theoretical sound field (GR).*

Figure 6.5: Value of the metric PER$^-$ and PER$^+$ over the measurement area.

6.4.2 Timbral Artifacts

In this section we present the results of the methodology introduced in Section 4.1 to evaluate the timbral artifacts introduced by a sound field rendering system under an objective and a psychoacoustic-based point of view. At first we present the results of the objective evaluation, then in a following paragraph we present the results of a psychoacoustic-based evaluation.

Objective Evaluation

In this paragraph we show the results of the application of the metrics *Peak-to-(pre)-Echo Ratio* PER$^-$ and *Peak-to-(post)-Echo Ratio* PER$^+$ on simulative sound fields.

In particular, in Figure 6.5 we show PER$^-$ and PER$^+$ as a function of the position inside the measurement area for sound field simulated with the procedure presented in Section 6.2.

It appears to be clear, even from a rapid analysis of the Figures 6.5a-6.5d, that an objective evaluation is not suitable for analyzing the behavior of pre-

echoes and post-echoes inside the listening area, due to the high variability of the results. For this reason, we resort a psychoacoustic-based evaluation, reported in the next paragraph.

Psychoacoustic-based Evaluation

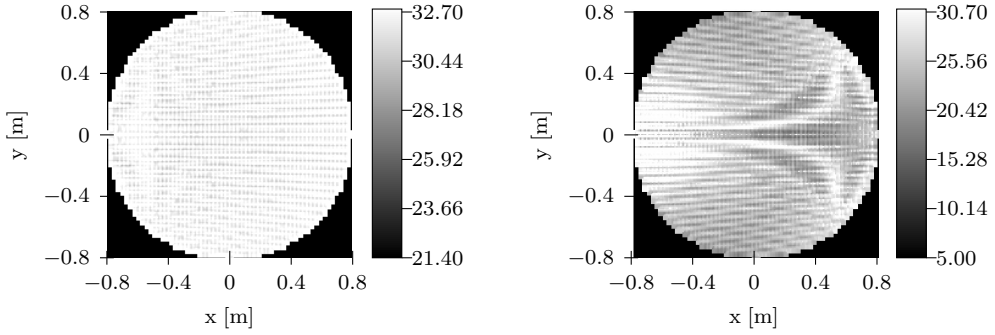
Since we want to assess the impact of pre-echoes and post-echoes on the perception of timbral quality inside the listening area, we apply the psychoacoustic-based metrics *Direct-to-(pre)-Echo Ratio* DER^- and *Direct-to-(post)-Echo Ratio* DER^+ on simulative and experimental sound fields. These metrics have a great advantage over PER^- and PER^+ since they take human perception into account. As widely explained in Chapter 4, the metrics DER^- and DER^+ are based on the temporal masking effect.

In particular, in Figure 6.6 we show the values of the metrics DER^- and DER^+ for all the positions inside the measurement area. The theoretical sound fields have been simulated with the procedure presented in Paragraph 6.2. Here we notice that, in general, the values of DER^- (Figures 6.6a and 6.6c) are approximately constant over the whole measurement area; this means that we expect the impact of pre-echoes to be independent on the position of the listener. On the other hand, considering DER^+ (Figures 6.6b and 6.6d) we notice a behavior that is dependent on the position; this behavior will be clear when we will analyze the results of DER^+ for a measured sound field.

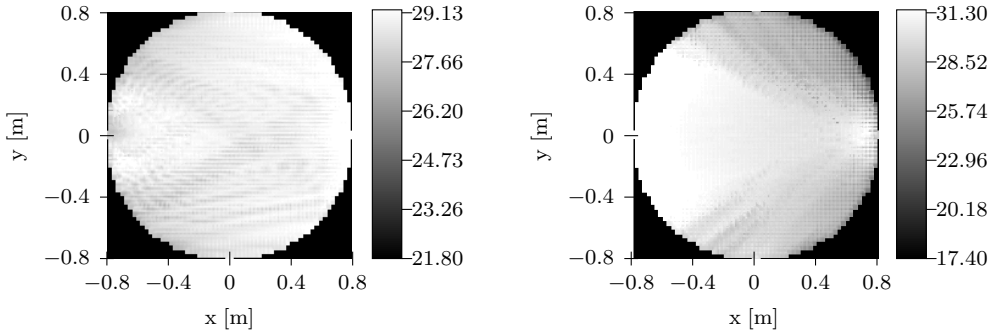
Through the comparison of Figure 6.7 with Figure 6.6 we will be able to characterize the inaccuracies introduced by the measurement technique on the metrics DER^- and DER^+.

From Figure 6.7a we notice that the values of DER^- for WFS seem to be lower in a precise region inside the listening area, i.e. far from the loudspeaker array, while for GR the distribution of values is almost constant over the whole measurement area. On the other hand, by analyzing the distribution of DER^+ we notice that for GR we have a slight degradation in the region far from the loudspeaker array. From the comparison with Figure 6.6 we can conclude that, in general, the measurement technique tends to affect the region of space far from the loudspeaker array. Notice that we are not able to separate the contributions of pre-echoes and post-echoes in the perception of timbral quality, so the only result we can expect from the listening tests will be that they will reveal a degradation of the perceptual performance in the region far from the loudspeaker array for both the rendering engines.

