Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
Doctoral Programme In Information Technology

Space-time Parametric approach to Extended Audio Reality (SP-EAR)

Doctoral Dissertation of:
Mirco Pezzoli

Supervisor:
Prof. Augusto Sarti

Co-supervisor:
Prof. Fabio Antonacci

Tutor:
Prof. Andrea Monti Guarnieri

The Chair of the Doctoral Program:
Prof. Barbara Pernici

Year 2020 – XXXIII Cycle

Abstract

THE term extended reality refers to all possible interactions between real and virtual (computed generated) elements and environments. The extended reality field is rapidly growing, primarily through augmented and virtual reality applications. The former allows users to bring digital elements into the real world, while the latter lets us experience and interact with an entirely virtual environment. While currently extended reality implementations primarily focus on the visual domain, we cannot underestimate the impact of auditory perception in order to provide a fully immersive experience. As a matter of fact, effective handling of the acoustic content is able to enrich the engagement of users. We refer to Extended Audio Reality (EAR) as the subset of extended reality operations related to the audio domain. In this thesis, we propose a parametric approach to EAR conceived in order to provide an effective and intuitive framework for the implementation of EAR applications. It is clear that the main challenges of EAR regard the processing of real sound fields and the rendering of virtual acoustic sources (VSs); hence, EAR requires the development of properly designed sound field representations.

As far as sound field representation is concerned, two main paradigms are present in the literature: parametric and non-parametric. The former describes the acoustic field assuming a signal model governed by few meaningful parameters, e.g., the source signal and location, while the latter relies on the solutions of the wave equation providing accurate results at the cost of higher complexity and lower model interpretability. Therefore, in the context of the EAR, parametric models represent an appealing approach. In fact, they provide a compressed and intuitive description of the sound field. This characteristic promotes the integration of VSs through the parameters of the model and their manipulation thereof.

Here, we introduce a novel parametric model for sound field representation based on few parameters. This model allows both the navigation and manipulation of a recorded sound scene. The main feature of the proposed solution is represented by the modeling of the acoustic source directivity integrated among the parameters of the representation. The directivity is a function describing the spatial property of the source sound radiation. As a matter of fact, sound sources typically present a directional acoustic

emission imposed by their physical characteristics. It follows that the source directivity information influences our acoustic scene perception. Therefore, the integration of the directivity is a fundamental aspect for providing a more natural and immersive EAR, enhancing the user experience. In order to analyze the sound field, we adopted spatially distributed acoustic sensors. This configuration allows us to evaluate the acoustic field from different observation points in order to estimate the parameters required by the proposed representation. Successively, we exploit the estimated parameters to provide a sound field reconstruction technique that enables the *six-degrees-of-freedom* interaction (virtual navigation) with the sound field.

Conveniently, the parameters adopted for describing the acoustic sources can be exploited for characterizing a VS. Therefore, we can seamlessly implement EAR within the same parametric representation. Here, the addition of the source directivity into the model is appealing since it allows the accurate rendering of VSs, including their directional characteristics. Hence, we can further lead the real-virtual interaction by implementing VS replicas of actual acoustic sources. A VS replica mimics the source spatial sound radiation through the VS directivity parameters. For instance, the VS parameters can be estimated from measurements on the real source. Conversely, we can rely on fully simulated acoustic sources, e.g., employing Finite Element Method (FEM) simulations, from which the VS parameters are derived. It follows that an accurate estimate, prediction, and analysis of the directivity of VSs are fundamental to obtain an effective EAR.

In this thesis, we studied the VS implementation through a case study. In particular, we focused on the VS implementation of violins. Whereas violins present a peculiar directional radiation characteristic, we need to carefully analyze and model their directivity in order to provide an accurate VS implementation. Regarding the analysis of the violin directivity, we can outline different solutions according to their invasiveness. In the first place, one can perform measurements directly on played violin. During our collaboration with *Museo del Violino* settled in Cremona (Italy), we had the unique opportunity to measure, for the first time, a relevant number of valuable historical violins made by the renowned masters of the Cremonese school such as *Antonio Stradivari* and played by professional violinists. From the acquired data, we derived a compressed representation of the violin directivity pattern based on the spherical harmonics expansion. Besides the VS modeling, the adopted representation allowed us to study and characterize the directivity patterns of the instruments, giving insights of their directional behavior. Although the measurement of played instruments allows an analysis scenario closer to the actual listening conditions, it might not be applicable for particularly fragile instruments.

Less invasive techniques, such as nearfield acoustic holography (NAH), can be employed when conventional measurements cannot be carried out. It is known that the acoustic radiation of vibrating objects, such as violins, is determined by their dynamical behavior. Hence, from the knowledge of the vibration velocity field, we can estimate the directivity of the source. NAH allows the contactless estimation of the velocity field of a vibrating source from acoustic pressure measured in its proximity. Here, we introduced a novel NAH technique based on deep learning. In particular, we proposed a convolutional neural network (CNN) with an autoencoder-inspired structure in order to estimate the velocity field of both rectangular and violin plates.

Alternatively, simulations allow us to predict the directivity of a source relying on the FEM simulation of its vibroacoustic behavior. This approach minimizes the invasiveness at the cost of reduced accuracy caused by inherent approximations of the simulated model. It follows that an effective violin simulation requires a 3D model of the instrument geometry and the mechanical parameters of the material. Unfortunately, we can typically only acquire the outer surface of existing instruments. Therefore, we developed a practical technique for reconstructing the 3D model of violin plates, starting from outer surface scans and sparse thickness measurements taken at reference points. Furthermore, as regards the estimation of the material mechanical parameters, we proposed the evaluation of the Young's modulus from the sound wave velocity of wood. As a matter of fact, the Young's modulus is a fundamental parameter for mechanical simulations. The developed technique estimates the sound wave velocity from responses of the wood to an impulsive excitation in a rake receiver fashion. Successively, from the knowledge of the sound wave velocity, the Young's modulus is indirectly derived.

Lastly, we propose an EAR proof of concept through which we showcase the benefit of the proposed parametric approach to EAR. We display an EAR scenario in which two VSs, a VS replica of a prestigious violin, and a simulated generic model of the instrument are virtually co-located in a real sound scene with the presence of actual sound sources. The results give a sneak peek of the power of EAR, showing that the proposed parametric approach is able to provide the blend between real and virtual sound elements. Hence, we envision that the proposed solutions will pave the way to the development of parametric EAR frameworks for extended reality applications.

Contents

CHAPTER 1

Introduction

Extended reality is the field in which all the applications concerning the interaction between real and computer-generated virtual elements and environments are contained. Therefore, both augmented and virtual reality fall within the scope of extended reality. On the one hand, augmented reality allows the addition of digital elements into the real world, while on the other hand, virtual reality lets a user interact with an entirely virtual environment.

Despite extended reality applications primarily involving the visual domain, auditory perception plays a fundamental role in order to achieve an immersive experience. As a matter of fact, through our sound perception, we make sense of surroundings, localizing sound sources and placing ourselves in the environment. Therefore, the correct handling of the audio contents coherently with the visual information is a key aspect in extended reality.

In the last decade, we experienced a rapidly growing development of extended reality applications, made possible by the introduction of head-mounted displays and the increase of computational power in smartphones and portable devices. At the same time, the interest in multichannel audio systems increased a great deal. A huge number of technological devices are now equipped with spatial distributions of microphones or loudspeakers, usually organized in arrays. This allows the acquisition or reproduction of sound, exploiting the additional information introduced by the spatial dimension. As a matter of fact, a wide range of tasks such as acoustic source localization, speech enhancement, and dereverberation, sound source separation, and others benefit from the space-time analysis of sound fields. Extended reality is no exception, and in the next future, we expect a growth of advanced extended reality applications made possible by the availability of low-cost sensors, e.g., MEMS microphones. The deployment of a large number of sensors, seamlessly integrated with the environment, enables the analysis of a sound scene from multiple points of view.

This possibility opens to the implementation of algorithms that analyze the entire acoustic scene, paving the way to powerful handling of the audio contents in extended reality applications.

1.1 Extended Audio Reality

With Extended Audio Reality (EAR), we refer to the audio branch of extended reality. More precisely, EAR concerns all the sound-related operations in extended reality applications. Hence, EAR aims at merging the reality and virtual elements with the focus on audio signals only.

One can easily understand EAR, imaging, for example, the virtual access to live concerts or events. In extended reality, the user is able to navigate with the environment, i.e., exploring the venue enjoying the event from different points of view. In fact, the goal is to provide an experience as close as possible to "*as you were there*". Therefore, EAR must be able to let a listener explore the recorded scene, and ideally, provide an accurate "*sound experience*" independently from the user position. Nonetheless, this is only a part of the picture. We can also think to add virtual contents to the real sound scene, for instance, remotely played instruments. It follows that the additional virtual data must be coherently rendered in the EAR framework allowing the user to perceive the virtual content as it was present in the scene. It is clear that EAR brings a series of interesting challenges for the space-time acoustic signal processing research community.

On the one hand, EAR requires the processing of actual acoustic fields, enabling user interaction with the sound scene. Here, the sound field navigation problem is fundamental since it allows a user to interact with a recorded scene with *six-degrees-of-freedom* (6DOFs) to change its listening position arbitrarily. Additionally, one might actively manipulate a sound scene, for instance, excluding one acoustic source or reducing the reverberant component of the sound field.

On the other hand, EAR provides the interaction between virtual acoustic elements, such as a virtual acoustic source (VS) and real sound scenes. Therefore, we need to accurately model the VSs in order to correctly include them in the scene, i.e., providing the same perception as they were actually there.

In order to provide an effective EAR experience, the sound field at the target (listening) position must be accurately rendered. We can tackle this problem from a sound field reconstruction perspective. With the aim to let the user freely move in the scene, we have to estimate the acoustic field at the listening location exploiting the information acquired by the microphones at different positions.

As regards the VSs, a key aspect is represented by the effective modeling of the spatial sound radiation of the VS. In fact, this allows a better perception of the source position and orientation, hence enabling more natural navigation in the environment. It is clear that EAR requires the development of a suitable sound field representation in order to solve its main challenges.

In this thesis, we propose an EAR approach based on a parametric sound field representation. This allows us to perform the sound field reconstruction and interaction and the addition of VSs in a compact fashion.

In general, EAR and extended reality might potentially have a significant impact on society and our everyday life. It is sufficient to think of potential applications concerning social media, gaming, and, generally speaking, the entertainment business. Limiting our analysis to music-related applications, we can expect a growth of new media contents and events that can be enjoyed through extended reality, e.g., live concerts and shows. Therefore, we foresee the development of a market for accessing entertaining contents or purchasing virtual objects (e.g., virtual musical instruments) to be used in extended reality applications.

1.2 Goals and Methodology

The goal of this thesis is to propose an approach to EAR that paves the way to the implementation of EAR frameworks. We aim to present a compact and intuitive paradigm to EAR that is able to provide both sound field reconstruction and seamless integration of VSs.

Our solution relies on a novel parametric sound field representation. This description is characterized by the inclusion of the sound source directivity into the sound field model. Therefore, we make explicit expression of the source directional sound emission in order to improve the sound field reconstruction on the one side and provide a directional VS model on the other side. Moreover, we discuss the implementation of VSs through the analysis of a case study. We analyze the VS implementation of violins; in particular, we focus on different strategies for the estimation of VS parameters.

In the following, we introduce the main challenges of the proposed EAR approach regarding the sound field processing and the VS modeling.

1.2.1 Sound Field Processing

In the literature, two main approaches to sound field reconstruction emerged: non-parametric and parametric. Both paradigms are based on their inherent sound field representation.

As regards non-parametric methods, they rely on the sound field decomposition into basis functions. Hence, non-parametric sound field reconstruction techniques are directly derived from the solution of the wave equation [1, 3, 7, 29, 36, 68, 80, 120, 167, 167, 171, 190, 195, 205, 215, 228, 229, 262]. Typically, plane waves [80, 120, 190, 215] or spherical harmonics [1, 88, 228, 229] are adopted as basis functions, according to the geometry of the employed configuration of acoustic sensors. Therefore, the coefficients of the basis functions, estimated at the locations of the sensors, are exploited in order to provide the description of the sound field.

From the knowledge of the expansion coefficients, the acoustic field is thus potentially reconstructed at arbitrary positions. In practical scenarios, only a finite-order expansion of the sound field can be adopted, due to the physical limitation of the spatial sampling. It follows that only an approximation of the sound field can be provided with an accuracy that depends on the number and spatial arrangement of the sensors. Recent works cope with these limitations employing multiple distributed microphone arrays able to capture first or higher spherical harmonics expansion [88, 228, 229, 262] or single sensors distributed in the scene [23, 24, 131, 142, 143, 256]. Interesting methods [88, 228, 229, 262] rely on the so-called spherical harmonic translation theorem;

this technique relates the local (low-order) spherical harmonics coefficients with the global (high-order) spherical harmonic coefficients obtained in a convenient center of expansion. Other approaches [23, 24, 131, 142, 143, 256] rely on sparsity assumptions of the sound field and compressed sensing processing in order to relax the sampling requirements. Although compressed sensing allows us to reduce the spatial sampling, a main limitation for adopting non-parametric techniques in EAR remains; this concerns the interpretability of the signal representation. In general, the sound field is described by a set of coefficients estimated from the acquired data, but the interpretation of the coefficient values is not often straightforward (e.g., differentiate between the values of two sources).

An alternative approach to the sound field reconstruction is given by parametric or model-based methods.

Generally, parametric techniques rely on a compact model of the sound field, defined by the parameters of the model itself. Although parametric methods are not based on an exact solution of the wave equation, they aim at accurately recreating the perceptual spatial sound features at the target locations. A great range of parametric methods are available in the literature [33,70,72,89,90,111,124,141,146,206,211,212,260], among which the directional audio coding (DirAC) [211] and the high angular resolution plane wave expansion (HARPEX) [33] have been widely adopted.

In [211] and [33], the sound field is represented as the sum of two distinct components: the direct and the diffuse sounds. The direct sound represents the signal associated to the direct path between a source and the recording location, while the diffuse component models the reverberation and noise. The model is completed by the additional information of the source location or the direction of arrival.

Through their original definition, [211] and [33] provide a sound field reconstruction limited to the recording location, i.e., they enable a *three-degrees-of-freedom* interaction (rotation) of the user. In order to let the user move in the scene (6DOFs interaction), improved solutions have been proposed [72, 141, 206, 260].

For instance, in [206], the authors exploit the a-priori information on the source location in order to allow the translation of a first-order microphone signal in space. Differently, in [72, 141, 260] spatially distributed sensor arrays are employed for analyzing the sound scene. The source location and the direct and diffuse components of the sound field are estimated and exploited for reconstructing the signal of a virtual microphone (VM), namely, a sensor virtually placed in the scene.

In this thesis, we introduce an improved parametric sound field reconstruction technique inspired by the solution in [260]. In particular, we generalize and extend the sound field model typically adopted in the literature, which assumes an omnidirectional acoustic radiation of the sources, i.e., the emitted acoustic energy is isotropic. As a matter of fact, actual sound sources are usually characterized by a directional acoustic emission inherently caused by their physical properties.

The spatial characteristic of the source sound radiation is generally described by the directivity function. Therefore, in order to accurately reconstruct the sound field we have to include the directivity of the sources into the underlying model. As a result, our sound scene perception is also influenced by the source directivity. For instance, we can think of the perceptual difference between being behind or in front of a speaker. It follows that, in the context of sound field reconstruction for EAR, the directivity information is a key aspect for providing a more natural experience.

We take advantage of distributed microphone arrays for analyzing the sound field estimating the parameters of the model. In particular, during the analysis phase we estimate the location of the sources and we derive the direct and the diffuse sound field components of the microphone signals. From the direct signal, we then retrieve the parameters describing the directional sources as the solution of a sparsity-based optimization. Modeling the direct sound field as the exterior field of the sources we include their directional properties, overcoming the omnidirectional source model typically assumed in the literature.

Finally, we exploit the estimated parameters in order to synthesize the signal of a VM arbitrary located in the scene.

In the context of EAR parametric models represent an appealing approach. In fact, they provide a description of the sound field by means of few meaningful features, i.e., the model parameters. This might promote an intuitive manipulation of the sound field and ease the integration of VSs. For instance, during the VM synthesis phase, we can remove acoustic sources from the scene, simply by selecting the parameters of the target sources only, or we can remove the reverberant component of the sound field discarding the diffuse component from the model.

With the goal of enabling the extraction of sound sources from an existing sound field, we explored solutions based on non-parametric sound field representations. In particular, we focused on the sound source separation problem.

The adoption of a non-parametric sound field representation allowed us to reduce the requirements in terms of sensor setup and a priori information. Therefore, we developed a blind source separation (BSS) algorithm that can be applied on a single extended linear microphone array. The aim of BSS is to extract multiple unknown audio signals (sources) by processing a set of mixture thereof. The separation process is defined as “blind" since no other information, except from the multichannel signal, is available. We tackled BSS by bringing a new non-parametric sound field representation into the state-of-the-art BSS technique known as Multichannel Nonnegative Matrix Factorization (MNFM) [189, 232]. Nonnegative matrix factorization (NMF) consists of a data decomposition technique that factorizes a nonnegative matrix into a sum of rank-1 components. In the context of BSS this factorization is applied to multichannel signals in order to decompose each source contribution at the microphones. Different MNMF techniques [57, 127, 135, 150, 172, 180, 242] take advantage of the spatial information contained in the signals in order to improve the separation performance. However, the spatial information these methods is typically limited to the Direction of Arrival (DoA). In this thesis, we propose to adopt the ray space (RS) [37, 160] sound field representation in order to better exploit the spatial information in the MNMF framework. The RS maps the domain of the plenacoustic function [7] parametrizing the directional components of the sound field according to their direction and position of analysis. Recently, the RS emerged as an effective tool for representing both far and near acoustic fields since, thanks to its parametrization, the location of the acoustic sources in space is inherently represented in the data [37, 159]. Therefore, we exploit this characteristic adopting the RS as the domain of the MNMF, overcoming the limited representation of the position given by the DoA. As regards EAR, we can employ BSS for the static manipulation of the scene, namely, separating one or more sources. Additionally, the proposed BSS solution can be used for extracting a source signal from a given sound field that can be later rendered as a VS in a different sound scene as described in the next section.

1.2.2 Virtual Source Modeling

As underlined in the previous section, acoustic sources usually present a directional sound radiation. It follows that in order to accurately implement VSs in a given sound scene, the directivity of the VSs cannot be discarded.

Conveniently, the parametric model introduced for describing directional sources can be exploited for characterizing a VS. In particular, we model the directivity pattern of a source or VS by means of its spherical harmonics expansion. This provides a compact description of the directional behavior of a source through the coefficients of the expansion. Therefore, we can seamlessly achieve the EAR mixture of actual sound scenes and VSs within the same parametric model. As a consequence, a further level of interaction between real and virtual elements can be achieved by implementing VS replicas of actual sound sources. In practice, a VS replica aims at reproducing the source acoustic emission through the VS directivity model.

It is clear that the accuracy of the VS model depends on the available data. For instance, when the source to be modeled as VS is extracted from another sound scene, for instance by means of BSS, usually no information about its directivity is given. In this scenario, the VS can be simply rendered with an omnidirectional radiation characteristics.

When possible, instead, we can rely on a set of measurements on the real source for estimating the directional parameters of the VS. This allows the implementation of VS replicas of real sound sources.

Another option is represented by simulations. One can derive the VS parameters from fully simulated sources, e.g., adopting finite element method (FEM) simulations. Here, the effectiveness of the derived VS is governed by the accuracy of the simulated source model.

In this thesis we studied the problem of the implementation of VSs by means of a case study. Our goal is the analysis of the VS implementation of violins. In fact, the violin is an interesting musical instrument that present a peculiar acoustic behavior. Hence, we need to carefully analyze and model the directivity of violins in order to provide their relative VS implementations. We outlined different solutions for the analysis of the violin directivity, which can be grouped according to their level of invasiveness.

First of all, we can rely on measurement directly taken on played instruments. Thanks to our collaboration with *Museo del Violino* setteld in Cremona (Italy), we had the possibility of measuring the directivity pattern of a relevant number of prestigious historical instruments played by professional violinists. The violins under analysis were made by the great masters from Cremona such as *Antonio Stradivari* and *Giuseppe Guarneri "del Gesù"*.