Finally, in Figure 6.8 we show the values of the metrics DER^- and DER^+



(a) DER$^-$, *theoretical sound field (WFS).* **(b)** DER$^+$, *theoretical sound field (WFS).*

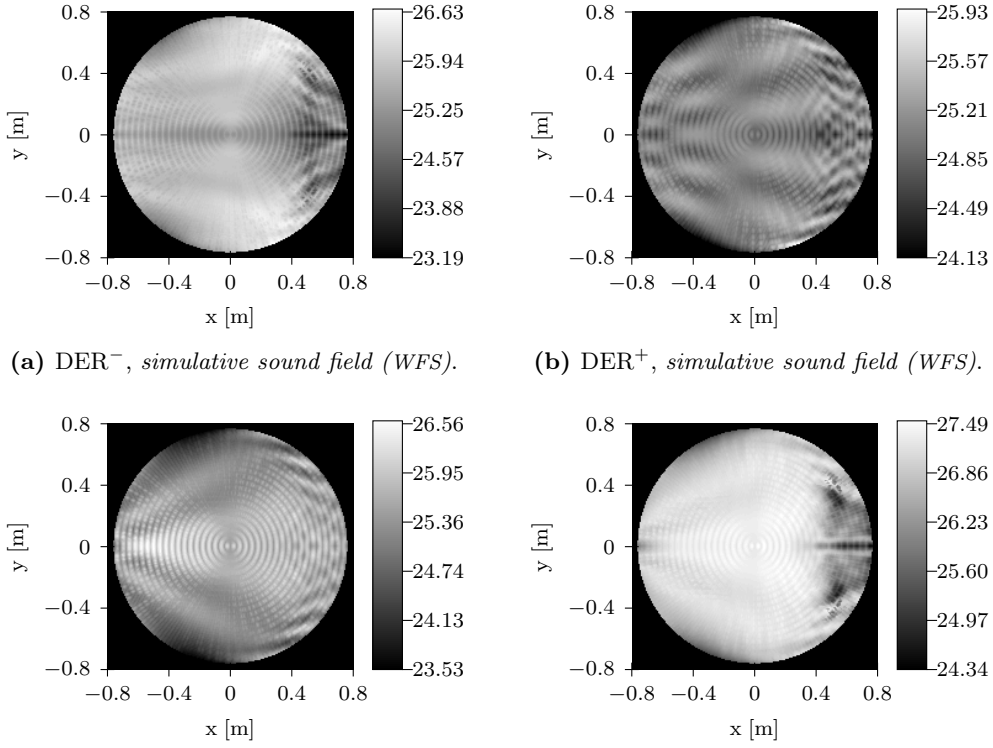


(c) DER$^-$, *theoretical sound field (GR).* **(d)** DER$^+$, *theoretical sound field (GR).*

Figure 6.6: Value of the metrics DER$^-$ and DER$^+$ over the measurement area for theoretical sound fields rendered by WFS and GR.

for sound fields measured in the environment described in Paragraph 6.1.1 with the procedure presented in Paragraph 6.3. The values of DER$^-$ for both WFS (Figure 6.8a) and GR (Figure 6.8c) exhibit a behavior that is approximately constant over the whole measurement area, except for some isolated black spots on the right (i.e., in the region far from the loudspeaker array). Most of these artifacts are introduced by the measurement technique, as noticed in Figure 6.7. As a consequence, neglecting such artifacts, we can state that pre-echoes due to the real acoustic environment will impair the perception of timbral quality, but independently from the listener position.

On the other hand, observing the values of DER$^+$ in Figures 6.8b and 6.8d, we notice that the artifacts introduced by the measurement methodology are still present. However, in this case, we observe that the values of DER$^+$ depends on the listener position, with a degradation that is more evident far from the loudspeaker array and close to the wall. In particular, the strong degradation in the region between the points $(0.5, -0.4)$ m and $(0.5, 0.4)$ m in Figures 6.8b and 6.8d are due to the presence of a close wall, as it can



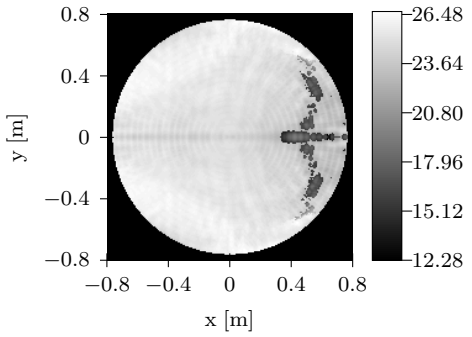
(a) DER$^-$, *simulative sound field (WFS).* **(b)** DER$^+$, *simulative sound field (WFS).*

(c) DER$^-$, *simulative sound field (GR).* **(d)** DER$^+$, *simulative sound field (GR).*

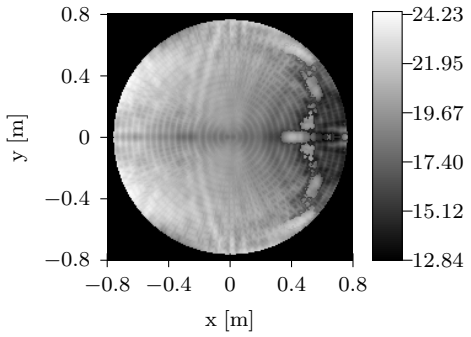
Figure 6.7: Value of the metric DER$^-$ and DER$^+$ over the measurement area for simulative sound fields rendered by WFS and GR.

be clearly noticed looking at the geometry of the rendering room depicted in Figure 6.2. Thus we can state that, even in a semi-anechoic environment, the reflections of the walls may cause post-echoes that will impair the perception of timbral quality.

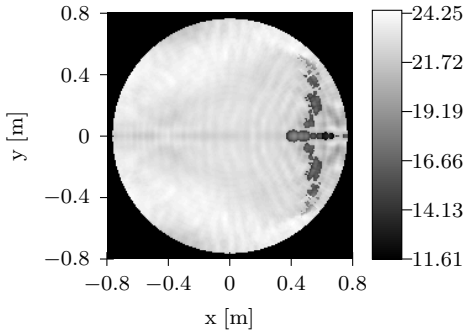
From a direct comparison of the results shown in Figures 6.6 and 6.8 we can conclude that artifacts due to the real acoustic environment (above all the reflections of the walls) will considerably affect the perception of quality, so that post-echoes are the most important time-domain artifacts to be considered in real-world sound field rendering applications.