The acquired data allowed the analysis and the characterization of the directivity pattern of the instruments, besides the VS modeling. This gave us insights on the directional behavior of such valuable instruments. The set of violins under analysis represents an unicum of great scientifc and cultural value. Moreover, to the best of our knowledge this is first time that the directivity of violins is analyzed in a systematic fashion, introducing new tools for the comparison of the directivities. As a matter of fact, in the literature of violin acoustics only qualitative evaluations of a limited number of instruments are available.

In this scenario, the analysis is performed in a situation close to the actual listening conditions. Unfortunately, this might not be applicable for particularly fragile instruments, due to the invasive interaction with the player.

In the case in which customary measurements cannot be employed, we can rely on less invasive techniques such as nearfield acoustic holography (NAH) [165, 286].

NAH is an interesting method for the analysis of acoustic sources. It allows the contactless estimation of the velocity field of a vibrating source from acoustic pressure measured in its proximity. As a matter of fact, the acoustic radiation of a vibrating sound source, such as the violin, is determined by its mechanical behavior. Therefore, the directivity of a sound source can be inferred from the knowledge of its vibration velocity field.

In practice, NAH primarily comprises the inversion of the Kirchhoff-Helmholtz integral formulation of the radiated sound field. In the literature this problem is known to be ill-posed, and different solutions based on its regularization have been presented [56, 58, 94, 237, 270, 287, 295].

Interesting advanced approaches are introduced in [56,94], where the authors exploit compressed sensing principles and the Equivalent Source Method (ESM) [139, 149] in order to solve the NAH inversion. The solution is then given by means of a sparse set of fictitious dimensionless sound sources that equivalently represent the acoustic pressure field. Nonetheless, the main limitation of ESM techniques resides in the computation of the optimal set (in terms of number and location) of equivalent sources. Although the authors in [56] alleviate this limitation by restricting the ESM solution space to a suitable learned dictionary, the technique is derived solely for square plates with fixed geometry.

Here, we proposed a novel data-driven NAH technique. In particular, we adopted a deep learning approach to NAH, where the regularization is performed by exploiting the knowledge given by the data during the training. Inspired by ESM solutions, where the acoustic field is represented by a set of feature components, namely, the equivalent sources, we employ a convolutional neural network (CNN) in order to learn a potentially more powerful feature representation for performing NAH directly from the data itself.

Among the different CNN architectures, we focused on a structure similar to an autoencoder. This network allows us to learn useful properties of the data and they are adopted for dimensionality reduction and features learning other than denoising. The basic structure of these networks is composed by a compressing phase, known as encoder, in which a compressed encoded representation of the input is learned, followed by a decoder where the encoded data is expanded to obtain the desired output. The CNN is trained using datasets of synthetic data generated using FEM simulations. We show that the proposed approach is able to work with different geometries and material properties by applying the CNN to both rectangular and violin plates.

In order to minimize the invasiveness of the analysis on the directivity of the instrument, we can also rely solely on simulations. Through the FEM simulation of the vibroacoustic behavior of a sound source, we can predict its directivity.

Clearly, the accuracy of such analysis is reduced, with respect to measurements carried out on actual instruments, due to the approximations of simulated models. Nevertheless, simulations allow more flexibility in the design of the object and in some contexts this can be useful for the analysis of prototypes.

In order to obtain an effective simulation of a violin, we need the accurate 3D model of the instrument geometry and the mechanical parameters of the materials. In the case of existing violins, we can acquire its geometry be means of 3D scans. Usually, we cannot access to the single body parts, hence we have to rely on the outer surface geometry only. Therefore, we developed a practical technique for the reconstruction of the 3D model of a violin plate.

The proposed method requires the outer surface scan of the plate and sparse thickness measurements. In practice, we reconstructed the profile of the inner surface of a violin plate accordingly to the thickness at the reference points. Then, we combined the outer and inner surfaces in order to reconstruct the whole 3D geometry of the violin plate.

For what concerns, the analysis of the mechanical parameters, we introduced a technique for the estimation of the Young's modulus in wood. The Young's modulus represents a fundamental parameter for the correct simulation of the dynamic behavior of materials. A direct estimation of the Young's modulus can he obtain from the material by means of destructive tests [222]. More interestingly, we can indirectly derive the Young's modulus from the sound wave velocity with repeatable non-destructive techniques. Luthiers typically estimates the sound wave velocity through the *tap tone* technique [128] or by measuring the time of flight (TOF) of an impulsive wave in the wood. On the one side the *tap tone* is highly affected by the manual skill of the luthier that is required to estimated the resonance frequency of the wood. On the other side, TOF estimation is performed with expensive instrumentation and it is sensitive to the measurement noise. Here, the proposed technique relies on the impulse responses of the wood measured by means of accelerometers. Hence, no particular skill is required for performing the measurements. The acquired signals are processed in a rake receiver fashion in order to identify the sound wave velocity associated to the measurements. This allows us to work in the audio bandwidth adopting a less expensive hardware with respect to TOF devices.

1.2.3 EAR Proof of Concept

A proof-of-concept simulation is proposed in order to demonstrate the parametric approach to EAR.

We discussed the main challenges of the implementation of a framework for EAR, in particular as regards the VS synthesis. As a matter of fact, the rendering of the VS is conditioned by the actual EAR application scenario. In particular, while the VS modeling allows us to provide an accurate description of the VS spatial sound radiation, the effective rendering in the EAR scene is governed by the available environment information.

The direct sound generated by the VS is obtained directly from the VS modeling. Here, we take advantage of the directivity information in order to accurately render this component of the sound field. Conversely, as regards the synthesis of the diffuse component generated by the VS, different strategies arise according to the available

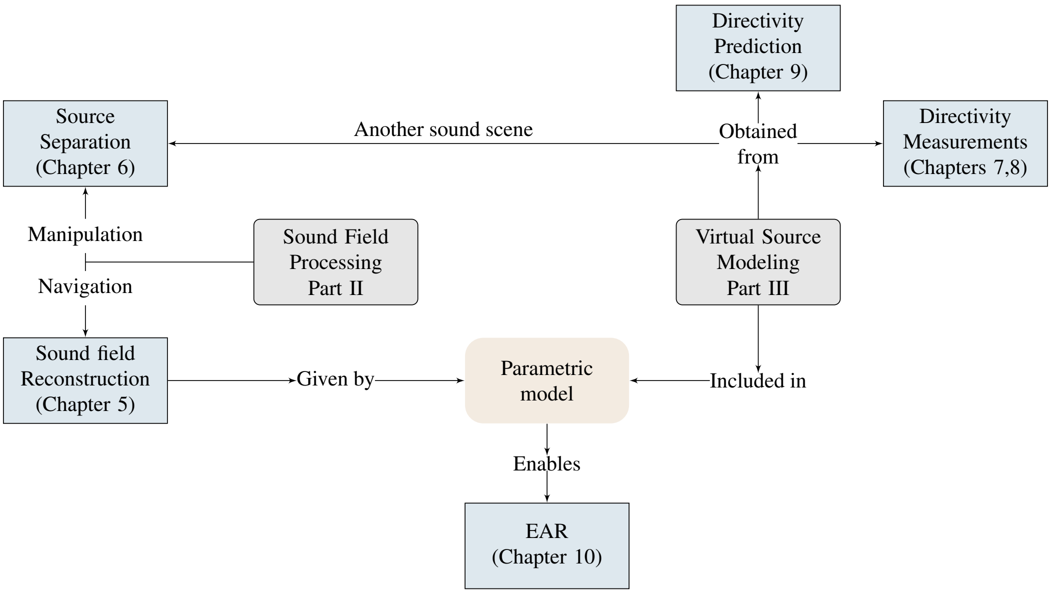


Figure 1.1: *A block diagram of the overall space-time parametric approach to EAR developed in this thesis. Gray-shaded blocks represent parts of the thesis, while blue-shaded elements refers to chapters.*

information about the environment.

The scenario of the EAR proof of concept comprises a string trio setup. In particular, we included two violins as VSs in a sound scene with a cello, as expected in a string trio. Hence, we analyze the sound field generated by the actual acoustic source, the cello, through the proposed parametric sound field reconstruction technique computing the signal of a VM in order to navigate the scene. Successively, the violin VSs are added to the VM signals assuming, for simplicity that the environment is known. We present two violin VS models. The first VS implements a virtual replica of the prestigious *Il Cremonese* violin by *Antonio Stradivari* derived from measurements on the instrument. The second VS is computed from a simulated generic violin model. Here, we estimated the VS parameters from the FEM simulation of a simplified violin body model. The goal of this proof-of-concept simulation is to display the power of the parameteric EAR approach, by including different VS models in an interesting scenario. The results showed that the parametric EAR approach allows the interaction between virtual elements and real sound scenes. Therefore, we envision that the proposed parameteric model could provide an effective approach to the implementation of EAR frameworks.

1.3 Thesis Outline and Contributions

In this section, we present an overview of the thesis.

In Figure 1.1 a summary of the overall developed approach to EAR is provided. The thesis outline directly reflects the main EAR tasks introduced in the previous sections. Besides a first preliminary part covering the basic definitions of signals and acoustics, the thesis is divided in two main independent contributions (see Figure 1.1):

- Part II concerning the sound field processing, namely the techniques proposed for the navigation and manipulation of real recorded sound fields;

- Part III devoted to the modeling of virtual sources where techniques for the measurement or prediction of directivity of violins are considered.

Although considered as two independent parts the combination of the sound field processing and virtual source modeling allows us to implement a EAR proof-of-concept simulation in Chapter 10.

We summarize the content for each chapter, underlying its original contribution and the relation with respect to the overall thesis. In Part I we provide the preliminaries required for the comprehension of the rest of the thesis.

Chapter 2 introduces the definition of fundamental background elements and the notation adopted throughout the thesis. Both continuous and discrete signals and their related operations such as the Fourier transforms are defined.

Chapter 3 provides the relevant acoustic background. The concepts and definitions introduced in this chapter are the preliminaries of the following chapters. First, the notation adopted for both 2D and 3D spatial coordinate systems is given. Successively, the solutions of the wave equation are presented according to the Cartesian and spherical coordinate systems. The introduced sound field representations are at the basis of the concepts described in Chapter 6 and Chapter 4.

Part II address the sound field processing in the EAR context. In particular, the sound field reconstruction problem is examined with the introduction of a novel parametric technique. Moreover, a blind sound source separation technique based on a non-parametric sound field representation is presented.