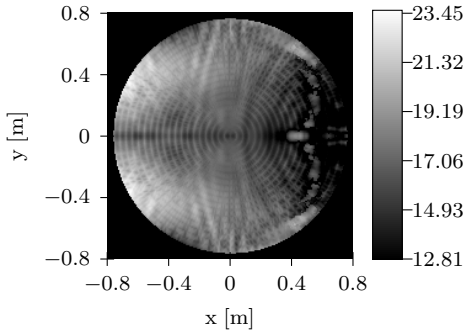
(a) DER^{-}, *measured sound field (WFS).*



(b) DER^{+}, *measured sound field (WFS).*

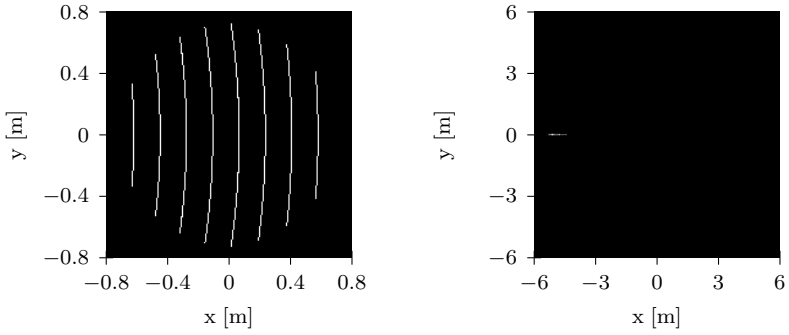


(c) DER^{-}, *measured sound field (GR).*



(d) DER^{+}, *measured sound field (GR).*

Figure 6.8: Value of the metric DER^{-} and DER^{+} over the measurement area for experimental sound fields rendered by WFS and GR.



(a) *Target sound field (edge image).* **(b)** *Target sound field (2-d accumulator space).*

Figure 6.9: Localization methodology applied on the target sound field (frequency 1 kHz).

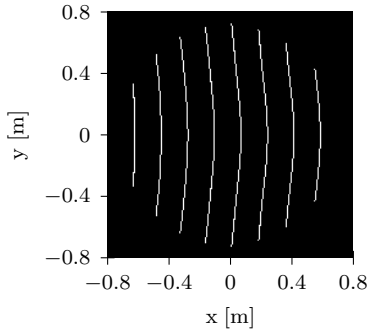
6.4.3 Localization Artifacts

In this section we present the results of the evaluation methodology introduced in Section 4.2 to assess the impact of rendering artifacts on the localization of a single virtual source.

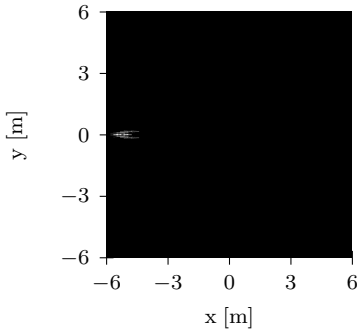
In Figure 6.9a we show the space-frequency image of the target sound field at frequency 1 kHz. The pre-processing stage described in Section 4.2 is applied in order to obtain an edge image with ones on the points where the intensity of the gradient is maximum. Notice that the shapes of the wave fronts are correctly extracted. On the other hand, Figure 6.9b shows the two-dimensional accumulator space for the sound field image of Figure 6.9a. Thanks to the specific processing procedure described in Section 4.2, the candidate points are all clustered in small region around the point $(-5, 0)$ m, i.e. the position of the virtual source.

Considering a sound field obtained with the simulative procedure presented in Paragraph 6.2, in Figure 6.10 we show the results of the localization methodology applied on sound fields generated by WFS (Figures 6.10a and 6.10b) and GR (Figures 6.10c and 6.10d). We notice that the shape of the wave fronts is slightly distorted by WFS and this artifact causes a slightly more sparse clustering in Figure 6.10b. This artifact arises because of the finite aperture of the loudspeaker distribution adopted for these simulations, which deviates from the theoretical formulation of WFS (see Section 2.3). On the other hand, the sound field generated by GR appears to be less affected by such non ideality since in its formulation no assumptions are made on the geometry of the loudspeaker distribution.

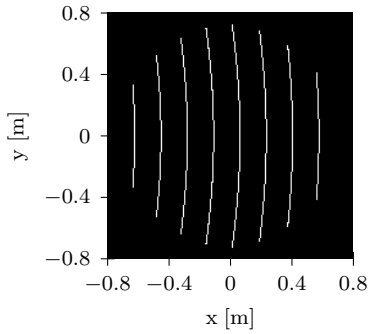
Figure 6.11 shows the localization methodology applied on sound field



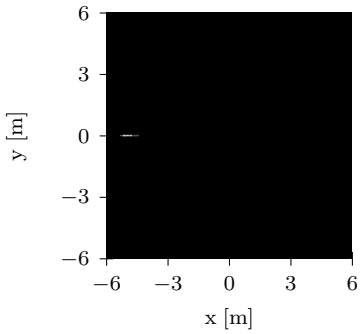
(a) *Theoretical sound field WFS (edge image).*



(b) *Theoretical sound field WFS (2-d accumulator space).*



(c) *Theoretical sound field GR (edge image).*

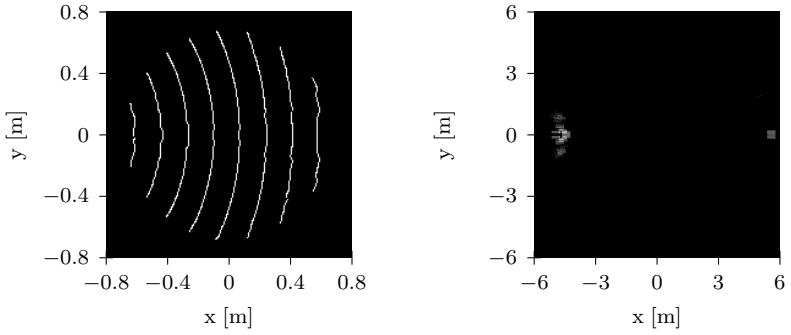


(d) *Theoretical sound field GR (2-d accumulator space).*

Figure 6.10: Localization methodology applied on theoretical sound fields at frequency 1 kHz.