Chapter 4 provides an overview of the sound field reconstruction problem. This problem is directly related to the navigation of a recorded sound field in EAR. According to the paradigms of the sound field representation, two main categories of sound field reconstruction methods can be defined: non-parametric and parametric. We provide a review of state-of-the-art techniques in both categories.

In Chapter 5, we propose parametric sound field reconstruction techniques [199] that represent the basis of the EAR approach. The methods we developed in [197–199], include the directivity pattern of the sound source into the parametric model. [197,198] consider controlled environments, while the technique presented in [199] extends the parametric model to reverberant sound fields.

Chapter 6 addresses the sound field manipulation through the development of a novel blind source separation approach. Here, we exploit a non-parametric sound field representation, known as ray space, in order to perform BSS. First, we describe the process of mapping the signal of a uniform linear array to the ray space domain, and we present a new computationally efficient implementation [44] of [37]. The BSS in [201] is performed applying the MNMF algorithm to the ray space data in order to estimate the ray space representation of each source present in the environment. This allows us to extract the content of an acoustic source when no a-priori information on the sound scene is given.

Part III address the implementation of virtual sources in the EAR context. We tackle the virtual source modeling through the analysis of a case study concerning the violin as VS.

In Chapter 7 we describe the estimation of the directivity pattern of violins and its characterization given in [200]. First, we review the acoustics of violins with special focus on the sound radiation characteristics. The literature on violin acoustics underlines the relevant directional radiation of this kind of instruments. Therefore, it is fundamental to model their directional behavior in a VS. We performed mesurements and analysis on directivity patterns of historical violins. This provided insights on the directional radiation characteristics of violins and thus the importance of their accurate modeling.

In Chapter 8, before the introduction of the novel data-driven NAH technique of [185], a review of the CNN fundamentals is given. This type of architectures have been employed in [185] for estimating the vibrating field of rectangular and violin plates from the acoustic pressure measured in their proximity.

Chapter 9, introduced practical techniques that can be employed in order to obtain improved models for the FEM simulation of violins. The method developed in [203] allows the reconstruction of a violin plate 3D geometry from its outer surface scan. In [273] we introduced a practical technique for the estimation of the sound wave velocity in wood.

Lastly, Chapter 10 discuss the implementation of VSs in a proof-of-concept simulation. Here, we introduce the VS model that allows the implementation of directional VSs. We describe how a-priori information about the acoustic environment drives the rendering of VSs. Later, we present a simulation that shows the EAR approach. In particular, a string trio scenario is considered, where a cello is simulated as a “real” source in the scene, while two violins are added as VSs. The VS models are derived both from measurements on a real instruments and simulations. The first violin represents a virtual replica of an historical instrument, obtained from the directivity analysis described in Chapter 7. Conversely, the second VS source is derived from the FEM simulation of a simplified violin body model.

Chapter 11 draws final considerations and outlines the directions of possible future research and development activities.

1.4 List of the included publications

In this thesis we present a work that includes original contents from the following peer-reviewed articles. The papers have been published or planned to be submitted by the author to international journals and conferences.

- [197] M. Pezzoli, F. Borra , F. Antonacci, A. Sarti, and S. Tubaro. *Estimation of the soundfield at arbitrary positions in distributed microphone networks based on distributed ray space transform.* In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 186–190. IEEE, 2018.
  Abstract
  In this paper we propose a parametric sound field reconstruction approach. In particular, the technique is based on the estimation of three parameters for each acoustic source (source position, radiation pattern and source signal) given the signals acquired by few arbitrarily placed microphone arrays. This allows us to synthesize the signal of a virtual microphone placed in any point of the acoustic scene.

- [198] M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, and S. Tubaro. *Reconstruction of the virtual microphone signal based on the distributed ray space trans-*

form. In 2018 26th European Signal Processing Conference (EUSIPCO), pages 1537–1541. IEEE, 2018.
Abstract
In this paper we propose a technique for the reconstruction of the sound field at arbitrary positions based on a parametric sound field description. The methodology consists in the estimation of the sources model parameters (source position, radiation pattern and source signal), starting from the signals acquired by arbitrarily distributed microphone arrays. Given the model parameters it is possible to synthesize the signal of a virtual microphone at an arbitrary position and with an arbitrary pick-up pattern.

- [199] M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, and A. Sarti. *A parametric approach to virtual miking for sources of arbitrary directivity*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol 28, pages 2333–2348, 2020.
Abstract
In this article we propose a methodology for the reconstruction of sound fields in arbitrary locations based on the signals acquired by a spatial distribution of compact microphone arrays (virtual miking). The proposed method is suitable for operating in reverberant environments, thanks to a two-stage analysis process, the former of which aims at separating the direct and the diffuse components of the sound field. The method that we propose is inherently parametric, as the sources of the acoustic scene are characterized by parameters describing location and directivity (spherical harmonics expansion), which are extracted from the exterior model of the direct component of the sound field. Once the parameters of the sources are extracted, the direct sound field at an arbitrary location is reconstructed. The diffuse component is reconstructed from the joint knowledge of the diffuse component at the locations of the distributed microphone arrays, under the assumption of isotropic behavior. Results show that the proposed technique is able to analyze the sound field and reconstruct the parameters of the sources that are active in the scene. In addition, the synthesis of the signals at the virtual microphone locations turns out to accurately match (in terms of spatial cues) the actual sound field, as measured by a microphone places in the desired location.

- [44] F. Borra, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, S. Tubaro, and A. Sarti. *A fast ray space transform for wave field processing using acoustic arrays*. In 2020 28th European Signal Processing Conference (EUSIPCO), pages 186–190. IEEE, 2020.
Abstract
The importance of soundfield imaging techniques is expected to further increase in the next few years thanks to the ever-increasing availability of low-cost sensors such as MEMS microphones. When it comes to processing a relevant number of sensor signals, however, the computational load of space-time processing algorithms easily grows to unmanageable levels. The Ray Space Transform (RST) was recently introduced as a promising tool for soundfield analysis. Given the collection of signals captured by a uniform linear array of microphones, the RST allows us to collect and map the directional components of the acoustic field onto

a domain called “ray space”, where relevant acoustic objects are represented as linear patterns for advanced acoustic analysis and synthesis applications. So far the computational complexity of the RST linearly increases with the number of microphones. In order to alleviate this problem, in this paper we propose an alternative efficient implementation of the RST based on the Non Uniform Fast Fourier Transform.

- [201] M. Pezzoli, J. J. Carabias-Orti, M. Cobos, F. Antonacci,and A. Sarti. *Ray-space-based multichannel nonnegative matrix factorization for audio source separation*. IEEE Signal Processing Letters, vol. 28, pp. 369-373, 2021.
  Abstract
  Nonnegative matrix factorization (NMF) has been traditionally considered a promising approach for audio source separation. While standard NMF is only suited for single-channel mixtures, extensions to consider multi-channel data have been also proposed. Among the most popular alternatives, multichannel NMF (MNMF) and further derivations based on constrained spatial covariance models have been successfully employed to separate multi-microphone convolutive mixtures. This letter proposes a MNMF extension by considering a mixture model with Ray-Space-transformed signals, where magnitude data successfully encodes source locations as frequency-independent linear patterns. We show that the MNMF algorithm can be seamlessly adapted to consider Ray-Space-transformed data, providing competitive results with recent state-of-the-art MNMF algorithms in a number of configurations using real recordings.

- [200] M. Pezzoli, A. Canclini, F. Antonacci, and A. Sarti. *A Comparative Analysis of the Directional Sound Radiation of Historical Violins*. The Journal of the Acoustical Society of America, (submitted).
  Abstract
  The directivity pattern of a violin describes the sound energy radiation of the instrument as a function of frequency and direction of emission. Violins exhibit a rather complex directivity pattern, which is known to exhibit rapid variations across the frequency, and whose behavior cannot be easily predicted except in the lowest frequency range. The acoustic behavior of the violin is a fascinating research topic that has prompted numerous published works, but a thorough, comprehensive and comparative analysis of violin directivity patterns, is long overdue. In this article, we propose a set of tools for characterizing the radiative behavior of musical instruments and, in particular, for comparing their directivity patterns. We apply such tools for a comparative analysis of the directivity patterns of some of the most prestigious historical violins ever made, coming grand masters such as Antonio Stradivari, Giuseppe Guarneri “del Gesú" and members of the Amati family; which are all preserved in the Violin Museum of Cremona, Italy, where our lab is located. The tools introduced in this work allowed us to gain insight on the acoustic behavior of such extraordinary instruments, quantitatively confirm some widespread beliefs, and draw some surprising conclusions.

- [185] M. Olivieri, M. Pezzoli, R. Malvermi, F. Antonacci, and A. Sarti. *Near-field acoustic holography analysis with convolutional neural networks*. In INTER-NOISE and NOISE-CON Congress and Conference Proceedings, volume 261,

pages 5607–5618, Institute of Noise Control Engineering, 2020.
Abstract
Near-field Acoustic Holography (NAH) enables the contactless analysis of the vibrational field on plates and shells from the acoustic data captured in proximity of the vibrating object. In this paper, we propose a data-driven approach to NAH by using a Convolutional Neural Network (CNN) that predicts the vibrational field on the object from the acoustic pressure field captured by a microphone array deployed in its proximity. We have conducted an extensive simulation campaign on rectangular plates of different dimensions, boundary conditions and mechanical properties. This dataset has been generated using Finite Element Method simulation for predicting both vibrational and acoustic pressure fields. The performance of the proposed data-driven NAH method is assessed by comparing the estimated vibrational field with the ground truth. Moreover, we offer an analysis of the robustness of the estimate against noisy input data.

- [203] M.Pezzoli, R. R. DeLucia, F. Antonacci, and A. Sarti. *Predictive simulation of mechanical behavior from 3D laser scans of violin plates.* In 23rd International Congress on Acoustics, pages: 7577-7583, ICA, 2019.
  Abstract
  In this paper we present a methodology for the predictive simulation of the vibrometric behaviour of a violin plate. The 3D outer shape of the plate is acquired by means of a 3D laser scanner and then smoothed in order to remove artefacts and details that are unnecessary for the acoustics simulation. The thickness of the plate is incorporated into the model through a technique that receives as input the thickness sampled at some points of the plate and interpolates it over the entire surface. We validate this 3D reconstruction technique by comparing the vibrometric behaviour of the 3D model with data measured on the reference plate, and with simulations on a model with uniform thickness.