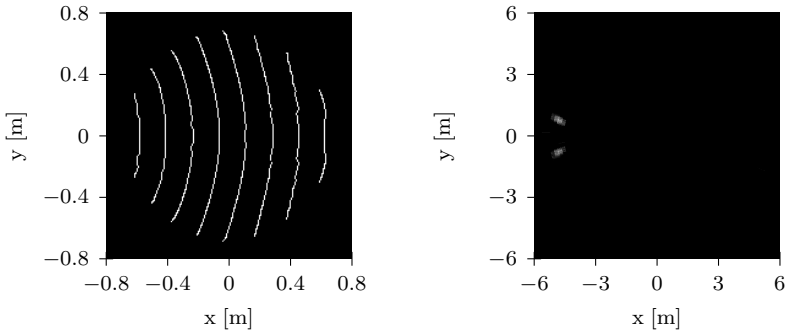
simulated with the intend of modeling the measurement procedure. From the edge images (Figures 6.11a and 6.11c) we notice that the wave fronts appear to be more curved than in the target and in the theoretical sound fields; this deviation of the shape of the wave fronts causes an erroneous localization of the virtual sound source, which is localized closer to the listener. Moreover, other distortions of the shape of the wave fronts cause a dispersion of the candidate points in the accumulator spaces of Figures 6.11a and 6.11c, which causes a less precise identification of the source position.

In order to highlight the deviations introduced by the real environment (described in Paragraph 6.1.1), in Figure 6.12 we show the results of the localization methodology applied on measured sound fields. We notice that the measurements show a result similar to the one obtained with a simulation of the measurement methodology (Figure 6.11). Thus, neglecting the artifacts introduced by the measurement technique, we expect that the sound source is rendered in the correct position.



(a) *Simulated sound field WFS (edge image).*

(b) *Simulated sound field WFS (2-d accumulator space).*



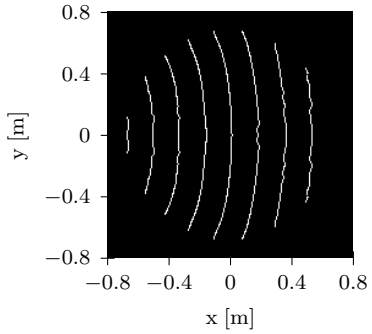
(c) *Simulated sound field GR (edge image).*

(d) *Simulated sound field GR (2-d accumulator space).*

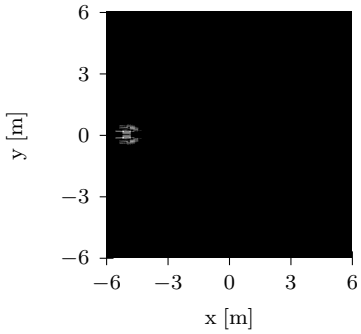
Figure 6.11: Localization methodology applied on sound fields simulated to reproduce the measurement procedure. The frequency is 1 kHz.

The precision of the localization reveals to be considerable, even for real-world data. In Table 6.1 we show the results of the localization methodology applied on the sound field shown in Figure 6.9-6.12. We recall that all the considered sound fields were generated by a virtual source placed at point $S = (-5, 0)$ m. The resolution of the localization methodology is fixed to 0.001 m.

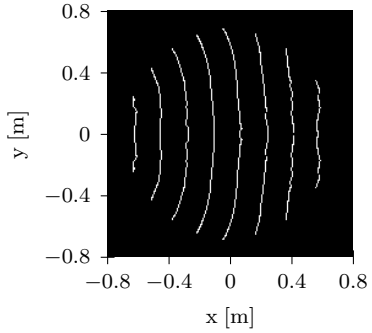
We notice that in all cases the position of the virtual source is estimated with an error at most in the order of tenth of centimeter. In particular, we highlight the fact that WFS seems to deliver a more precise spatial impression than GR. In the next Section we will show that these conclusions are supported by subjective results.



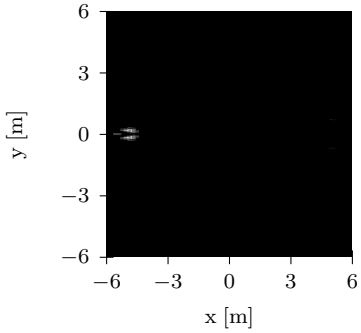
(a) *Measured sound field WFS (edge image).*



(b) *Measured sound field WFS (2-d accumulator space).*



(c) *Measured sound field GR (edge image).*



(d) *Measured sound field GR (2-d accumulator space).*

Figure 6.12: Localization methodology applied on measured sound fields at frequency 1 kHz.

Table 6.1: Estimated position of the virtual sound source for target, simulative and measured sound fields.

Sound field	**Estimated position** $\hat{S}$ (m)	**Error** $\|S - \hat{S}\|$ (m)
Target (Figure 6.9a)	$(-4.91, 0.002)$	0.09
Theoretical WFS (Figure 6.10a)	$(-5.12, 0.002)$	0.12
Theoretical GR (Figure 6.10c)	$(-4.89, 0.002)$	0.11
Measured WFS (Figure 6.12a)	$(-5.17, -0.002)$	0.17
Measured GR (Figure 6.12c)	$(-4.70, 0.002)$	0.30

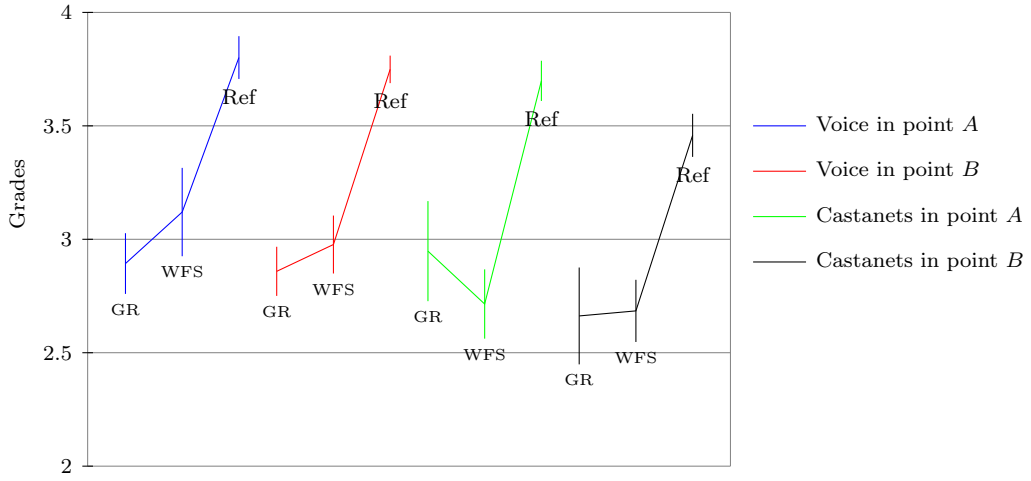


Figure 6.13: Result of the listening tests aimed at assessing the impact of rendering artifacts on the perceived timbral quality.