- [273] L. Villa, M. Pezzoli, F. Antonacci, and A. Sarti. *A methodology for the estimation of propagation speed of longitudinal waves in tone wood.* In 2020 28th European Signal Processing Conference (EUSIPCO), pages 66–70. IEEE, 2020.
  Abstract
  In this paper we propose a methodology for the estimation of the longitudinal wave velocity in tone wood. Differently from techniques adopted in the field of luthiery, the proposed estimation method does not require neither specific user skill nor expensive instrumentation. The introduced method exploits the impulse response of the wood block, acquired by means of accelerometers. The measured signals are processed in order to compute an estimate of the longitudinal wave velocity of the tone wood in a rake receiver fashion. We tested the technique both on synthetic data and measurements of actual tone wood blocks, showing the effectiveness of the proposed solution with respect to state-of-the-art methods.

Part I

Preliminaries

CHAPTER 2

Signals and Transformations

This chapter offers a review of signals and their representation, the definition of basic operations on the signals and the Fourier transformations. Signals are defined as functions with a proper domain that can be continuous or discrete, one-dimensional or multi-dimensional and in the scope of this thesis they represent acoustic phenomena.

Both one-dimensional and multi-dimensional continuous signals are introduced in Section 2.1 along with the definition of the inner product, the Euclidean norm and the convolution operation.

Additionally, one-dimensional continuous periodic signals and the circular convolution are defined in Section 2.1.2.

While reviewing basic definitions of signals and Fourier transformation, this chapter allows us to introduce and defined a unified set of notations that is adopted throughout the thesis. Nonetheless, the reader accustomed to these topics can skip that chapter.

In Section 2.2 the discrete counterparts of the one-dimensional and multi-dimensional signals are introduced and the definitions of inner product, Euclidean norm, linear and circular convolutions of discrete signals are provided.

Finally, in Section 2.3 we define the Fourier transformations for both continuous and discrete signals. Moreover, we present the local Fourier transform and its inverse operation, a useful and fundamental tool for the processing of signals that show localized information, e.g. time varying acoustic signals.

2.1 Continuous Signals

2.1.1 One-Dimensional Continuous Signals

Let us define as one-dimensional continuous signals the set of complex functions of a single continuous variable defined on the domain of real numbers that form the vector space $\mathbb{C}^{\mathbb{R}}$ as [271]

$$f(t), \quad f : \mathbb{R} \to \mathbb{C}. \tag{2.1}$$

The inner product of a pair of one-dimensional functions (e.g. $f_1(t),\ f_2(t) \in \mathbb{C}^{\mathbb{R}}$) is defined as [271]

$$\langle f_1(t), f_2(t) \rangle = \int_{-\infty}^{+\infty} f_1(t) f_2^*(t) dt, \tag{2.2}$$

where $(\cdot)^*$ represents the complex conjugate operator. From (2.2), the Euclidean norm ℓ_2 is derived as [271]

$$\|f\| = \sqrt{\langle f, f \rangle} = \left(\int_{-\infty}^{+\infty} |f(t)|^2 dt \right)^{\frac{1}{2}}. \tag{2.3}$$

The energy of the signal f is defined as the square of (2.3) $\|f\|^2$. In this manuscript, we focus on finite-energy one-dimensional continuous time signals for which the ℓ_2 is finite, namely $\|f\| < \infty$, and t represents the time.

The convolution between two one-dimensional continuous signals $f_1(t)$ and $f_2(t)$ is defined as [271]

$$(f_1 * f_2)(t) = \int_{-\infty}^{\infty} f_1(\tau) f_2(t - \tau) d\tau = \int_{-\infty}^{\infty} f_1(t - \tau) f_2(\tau) d\tau. \tag{2.4}$$

2.1.2 One-Dimensional Continuous Periodic Signals

A one-dimensional continuous periodic signal is a continuous function that satisfies the following equation

$$f(t + T) = f(t), \quad t \in \mathbb{R}, \tag{2.5}$$

where T refers to the period of the function. In general, the ℓ_2 norm of a continuous periodic signal is not finite [271], hence, the energy of such signals is defined over one period as

$$\|f\|^2 = \int_{-\frac{T}{2}}^{\frac{T}{2}} |f(t)|^2 dt. \tag{2.6}$$

The circular convolution between two one dimensional periodic signals is defined in [271] as

$$(f_1 \circledast f_2)(t) = \int_{-\frac{T}{2}}^{\frac{T}{2}} f_1(\tau) f_2(t - \tau) d\tau = \int_{-\frac{T}{2}}^{\frac{T}{2}} f_1(t - \tau) f_2(\tau) d\tau. \tag{2.7}$$

The result of the circular convolution (2.7) is still periodic of period T, thus $(f_1 \circledast f_2)(t) = (f_1 \circledast f_2)(t + T)$.

2.1.3 Multi-Dimensional Continuous Signals

The multi-dimensional continuous signals are defined as a set of functions of D variables in the domain of real numbers forming the vector space $\mathbb{C}^{\mathbb{R}^D}$

$$f(\mathbf{r}), \quad f : \mathbb{R}^D \to \mathbb{C}, \tag{2.8}$$

where $\mathbf{r} = [r_0, \ldots, r_{D-1}] \in \mathbb{R}^D$ is the vector of the independent variables. When two multi-dimensional signals of D variables are considered, the inner product is defined as

$$\langle f_1, f_2 \rangle = \int_{\mathbf{r} \in \mathbb{R}^D} f_1(\mathbf{r}) f_2^*(\mathbf{r}) d\mathbf{r}, \tag{2.9}$$

while the ℓ_2 norm (Euclidean norm) is given by

$$\|f\| = \sqrt{\langle f, f \rangle} = \left(\int_{\mathbf{r} \in \mathbb{R}^D} |f(\mathbf{r})|^2 d\mathbf{r} \right)^{\frac{1}{2}}. \tag{2.10}$$

Similarly, to the one-dimensional case, in this thesis we consider finite-energy signals for which the ℓ_2 norm (2.10) is finite.

2.2 Discrete Signals

2.2.1 One-Dimensional Discrete Signals

A one-dimensional discrete signal or sequence, is a single variable function defined on the domain of integer numbers that forms the vector space $\mathbb{C}^{\mathbb{Z}}$. Typically, a discrete signals is obtained from the sampling of continuous signal and we denote it as

$$\mathbf{f} = [\ldots, f[-2], f[-1], f[0], f[1], f[2], \ldots]^T, \tag{2.11}$$

where $(\cdot)^T$ is the transpose operator and $f[n], = f(nT_s)$ with $n \in \mathbb{Z}$ refers to the discrete version of the continuous signal sampled with a sampling interval T_s. Here, we associate the independent variable with time, hence, n and T_s are also referred as time index and sampling period, respectively.

Commonly, only a finite N-length section of the infinite-length sequence (2.11) is considered

$$\mathbf{f} = [f[0], \ldots, f[N-1]] \in \mathbb{C}^N. \tag{2.12}$$

The inner product between two N-length one-dimensional discrete signals $\mathbf{f}_1$ and $\mathbf{f}_2$ is defined as [271]

$$\langle \mathbf{f}_1, \mathbf{f}_2 \rangle = \sum_{n=0}^{N-1} f_1[n] f_2^*[n] = \mathbf{f}_1^H \mathbf{f}_2, \quad \mathbf{f}_1, \mathbf{f}_2 \in \mathbb{C}^{\mathbb{N}}, \tag{2.13}$$

where $(\cdot)^H$ is the conjugate transpose operator. The inner product of (2.13) induces the Euclidean norm definition in the case of finite-length sequences

$$\|\mathbf{f}\| = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle} = \left(\sum_{n=0}^{N-1} |f[n]|^2 \right)^{\frac{1}{2}}. \tag{2.14}$$

Note that in the case of infinite-length sequences, the indices of the summations in (2.13) and (2.14) go from $-\infty$ to $+\infty$.

The convolution between two sequences $\mathbf{f}_1$, $\mathbf{f}_2 \in \mathbb{C}^{\mathbb{Z}}$ is defined as [271]

$$(f_1 * f_2)[n] = \sum_{k \in \mathbb{Z}} f_1[k] f_2[n-k] = \sum_{k \in \mathbb{Z}} f_1[n-k] f_2[k], \tag{2.15}$$

while the circular convution between two N-length discrete signals is given by [271]

$$
\begin{aligned}
(f_1 \circledast f_1)[n] &= \sum_{k=0}^{N-1} f_1[k] f_2[(n-k) \bmod N] = \\
&= \sum_{k=0}^{N-1} f_1[(n-k) \bmod N] f_2[k],
\end{aligned} \tag{2.16}
$$

2.3 Fourier Transform

2.3.1 One-Dimensional Fourier Transform

The Fourier Transform of a one-dimensional complex continuous function $f(t) \in \mathbb{C}^{\mathbb{R}}$ with $t \in \mathbb{R}$ is defined as

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt, \tag{2.17}$$

where $\omega \in \mathbb{R}$ is the *angular frequency* and $j = \sqrt{-1}$ is the imaginary unit. In the scope of this thesis, we interpret t as time and it is standard practice to consider $\omega/2\pi$ as the temporal frequency measured in Hz, while ω is in $\mathrm{rad\,s^{-1}}$.

The inverse Fourier Transform of the signal $F \in \mathbb{C}^{\mathbb{R}}$ is defined as

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{j\omega t} d\omega, \tag{2.18}$$

with $t \in \mathbb{R}$. The following notation

$$
\begin{aligned}
F(\omega) &= \mathcal{F}_t\{f(t)\}, \\
f(t) &= \mathcal{F}_t^{-1}\{F(\omega)\},
\end{aligned} \tag{2.19}
$$

is adopted for referring to the Fourier Transform (2.17) and its inverse (2.18), respectively.

2.3.2 Multi-Dimensional Fourier Transform

The one-dimensional Fourier Transform (2.17) can be straightforwardly generalized to multi-dimensional signals i.e. $\mathbb{C}^{\mathbb{R}^D}$ as follows

$$F(\mathbf{k}) = \int_{-\infty}^{+\infty} f(\mathbf{r}) e^{-j\langle \mathbf{k}, \mathbf{r} \rangle} d\mathbf{r}, \quad \mathbf{k} \in \mathbb{R}^D. \tag{2.20}$$

Similarly, the inverse Fourier Transform of a multi-dimensional signal is defined as

$$f(\mathbf{r}) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\mathbf{k}) e^{j\langle \mathbf{k}, \mathbf{r} \rangle} d\mathbf{k}, \quad \mathbf{r} \in \mathbb{R}^D, \tag{2.21}$$

and with the compact notation operators (2.20), (2.21) become

$$
\begin{aligned}
F(\mathbf{k}) &= \mathcal{F}_{\mathbf{r}}\{f(\mathbf{r})\}, \\
f(\mathbf{r}) &= \mathcal{F}_{\mathbf{r}}^{-1}\{F(\mathbf{k})\}.
\end{aligned} \tag{2.22}
$$

2.3.3 Fourier Series

The Fourier transformation of periodic signals in the vector space $\mathcal{L}^2([-T/2,\,T/2])$ defines a discrete complex function known as the Fourier Series coefficient sequence [271]

$$F[k] = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)e^{-j(2\pi/T)kt}dt, \quad k \in \mathbb{Z}, \tag{2.23}$$

where $f(t) \in \mathbb{C}^{\mathbb{R}}$ is a periodic function with period T and k, defined on the domain of the integer numbers $\mathbb{Z}$, is the discrete frequency that indexes the multiples of the fundamental $\omega_0 = 2\pi/T$. The signal $f(t)$ can be reconstructed from the coefficients of (2.23) as

$$f(t) = \sum_{k\in\mathbb{Z}} F[k]e^{j(2\pi/T)kt}, \quad t \in [-T/2, T/2]. \tag{2.24}$$

2.3.4 Discrete-Time Fourier Transform

The discrete-time Fourier transform is the Fourier transformation related to infinite-length discrete signals in $\mathbb{C}^{\mathbb{Z}}$ and it is defined as

$$F(\omega) = \sum_{n=-\infty}^{+\infty} f[n]e^{j\omega n}, \quad \omega \in \mathbb{R}. \tag{2.25}$$

The discrete-time Fourier transform exists in the case when (2.25) converges for all $\omega \in \mathbb{R}$ and $F(\omega)$ is a periodic function of period 2π. The inverse of the discrete-time Fourier transform is defined as

$$f[n] = \frac{1}{2\pi}\int_0^{2\pi} F(\omega)e^{j\omega n}d\omega, \quad n \in \mathbb{Z}. \tag{2.26}$$

2.3.5 Discrete Fourier Transform

The discrete Fourier transform of a L-length discrete signal $\mathbf{f}$ is defined as

$$F[k] = \sum_{n=0}^{N-1} f[n]e^{-j(2\pi/N)kn}, \quad k \in \{0, \dots, N-1\}. \tag{2.27}$$

The inverse operator of (2.27), called inverse discrete Fourier transform is defined as

$$f[k] = \frac{1}{N}\sum_{n=0}^{N-1} F[k]e^{j(2\pi/N)kn}, \quad n \in \{0, \dots, N-1\}. \tag{2.28}$$

The following notation is introduced

$$\begin{aligned} F[k] &= \mathcal{DFT}_N\{f[n]\}, \\ f[n] &= \mathcal{DFT}_N^{-1}\{F[k]\}, \end{aligned} \tag{2.29}$$

to denote the direct and the inverse discrete Fourier transforms, respectively.

2.4 Local Fourier Transform

The different Fourier transformations reviewed in Sec. 2.3 represent a fundamental tool for the processing of signals. They perform a global analysis of the signal, since the information encoded by the time dependent function is integrated and projected onto the Fourier basis. Nevertheless, for some applications, it is interesting to analyze the local characteristics of the signals rather than their global properties. As an example, one might want to follow the temporal evolution of the frequency content of a signal. Therefore, the Fourier transform is not suitable for such task since it is defined on the whole signal duration. In order to processes the signals *locally*, in [106] a windowed Fourier transform, also referred as local Fourier transform, short-time Fourier transform or short-term Fourier transform was proposed. In the context of this thesis, we adopt the term local Fourier transform.

2.4.1 Local Continuous Fourier Transform

The local Fourier transform of a signal $f(t) \in \mathbb{C}^R$ is defined as

$$F(\omega,\tau) \int_{t\in\mathbb{R}} f(t)w(t-\tau)e^{-j\omega t}dt, \tag{2.30}$$

where $\omega, \tau \in \mathbb{R}$ and $w(t)$ is a real valued even function with support confined to a small interval about zero and usually, it is referred as window function. The inverse local Fourier transform is then defined as

$$f(t) = \frac{1}{2\pi} \int_{\omega\in\mathbb{R}} \int_{\tau\in\mathbb{R}} F(\omega,\tau)w(t-\tau)e^{j\omega t}d\omega d\tau. \tag{2.31}$$

In [49] the author proved that in order to have the inverse local Fourier transform to hold the following assumptions must be met

1. $\int_{\mathbb{R}} |w(t)|dt < \infty$;
2. $\int_{\mathbb{R}} |w(t)|^2 dt = C$;

where C is a constant different from zero and it is commonly assumed equal to one in order to have a window function of unitary energy.

2.4.2 Local Discrete Fourier Transform

The discrete version of the local Fourier transform is defined for a N-length sequence $f \in \mathbb{C}^N$ as

$$F[\tau,k] = \sum_{n=0}^{N-1} f[n]w[n-\tau]e^{-j(2\pi/N)kn}, \tag{2.32}$$

with $\tau \in \{0,\ldots,N-1\}$ and $k \in \{0,\ldots,N-1\}$. The discrete counterpart of the inverse local Fourier transform is defined as

$$f[n] = \frac{1}{N} \sum_{\tau=0}^{N-1} \sum_{k=0}^{N-1} F[\tau,k]w[n-\tau]e^{-j(2\pi/N)kn}. \tag{2.33}$$

Transform	Analysis/Synthesis	Characteristics
Fourier transform	$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt$ $f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{j\omega t} d\omega,$	
Fourier series	$F[k] = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-j(2\pi/T)kt}$ $f(t) = \sum_{k \in \mathbb{Z}} F[k] e^{j(2\pi/T)kt}$	Dual with DTFT $f(t+T) = f(t)$
Discrete-time Fourier transform	$F(\omega) = \sum_{n=-\infty}^{+\infty} f[n] e^{j\omega n}$ $f[n] = \frac{1}{2\pi} \int_{0}^{2\pi} F(\omega) e^{j\omega n} d\omega$	Dual with Fourier series $F(e^{j\omega + 2\pi}) = F(e^{j\omega})$
Discrete Fourier transform	$F[k] = \sum_{n=0}^{N-1} f[n] e^{-j(2\pi/N)kn}$ $f[k] = \frac{1}{N} \sum_{n=0}^{N-1} F[k] e^{j(2\pi/N)kn}$	
Local Fourier transform	$F(\omega, \tau) \int_{-\infty}^{\infty} f(t) w^*(t-\tau) e^{-j\omega t} dt$ $f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\omega, \tau) w^*(t-\tau) e^{j\omega t} d\omega d\tau$	
Local discrete Fourier transform	$F[\tau, k] = \sum_{n=0}^{N-1} f[n] w^*[n-\tau] e^{-j(2\pi/N)kn},$ $f[n] = \frac{1}{N} \sum_{\tau=0}^{N-1} \sum_{k=0}^{N-1} F[\tau, k]] w[n-\tau] e^{-j(2\pi/N)kn}$	

Table 2.1: *Summary of the introduced Fourier transform definitions given in Section 2.3 and Section 2.4.*

CHAPTER 3

Background on Acoustics

This chapter reviews the physical laws governing the phenomena of sound generation and propagation in space. We consider the sound or acoustic field as a scalar function. The domain of sound fields is represented by the union of the temporal and spatial domains, hence the sound field is a real-valued scalar function

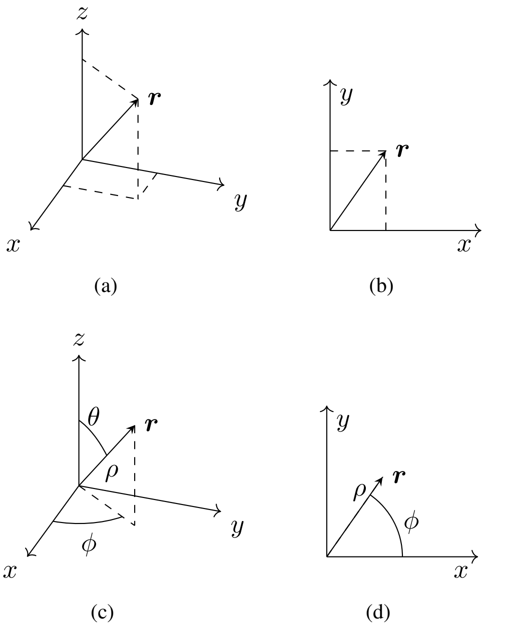
$$p(\boldsymbol{r}, t), \quad \boldsymbol{r} \in \mathbb{R}^3, \quad t \in \mathbb{R} \tag{3.1}$$

where $\boldsymbol{r}$ represents the spatial coordinates and t is time. Thanks to the property of invariance under space coordinates transformations [175], the sound field can be equivalently expressed adopting different representations of the spatial variable. In this chapter, basic linear acoustics is reviewed, therefore the reader accustomed to the topic can skip this chapter.

In Section 3.1, we review the common choices for 3D and 2D coordinate systems. A review of wave acoustics is given by the summary on the wave equation of Section 3.2, with the formulation of the homogeneous wave equation in Section 3.2.1 and the inhomogeneous wave equation in Section 3.2.2. Solutions to wave equations are given in Section 3.3 and Section 3.4 according to the considered formulation, while in Section 3.5 the acoustic field is expressed through the well-known Kirchoff-Helmholtz integral. Finally, in Section 3.6 the representation of acoustic fields in terms of plane waves and spherical waves are given and the exterior and interior problems are described.

3.1 Coordinate Systems

In this section, the notation adopted in the rest of the thesis to denote spatial coordinate systems is introduced both for 3D and 2D case.