6.4.4 Listening Tests

In this section we present the results of the formal listening tests conducted in order to validate the results presented in the previous paragraphs. In particular we will show the subjective evaluation of both timbral and localization artifacts, then we will correlate these results with the psychoacoustic-based evaluation methodology presented in Chapter 4.

Timbral artifacts

One subjective experiment evaluates the timbral quality of sound stimuli generated by a sound field rendering system by compairing them to an unimpaired reference and to a distorted reference. The details of the test method have been presented in Chapter 5. For the purpose of a timbral evaluation, the rendering system emits stimuli that are all generated by a virtual source placed at point $-5,0$ m.

In Figure 6.13 we show the results of the subjective test aimed at assessing the impact of rendering artifacts on the perceived timbral quality. We recall from Section 5.2 that we hav adopted two different sound excerpts: a vocal signal (*Tom's Diner* by Suzanne Vega) and a percussive signal (castanets from from EBU SQAM CD). For each listener we repeated the experiment in two positions inside the listening area; referring to the geometry depicted in Figure 6.2, the experiment have been repeated with the listener placed in points A and B. The figure shows the mean value of the normalized scores expressed by all subject and the confidence interval computed with the procedure explained

in Paragraph 5.2.3. The different colors indicate the different test conditions. In particular, the blue line reports the results of the timbral evaluation for the vocal signal, with the listener placed in point A (i.e. in the center of the listening area). The red line refers to the results for the same vocal signal but with the listener placed in point B, i.e. closer to a wall and far from the loudspeaker array. The green and the black lines show the results for the percussive signal with the listener placed in points A and B, respectively.

From an analysis of the results, we notice that the unimpaired reference signal is identified clearly by all subjects. Considering the sound stimuli rendered with both WFS and GR techniques, the confidence interval is bigger than the difference of the average values in all cases, so we cannot make a precise comparison of the two techniques. Despite of the confidence interval, one consideration that emerges clearly from Figure 6.13 is that the position of the listener inside the listening area affects the perceived timbral quality. As we have already highlighted in Section 6.4.2 while discussing the results of the psychoacoustic-based evaluation methodology, it is clear that the reflections of the walls produce noticeable post-echoes that impair the perceived timbral quality in a considerable way. This effect is clearly revealed by the results referred to the percussive signal, where we notice a clear downgrading even for the unimpaired reference signal.

Localization Artifacts

In Figure 6.14 we show the results of the subjective test aimed at the localization of an acoustic virtual source. This test has been conducted according to the guidelines presented in Chapter 5; in particular, the test method has been presented in Section 5.3. The figures 6.14a-6.14d show the mean of the judgments expressed by all the listeners and the relative confidence interval, computed with the procedure presented in Paragraph 5.2.3. In particular, we show the mean result and the confidence interval for three stimuli that have been presented to the listeners. These stimuli are generated by a virtual source placed at the same distance (5 m) but with different incoming angles: $-20°$, $2°$ and $17.5°$. We notice from an analysis of the results that the listeners are able to make judgments on the angular position of the virtual source with sufficient precision, despite of the presence of rendering artifacts. Moreover, no significant difference can be observed adopting WFS rendering technique or GR.

We want to remark that some listeners signaled some ambiguity in the localization of the source coming from a direction of $-20°$. This ambiguity did

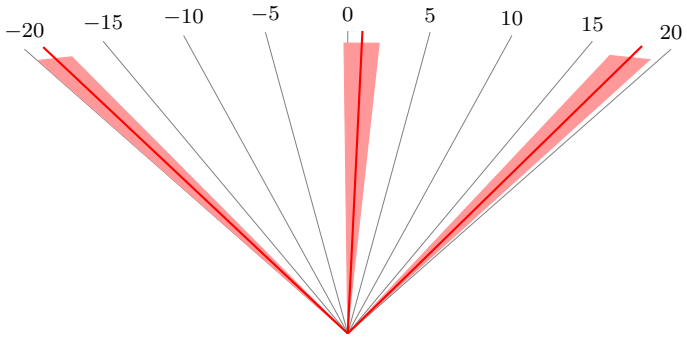
not prevent the listener to make a precise judgment, but it is important to be observed that it is due to the positioning of the virtual source close to the left limit of the loudspeaker array, so the wave fronts cannot be reproduced correctly.

Considering the vocal signal, we notice that the localization is almost correct for both WFS (Figure 6.14a) and GR (Figure 6.14b), but for GR the sound source placed at 17.5° is localized as incoming from a slightly minor angle. On the other hand, considering the percussive signal, we notice that the sound stimuli incoming from directions 2° and 17.5° are correctly localized, while the incoming angle of the sound source placed at 20° is underestimated in both WFS (Figure 6.14c) and GR (Figure 6.14d) cases.