(a)

(b)

(c)

(d)

Figure 3.1: *Spatial coordinate systems: (a) 3D Cartesian coordinate system, (b) 2D Cartesian coordinate system, (c) Spherical coordinate system and (d) Polar coordinate system.*

The position vector in 3D Cartesian coordinates $\boldsymbol{r} = [x, y, z]^T$ indicates the three spatial coordinates in a right-handed orthogonal coordinate system, as depicted in Figure 3.1(a). For a 2D Cartesian system of coordinates, the position vector is denoted as $\boldsymbol{r} = [x, y]^T$, with the spatial coordinates on the two perpendicular oriented axes (see Figure 3.1(b)).

For some applications, it is convenient to adopt a spherical reference frame, in which the position vector is defined as the radial distance ρ from the origin, the azimuth ϕ and the inclination (also called co-elevation) θ angles as shown in Figure 3.1(c). The relationship between the spherical coordinates and the 3D Cartesian coordinates is given by

$$\begin{aligned} x &= \rho \sin(\theta) \cos(\phi), \\ y &= \rho \sin(\theta) \sin(\phi), \\ z &= \rho \cos(\theta), \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} \rho &= \sqrt{x^2 + y^2 + z^2}, \\ \phi &= \arccos\left(\frac{x}{\sqrt{x^2 + y^2}}\right) = \arcsin\left(\frac{y}{\sqrt{x^2 + y^2}}\right), \\ \theta &= \arccos\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right). \end{aligned} \tag{3.3}$$

In a 2D setting, the polar reference system can be adopted, where the position vector is defined by the radius ρ and polar angle ϕ as depicted in Figure 3.1(d). The relationship between polar coordinates and 2D Cartesian coordinates is defined as

$$\begin{aligned}\rho &= \sqrt{x^2+y^2},\\ \phi &= \arctan\left(\frac{y}{x}\right).\end{aligned} \tag{3.4}$$

3.2 The Wave Equation

3.2.1 Homogeneous Wave equation

Under linear assumption, the sound field is described by small variations of the pressure amplitude $p(\boldsymbol{r},t)$ as a function of the space $\boldsymbol{r}$ and time t. In the case of a source-free volume of homogeneous medium, the pressure p must satisfy the homogeneous wave equation [286]

$$\nabla^2 p(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{r},t)}{\partial t} = 0, \tag{3.5}$$

where ∇^2 denotes the Laplace operator and c is the speed of sound in the medium. According to the adopted coordinate system, the Laplace operator assumes different formulations. In particular, when Cartesian coordinates are considered, ∇^2 becomes

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \tag{3.6}$$

while in spherical coordinates it reads

$$\nabla^2 = \frac{1}{\rho^2}\frac{\partial}{\partial \rho}\left(\rho^2\frac{\partial}{\partial \rho}\right) + \frac{1}{\rho^2 \sin(\theta)}\frac{\partial}{\partial \theta}\left(\sin(\theta)\frac{\partial}{\partial \theta}\right) + \frac{1}{\rho^2 \sin^2(\theta)}\frac{\partial^2}{\partial \phi^2}. \tag{3.7}$$

In many contexts, the sound field is generated by the harmonic motion of a source, hence it is convenient to consider the sound field as a function of the temporal frequency and space. It follows that the time-harmonic dependence of $p(\boldsymbol{r},t)$ can be expressed as

$$p(\boldsymbol{r},t) = P(\boldsymbol{r},\omega)e^{j\omega t}, \tag{3.8}$$

with $P(\boldsymbol{r},\omega) = \mathcal{F}_t\{p(\boldsymbol{r},t)\}$ the temporal Fourier transform of the pressure. Therefore, the wave equation (3.5) can be rewritten applying the Fourier transform (2.17) as

$$\begin{aligned}\mathcal{F}_t\left\{\nabla^2 p(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{r},t)}{\partial t^2}\right\} &= 0,\\ \nabla^2\mathcal{F}_t\left\{p(\boldsymbol{r},t)\right\} - \frac{1}{c^2}\mathcal{F}_t\left\{\frac{\partial^2 p(\boldsymbol{r},t)}{\partial t^2}\right\} &= 0,\end{aligned} \tag{3.9}$$

which results in the well-known homogeneous Helmholtz equation [286]

$$\nabla^2 P(\boldsymbol{r},\omega) + \left(\frac{\omega}{c}\right)^2 P(\boldsymbol{r},\omega) = 0. \tag{3.10}$$

3.2.2 Inhomogeneous Wave Equation

When the considered volumes are not free of sources, the homogeneous wave equation cannot be adopted for describing the sound field under analysis. In order to take into account of the energy emitted by the source, an excitation term is included in (3.5) yielding the inhomogeneous wave equation [286]

$$\nabla^2 p(\boldsymbol{r}, t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{r}, t)}{\partial t} = -q(\boldsymbol{r}, t), \tag{3.11}$$

where $q(\boldsymbol{r}, t) = Q(\boldsymbol{r}, \omega)e^{j\omega t}$ is the source function. Similarly to (3.10), the inhomogeneous Helmholtz equation is derived by applying the Fourier transform to (3.11)

$$\nabla^2 P(\boldsymbol{r}, \omega) + \left(\frac{\omega}{c}\right)^2 P(\boldsymbol{r}, \omega) = -Q(\boldsymbol{r}, \omega). \tag{3.12}$$

3.3 Solution to the Homogeneous Wave Equation

In this section, we review the solution of the homogeneous Helmholtz equation (3.10) that leads to well-known time-harmonic sound fields, characterized by the choice of the adopted coordinate system.

3.3.1 Plane Waves

When a Cartesian reference frame is adopted, candidate solutions of the homogeneous Helmholtz equation take the forms of complex exponential functions as [134, 286]

$$P(\boldsymbol{r}, \omega) = e^{j\langle \mathbf{k}, \boldsymbol{r}\rangle}, \tag{3.13}$$

where $\mathbf{k} = [k_x, k_y, k_z]^T$ denotes the wavenumber or propagation vector. The substitution of (3.13) in (3.10) leads to acceptable solutions if the dispersion relation is satisfied

$$\|\mathbf{k}\|_2^2 = \left(\frac{\omega}{c}\right)^2. \tag{3.14}$$

The condition in (3.14) relates the wavenumber vector with the frequency and given that $\left(\frac{\omega}{c}\right)$ is constant, the components of $\mathbf{k} = [k_x, k_y, k_z]^T$ are not independent of one another. It follows, that letting, as instance, k_z be the dependent variable, in order to satisfy (3.14) the following relation must hold

$$k_z^2 = \left(\frac{\omega}{c}\right)^2 - k_x^2 - k_y^2 \tag{3.15}$$

with k_x and k_y ranging from $-\infty$ to ∞.

Therefore, from (3.15) two distinct cases can be identified. When $(k_x^2 + k_y^2) \le \left(\frac{\omega}{c}\right)^2$, k_z assumes real values, i.e., $k_z \in \mathbb{R}$, otherwise it becomes a purely imaginary number, namely, $k_z = jk_z'$, $k_z' \in \mathbb{R}$.

In the first case, when $k_z \in \mathbb{R}$, the solutions of (3.10) are propagating plane waves, characterized by wavefronts that are defined by the location of points where the phase of the function is constant i.e.,

$$\angle P(\boldsymbol{r}, \omega) = \langle \mathbf{k}, \boldsymbol{r}\rangle = C, \tag{3.16}$$

with $C \in \mathbb{R}$ being a constant. It is clear from (3.16), that wavefronts are perpendicular to the wavenumber vector $\mathbf{k}$. Due to this property, (3.13) is named plane wave. It is worth noting that from the wavenumber vector $\mathbf{k}$, the Direction of Arrival (DoA) of the plane wave is derived as the unit vector $\hat{\mathbf{k}} = \mathbf{k}/\|\mathbf{k}\|$. We can express the solution of the homogeneous wave equation (3.5) by adding the time-harmonic dependence to (3.13)

$$p(\boldsymbol{r}, t) = P(\boldsymbol{r}, \omega)e^{j\omega t} = e^{j(\langle \mathbf{k}, \boldsymbol{r}\rangle + \omega t)}. \tag{3.17}$$

When k_z takes purely imaginary values, the solution of (3.10) corresponds to evanescent plane waves. The solution of the homogeneous wave equation (3.5) thus becomes

$$p(\boldsymbol{r}, t) = e^{j(\langle \mathbf{k}, \boldsymbol{r}\rangle + \omega t)} = e^{-jk'_z z} e^{j(k_x x + k_y y)} e^{j\omega t}. \tag{3.18}$$

In order obtain a solution that provide physical sense, it is customary to consider only $k'_z \in \mathbb{R}^+$. Through the term $e^{-jk'_z z}$ in (3.18), this choice leads to exponentially decaying plane waves as z increases, that would otherwise results in an energy gain as the plane waves propagate.

3.3.2 Spherical Waves

A second fundamental formulation of the solution to the homogeneous Helmholtz equation is expressed adopting a spherical reference frame. Employing the definition of the Laplace operator in spherical coordinates (3.7), equation (3.10) can be rewritten as

$$\begin{aligned} &\frac{1}{\rho^2}\frac{\partial}{\partial \rho}\left(\rho^2 \frac{\partial P(\boldsymbol{r}, \omega)}{\partial \rho}\right) + \frac{1}{\rho^2 \sin(\theta)}\frac{\partial}{\partial \theta}\left(\sin(\theta)\frac{\partial P(\boldsymbol{r}, \omega)}{\partial \theta}\right) \\ &+ \frac{1}{\rho^2 \sin^2(\theta)}\frac{\partial^2 P(\boldsymbol{r}, \omega)}{\partial \phi^2} + \left(\frac{\omega}{c}\right)^2 P(\boldsymbol{r}, \omega) = 0. \end{aligned} \tag{3.19}$$

The solution to (3.19) is obtained by separation of variables [286] and it consists in the multiplication of different functions each of them depending on one of the three variables (ρ, ϕ, θ) as

$$P(\boldsymbol{r}, \omega) = R(\rho)\Phi(\phi)\Theta(\theta). \tag{3.20}$$

For a detailed discussion and the mathematical derivation of (3.20) the reader is referred to [286].