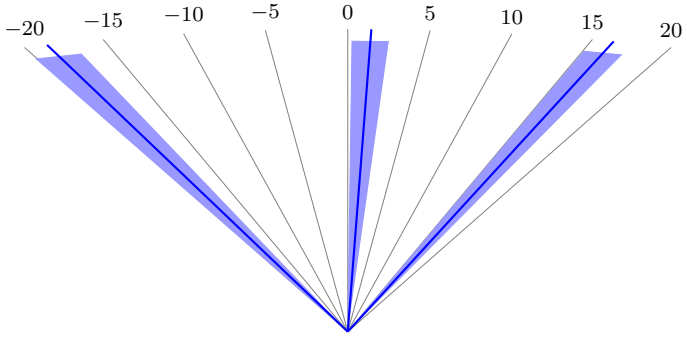
6.5 Conclusions

In this chapter we have presented the simulative and experimental results of the psychoacoustic-based evaluation methodology presented in Chapter 4 and of the subjective evaluation methodology described in Chapter 5. In particular, we have presented a comparison between a theoretical scenario and a simulation of the measurement procedure, in order to assess the impact of the errors introduced by the measurement methodology. Then we considered a real scenario, in which we have performed the measurements of a real sound field produced by a rendering system and then we have conducted the listening tests.

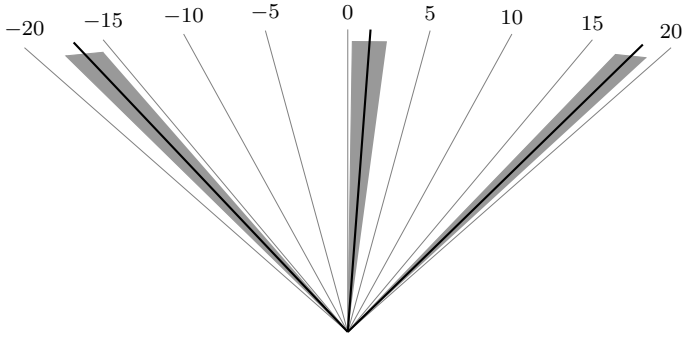
The comparison between the simulative and experimental results allowed us to assess the impact of the non-idealities arising from a practical implementation of a sound field rendering system operating in a real environment. Finally, the comparison of simulative, experimental and subjective results allows us to validate the psychoacoustic-based analysis introduced in Chapter 4.



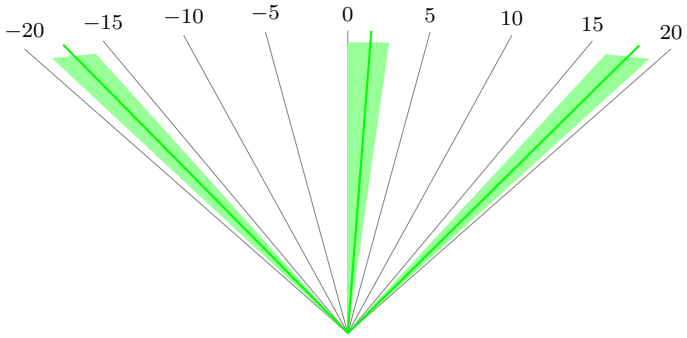
(a) *Voice (WFS).*



(b) *Voice (GR).*



(c) *Castanets (WFS).*



(d) *Castanets (GR).*

Figure 6.14: Result of the listening tests aimed at assessing the impact of rendering artifacts on the localization of a virtual source.

Chapter 7

Conclusions and Future Works

In this thesis we have proposed a psychoacoustic-based methodology to assess the impact of rendering artifacts on the perceptual performance of sound field rendering techniques.

In particular, we noticed that the practical realization of sound field rendering systems always imposes some constraints: the loudspeaker array only approximates a spatially continuous distribution of sound sources; the directivity pattern of the loudspeakers is not omnidirectional; D/A converters, power amplifiers and loudspeakers are all limited in bandwidth and their frequency response is not flat. The sum of all these constraints causes artifacts in the rendered sound field, which presents some distortions with respect to the desired one. Moreover, when the rendering system is operating in a real environment, even reflections and reverberation contribute to alter the rendered sound field.

In the literature we have found two classes of approaches aimed at assessing the impact that sound field distortions have on human perception. On the one hand there are objective approaches, which evaluate the differences of a measured sound field with respect to the desired one. These methods, however, are not informative enough since they do not take into account human perception.

On the other hand there are subjective approaches based on formal listening tests. Since the judgments of a human listener are involved, the results of these tests take human perception into account. However, their high cost prevents their use in a large number of context.

In this thesis we have proposed an alternative approach that provides a psychoacoustic evaluation obtained with metrics that are representative of human perceptive experience. In particular, we have considered two classes of artifacts: time-domain and localization artifacts and we have validated the

results of our psychoacoustic-based evaluation with subjective tests.

The proposed evaluation methodology relies on measurements of the sound field. In order to perform these measurements we have adopted a well known measurement procedure that employs a virtual circular array to sample the sound field on a circumference. Then the decomposition of the sound field into circular harmonics is exploited to reconstruct the sound field over the whole measurement area.

Considering time-domain artifacts, we have shown that the considered rendering techniques produce pre-echoes and post-echoes that alter the desired impulse response of the rendering system. Basing on a temporal masking curve, we have presented two metrics to assess the impact of pre-echoes and post-echoes on human perception. The results of these metrics are coherent with the judgments collected from listeners involved in formal listening tests. In particular, we have shown that post-echoes are the most important time-domain artifact that affects human perception. We notice that both the metrics and the subjective results highlight that the perceived amount of post-echoes is dependent on listener position, i.e. they are concentrated in the region close to the walls, due to the reflective behavior of the walls of the room in which both measurements and listening tests have been conducted.

On the other hand, considering localization artifacts we have shown that the wave fronts produced by a real rendering system are distorted by truncation effects and spatial aliasing, due to the finite and discrete nature of the loudspeaker array. In this thesis we have proposed a methodology to retrieve the position of a virtual source given the curvature of the rendered wave fronts. The proposed methodology is based on a generalized Hough Transform and allows to precisely determine the position of the virtual sound source. The listening tests conducted in order to validate the results of the evaluation have confirmed the reliability of our methodology.

Future works

In this thesis we have proposed a methodology to evaluate the quality of sound field rendering systems in a perceptually-related and cost-effective fashion. The proposed methodology may be applied in order to face with two different issues.

On the one hand, our methodology may be employed to assess the impact of rendering artifacts introduced by a simplification of the geometric model of the environment. In particular, an open issue in the geometrical modeling of

sound propagation in an environment is to decide which propagation paths can be neglected. Our methodology can be used to assess the impact of these approximations. In this way it can be ensured that, while reducing the complexity of the modeling, the perceptual qualities of the sound scene are preserved.

On the other hand, the quantitative description of masking in time-domain provided in this thesis can be exploited in order to hide time-domain rendering artifacts in such a way that they do not alter the perceived timbral quality. This is a strategy widely adopted in the field of perceptual audio coding and this thesis supports an extension of this strategy to sound field rendering.

Appendix A

Technical Specifications

A.1 Loudspeakers

Table A.1: Experimental setup: loudspeakers specifications.

Model	**Empire M2**
Total Power Output	RMS 3 W $\times$ 2
Total Harmonic Distortion	10%
Signal-to-Noise ratio	$\geq$ 90 dBA
Input Impedance	15 kΩ
Driver	$2'', 8\ \Omega$

A.2 Microphone

Table A.2: Experimental setup: microphone *AKG C1000s* specifications.

Model	**AKG C1000s**
Polar pattern:	cardioid hypercardioid (optional)
Frequency range:	50 Hz to 20 kHz
Sensitivity:	6 mV/Pa (−44 dBV)
Max SPL for 1% THD:	137 dB
Signal-to-Noise ratio (A-weighted):	73 dB
Impedance:	200 Ω
Powering:	9 V to 52 V phantom power
Connector:	3-pin XLR
Dimensions (diameter):	34 mm
Dimensions (length):	220 mm

A.3 Microphone Preamplifier

Table A.3: Experimental setup: microphone preamplifier *Focusrite Octopre LE* specifications.

Model	**Focusrite Octopre LE**
Gain:	+13 dB to +60 dB
Input Impedance:	2.5 kΩ
EIN:	124 dB @ 60 dB Gain
THD+N @ Min Gain (+13 dB):	0.0006% with 0 dBu input
THD+N @ Max Gain (+60 dB):	0.003% with −36 dBu input
THD+N @ Max Input (+9 dBu):	0.0008%
Frequency Response:	−0.4 dB @ 10 Hz & −3 dB @ 122 kHz
CMRR @ Max Gain (+60 dB):	80 dB

A.4 A/D and D/A Converter

Table A.4: Experimental setup: *Aurora Lynx 16* specifications (Analog I/0).

Model	**Aurora Lynx 16**
	16 inputs and 16 outputs
Type:	Electronically balanced or unbalanced
Level:	+4 dBu nominal / +20 dBu max.
	or −10 dBV nominal / +6 dBV max
Input Impedance (balanced mode):	24 kΩ
Input Impedance (unbalanced mode)	12 kΩ
Output Impedance (balanced mode):	100 Ω
Output Impedance (unbalanced mode)	50 Ω
Output Drive:	600 Ω impedance, 0.2 µF capacitance
A/D and D/A Type:	24 bit multi-level, delta-sigma

Table A.5: Experimental setup: *Aurora Lynx 16* specifications (Analog In performance).

Model	**Aurora Lynx 16**
Frequency response:	20 Hz to 20 kHz, +0/ − 0.1 dB
Dynamic range:	117 dB (A-weighted)
Channel crosstalk:	−120 dB maximum
THD+N @ −1 dBFS	−108 dB (0.0004%)
THD+N @ −6 dBFS	−104 dB (0.0006%)

Table A.6: Experimental setup: *Aurora Lynx 16* specifications (Analog Out performance).

Model	**Aurora Lynx 16**
Frequency response:	20 Hz to 20 kHz, +0/ − 0.1 dB
Dynamic range:	117 dB (A-weighted)
Channel crosstalk:	−120 dB maximum
THD+N @ −1 dBFS	−107 dB (0.0004%)
THD+N @ −6 dBFS	−106 dB (0.0006%)

Table A.7: Experimental setup: *Aurora Lynx 16* specifications (Digital I/O performance).

Model	**Aurora Lynx 16**
Number and Type:	16 inputs and 16 outputs 24 bit AES/EBU format transformer coupled
Channels	16 in/out in single-wire mode 8 in/out in dual-wire mode
Sample rates:	up to 192 kHz

Table A.8: Experimental setup: *Aurora Lynx 16* specifications (On-board Digital Mixer).

Model	**Aurora Lynx 16**
Type:	Hardware-based, low latency
Routing:	Ability to route any input to any or multiple outputs
Mixing:	up to four input or playback signals mixed to any output, 40 bit
Status	peak levels to −114 dB on all inputs and outputs

Table A.9: Experimental setup: *Aurora Lynx 16* specifications (Connections).

Model	**Aurora Lynx 16**
Digital I/O Ports:	25-pin female D-sub connectors Port A: channels 1-8 I/O Port B: channels 9-16 I/O Yamaha pinout standard
Analog I/O Ports:	25-pin female D-sub connectors Analog In 1-8, Analog In 9-16 Analog Out 1-8, Analog Out 9-16 Tascam pinout standard
External Clock	75 Ω BNC word clock input and output
MIDI	1 In and 1 Out, 5-pin female DIN connectors

Bibliography

[1] R. Rabenstein and S. Spors. *Springer Handbook on Speech Processing*, chapter Sound field reproduction, pages 1095–1113. Springer, 2007.

[2] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778, May 1993.

[3] F. Antonacci, A. Canclini, A. Galbiati, A. Calatroni, A. Sarti, and S. Tubaro. Soundfield rendering with loudspeaker arrays through multiple beamshaping. In *Proc. of 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2009.

[4] R. Rabenstein and S. Spors. Spatial aliasing artifacts produced by linear and circular loudspeaker arrays used for wave field synthesis. In *Audio Engineering Society Convention 120*, 5 2006.

[5] J. Ahrens, S. Spors, and H. Wierstorf. Comparison of higher order ambisonics and wave field synthesis with respect to spatial discretization artifacts in time domain. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, 10 2010.

[6] M. Geier, H. Wierstorf, J. Ahrens, I. Wechsung, A. Raake, and S. Spors. Perceptual evaluation of focused sources in wave field synthesis. In *Audio Engineering Society Convention 128*, 5 2010.

[7] S. Kerber, H. Wittek, H. Fastl, and G. Theile. Experimental investigations into the distance perception of nearby sound sources: Real vs. wfs virtual nearby sources. Number Table 1, pages 2–3, 2004.

[8] S. T. Birchfield and R. Gangishetty. Acoustic localization by interaural level difference. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP*, 2005.

[9] A. Kuntz and R. Rabenstein. Cardioid pattern optimization for a virtual circular microphone array. In *EAA Symposium on Auralization*, pages 1–4, Helsinki, June 2009.

[10] A. Canclini, P. Annibale, F. Antonacci, A. Sarti, R. Rabenstein, and S. Tubaro. A methodology for evaluating the accuracy of wave field rendering techniques. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pages 69–72, 2011.

[11] G. Stan, J. Embrechts, and D. Archambeau. Comparison of different impulse response measurement techniques. *J. Aud. Eng. Soc.*, vol. 50, no. 4:249–262, 2002.

[12] ITU-R BS.1116 Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 10 1997.

[13] ITU-R BS.1534 Method for the subjective assessment of intermediate quality level of coding systems, 1 2003.

[14] A. D. Blumlein. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems, 1933.

[15] M. A. Gerzon. Periphony: With-height sound reproduction. *J. Audio Eng. Soc*, 21(1):2–10, 1973.

[16] D. T. Blackstock. *Fundamentals of Physical Acoustics*. John Wiley and Sons, New York, NY, 2000.

[17] E. G. Williams. *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[18] P. M. Morse and H. Feshbach. *Methods of theoretical physics. Part I.* McGraw-Hill, New York, NY, 1953.

[19] S. Spors, R. Rabenstein, and J. Ahrens. The theory of wave field synthesis revisited. In *Proc. 124th Convention of the Audio Engineering Society, Amsterdam, May 17-20*, 2008.

[20] N. Gumerov and R. Duraiswami. *Fast Multipole Methods for Helmholtz Equation in three Dimensions*. Elsevier, Oxford, UK, 2004.

[21] I. Gradshteyn and I. Ryzhik. *Tables of Integrals, Series and Products*. Academic Press, 1965.

[22] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1972.

[23] A. J. Berkhout and D. de Vries. Acoustic holography for sound control. In *Audio Engineering Society Convention 86*, 3 1989.

[24] E. M. Hulsebos and D. de Vries. Parameterization and reproduction of concert hall acoustics measured with a circular microphone array. In *Audio Engineering Society Convention 112*, 4 2002.

[25] D. T. Paris and F. K. Hurd. *Basic Electromagnetic Theory*. McGraw-Hill, 1969.

[26] Scenic: Self-configuring environment-aware intelligent acoustic sensing. deliverable 4.3. final report on rendering methodologies and delivery of developed rendering software. Technical report, PoliMi, September 2011.

[27] A. Krokstad, S. Strom, and S. Sorsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, July 1968.

[28] T. Lewers. A combined beam tracing and radiant exchange computer model of room acoustics. *Applied Acoustics*, 38, 1993.

[29] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[30] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '98, pages 21–32, New York, NY, USA, 1998. ACM.

[31] P. S. Heckbert and P. Hanrahan. Beam tracing polygonal objects. *SIGGRAPH Comput. Graph.*, 18(3):119–127, January 1984.

[32] F. Antonacci, M. Foco, A. Sarti, and S. Tubaro. Fast tracing of acoustic beams and paths through visibility lookup. *IEEE Transactions on Audio, Speech & Language Processing*, 16(4):812–824, 2008.

[33] P. Annibale, A. Canclini, F. Antonacci, R. Rabenstein, A. Sarti, and S. Tubaro. An angular frequency domain metric for the evaluation of wave

field rendering techniques. In *Proc. of 19th European Signal Processing Conference (EUSIPCO'11)*, 2011.

[34] H. Wittek, S. Kerber, F. Rumsey, and G. Theile. Spatial perception in wave field synthesis rendered sound fields: Distance of real and virtual nearby sources. In *116th AES Convention*, Berlin, Germany, 2004. Audio Engineering Society (AES).

[35] H. Teutsch. *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Springer, 2007.

[36] A. Kuntz. *Wave Field Analysis Using Virtual Circular Microphone Arrays*. PhD thesis, Verlag Dr. Hut, München, 2009.

[37] M. Kahrs and K. Brandenburg, editors. *Applications of digital signal processing to audio and acoustics*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[38] M. Kolundzija, C. Faller, and M. Vetterli. Designing Practical Filters For Sound Field Reconstruction. In *AES 127th Convention*, 2009.

[39] M. Triki and D. T. M. Slock. Iterated delay and predict equalization for blind speech dereverberation. In *IWAENC 2006, International Workshop on Acoustic Echo and Noise Control, September 12-14, 2006, Paris, France*, Paris, FRANCE, 09 2006.

[40] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models (Springer Series in Information Sciences) (v. 22)*. Springer, 2nd updated ed. edition, April 1999.

[41] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986.

[42] P. V. C. Hough. Method and means for recognizing complex pattern, 1962.

[43] R. O. Duda and P. E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, January 1972.

[44] S. Vega. Tom's diner. From the album Solitude Standing. A&M/Polygram Records, 1987.

[45] EBU Tech. 3253 - Sound Quality Assessment Material: recordings for subjective tests, 09 2008.

[46] B. Klehs and T. Sporer. Wave field synthesis in the real world: Part 1 - in the living room. In *Audio Engineering Society Convention 114*, 3 2003.

[47] ITU-R BS.1284 General method for the subjective assessment of sound quality, 1997.

[48] R. Walpole, R. Myers, S. Myers, and K. Ye. *Probability and Statistics for Engineers and Scientists*. Pearson Education, 2002.

[49] Playrec: Multi-channel matlab audio. `http://www.playrec.co.uk`.

[50] Portaudio: Portable cross-platform audio i/o. `http://www.portaudio.com`.