Sometime, (see Section 3.6.2) it is useful to represent the solution for the angular variables in terms of the spherical harmonics functions

$$Y_{lm}(\theta, \phi) = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}}\mathcal{P}_{lm}\left(\cos(\theta)\right)e^{jm\phi}, \tag{3.21}$$

where $\mathcal{P}_{lm}$ is the associated Legendre function of order l and degree m with $|m| < l$ and it is defined for positive values of m as

$$\mathcal{P}_{lm}(x) = (-1)^m(1-x^2)^{m/2}\frac{\partial^m}{\partial x^m}\mathcal{P}_l(x) \tag{3.22}$$

with $\mathcal{P}_l(x)$ the associated Legendre function, while, for negative values of m it is given by

$$\mathcal{P}_{l(-m)}(x) = (-1)^m \frac{(l-m)!}{(l+m)!}\mathcal{P}_{lm}(x). \tag{3.23}$$

Given (3.23) one obtains that

$$Y_{l(-m)}(\theta, \phi) = (-1)^m Y_{lm}^*(\theta, \phi), \quad m > 0. \tag{3.24}$$

Hence, a convenient expression of the spherical harmonics is the following

$$Y_{lm}(\theta, \phi) = \sqrt{\frac{(2l+1)}{4\pi} \frac{(l-|m|)!}{(l+|m|)!}} \mathcal{P}_{l|m|}\left(\cos(\theta)\right) e^{jm\phi}. \tag{3.25}$$

The dependence on the radial distance in (3.20) is expressed by $R(\rho)$ in terms of the spherical Bessel function as

$$R(\rho) = R_1 j_l((\omega/c)\rho) + R_2 y_l((\omega/c)\rho) \tag{3.26}$$

where R_1 and R_2 are constants, $j_l(\cdot)$ is the spherical Bessel function of the first kind and $y_l(\cdot)$ is the spherical Bessel function of the second kind. As an alternative, $R(\rho)$ can be expressed in terms of the spherical Hankel function as

$$R(\rho) = R_3 h_l^{(1)}((\omega/c)\rho) + R_4 h_l^{(2)}((\omega/c)\rho) \tag{3.27}$$

with $h_l^{(1)}(\cdot)$ the spherical Hankel functions of the first kind and $h_l^{(2)}(\cdot)$ the spherical Hankel functions of the second kind. The values R_1, R_2, R_3 and R_4 are properly set according to the location of the acoustic sources in the scene. In particular, when the sound field is generated by sources far from the origin and observed around the origin, condition also known as incoming sound field, the spherical Bessel functions of the first kind are adopted. On contrary, in an outgoing sound field case, namely a sound field generated by sources closed to the origin and observed in a region far from the sources, spherical Hankel functions are employed. More details on the differences between the two settings are given in Section 3.6.2, where the concepts of the interior and exterior problem are described.

3.4 Solution to the Inhomogeneous Wave Equation

In this section, we introduce the solutions to the inhomogeneous wave equation (3.11), which is adopted when the sound field is generated by a sound source in the volume under analysis. Again, we assume a time-harmonic source, that let us focus on the solutions to the inhomogeneous Helmholtz equation (3.12).

A remarkable solution to (3.12), that can be used in the construction of arbitrary solutions, is known as Green's function and it is given by the adoption of a spatial impulse at $\boldsymbol{r}'$ as the source term $Q(\boldsymbol{r}, \omega) = \delta(\boldsymbol{r} - \boldsymbol{r}')\delta(\omega)$. The inhomogeneous Helmholtz equation is thus rewritten as

$$\nabla^2 G\left(\boldsymbol{r}|\boldsymbol{r}', \omega\right) + \left(\frac{\omega}{c}\right)^2 G\left(\boldsymbol{r}|\boldsymbol{r}', \omega\right) = -\delta\left(\boldsymbol{r} - \boldsymbol{r}'\right) \tag{3.28}$$

whose solution is given by Green's function

$$G\left(\boldsymbol{r} \mid \boldsymbol{r}', \omega\right) = \frac{e^{-j\frac{\omega}{c}\|\boldsymbol{r}-\boldsymbol{r}'\|_2}}{4\pi\left\|\boldsymbol{r} - \boldsymbol{r}'\right\|_2}. \tag{3.29}$$

More specifically, the Green's function (3.29) provides the sound field of a point source placed at $\boldsymbol{r}$ radiating in the free space. It is worth noting that the radiation described by (3.29) is omnidirectional and that $G(\boldsymbol{r}|\boldsymbol{r}') = G(\boldsymbol{r}'|\boldsymbol{r})$ since it depends only on the distance $\|\boldsymbol{r} - \boldsymbol{r}'\|$.

3.5 Integral Formulation of Sound Fields

This section reviews the formulation of acoustic fields in a confined volume. It follows that while the sound field is described by the Helmholtz equation inside the volume, some conditions have to be specified on the boundary. Different conditions exist according to the constrained quantity and the constraints specification. Hence, the acoustic field in an confined volume must also satisfy the boundary conditions.

3.5.1 Kirchoff-Helmholtz Integral Equation

Let us consider the sound field $P(\boldsymbol{r}, \omega)$ generated by a time-harmonic source excitation $F(\boldsymbol{r}, \omega)$ in a confined volume $\mathcal{V}$ with boundary $\partial\mathcal{V}$. The acoustic field must be a solution to the boundary value problem

$$\nabla^2 P(\boldsymbol{r}, \omega) + \left(\frac{\omega}{c}\right)^2 P(\boldsymbol{r}, \omega) = F(\boldsymbol{r}, \omega), \quad \boldsymbol{r} \in \mathcal{V}, \tag{3.30}$$

$$\alpha \langle \nabla P(\boldsymbol{r}, \omega), \hat{\mathbf{n}}(\boldsymbol{r}) \rangle + \beta P(\boldsymbol{r}, \omega) = 0, \quad \forall \boldsymbol{r} \in \partial\mathcal{V} \tag{3.31}$$

where $\hat{\mathbf{n}}(\boldsymbol{r})$ is a unit vector normal to $\partial\mathcal{V}$ at the location $\boldsymbol{r}$. We can recognize two contributions in the sound field $P(\boldsymbol{r}, \omega)$. The first component $P_0(\boldsymbol{r}, \omega)$ is given by the radiation of a source $F(\boldsymbol{r}, \omega)$ in an unbounded domain expressed as (3.29). The second component $P_1(\boldsymbol{r}, \omega)$ instead is a solution of the homogeneous Helmholtz equation that satisfies the boundary condition in (3.31). In fact, the expression in (3.31) provides the conditions for the pressure and its relative gradient on the boundary of the volume $\partial\mathcal{V}$.

By properly setting the parameters α and β different boundary conditions can be achieved, in particular,

- $\alpha = 0$, $\beta = 1$ results in the Dirichlet boundary condition (pressure-release boundary) [6];
- $\alpha = 1$, $\beta = 0$ results in the Neumann boundary condition (sound-hard boundary) [6];
- $\alpha = 1$, $\beta \neq 0$ results in the Robin boundary condition (absorbing boundary) [96];

The solution of (3.30) that satisfies the boundary condition (3.31) is given by the well-known Kirchoff-Helmholtz integral equation [286] defined as

$$\begin{aligned} a(\boldsymbol{r}) P(\boldsymbol{r}, \omega) = P_0(\boldsymbol{r}, \omega) + \int_{\partial\mathcal{V}} & G\left(\boldsymbol{r} \mid \boldsymbol{r}', \omega\right) \left\langle \nabla P\left(\boldsymbol{r}', \omega\right), \hat{\boldsymbol{n}}\left(\boldsymbol{r}'\right) \right\rangle \\ & - P\left(\boldsymbol{r}', \omega\right) \left\langle G\left(\boldsymbol{r} \mid \boldsymbol{r}', \omega\right), \hat{\boldsymbol{n}}\left(\boldsymbol{r}'\right) \right\rangle dA\left(\boldsymbol{r}'\right) \end{aligned} \tag{3.32}$$

where $A(\boldsymbol{r}')$ is an infinitesimal portion of $\partial\mathcal{V}$ while $a(\boldsymbol{r})$ is a parameter that takes the values [286]

$$a(\boldsymbol{r}) = \begin{cases} 1, & \text{for } \boldsymbol{r} \in \mathcal{V} \\ 0.5, & \text{for } \boldsymbol{r} \in \partial\mathcal{V} \\ 0, & \text{for } \boldsymbol{r} \notin \mathcal{V}. \end{cases} \tag{3.33}$$

Therefore, the sound field is composed by the superposition of three contributions

- the radiating term $P_0(\boldsymbol{r}, \omega)$ given by the source $F(\boldsymbol{r}, \omega)$ in an unbounded domain (3.30),
- the acoustic pressure on the boundary surface $\partial\mathcal{V}$,
- the pressure gradient in the direction normal to $\partial\mathcal{V}$.

The Kirchoff-Holmholtz integral equation (3.32) is at the foundation of many acoustic applications, among which Nearfield Acoustic Holography (NAH) [286] that is in the scope of this thesis (see Section 8.2).

3.6 Acoustic Field Representation

This section provides a review of the representations for sound fields that are directly derived from the solutions of the Helmholtz equation given in Section 3.3 and Section 3.4. The description of the acoustic field is therefore based on a set of independent basis functions and their relative coefficients, that are derived from such solutions and they differ according to the adopted reference system.

3.6.1 Plane Waves Representation

The adoption of plane waves as basis functions leads to the Whittaker's representation, in which an arbitrary sound field, satisfying the homogeneous Helmholtz equation in a region $\mathcal{V}$, is represented as an integral expansion of plane waves propagating in all directions. The plane waves representation provides the advantage of being simple and easily interpreted, since a plane wave is fully characterized by its direction of propagation and temporal frequency. Let us assume the sound field of a propagating plane wave in a Cartesian reference system

$$P(\boldsymbol{r}, \omega) = e^{j\langle \mathbf{k}, \boldsymbol{r}\rangle} \tag{3.34}$$

where $\mathbf{k} \in \mathbb{R}^3$ with the constraint of the dispersion relation (3.14) $||\mathbf{k}|| = \frac{\omega}{c}$. The sound field can be decomposed on a complete set of basis functions obtained by varying $\mathbf{k}$ fulfilling the dispersion relations. Therefore, the Whittaker's representation is defined as the multi-dimensional inverse Fourier transform with respect to space

$$P(\boldsymbol{r}, \omega) = \left(\frac{1}{2\pi}\right)^3 \iiint_{\mathcal{D}} \tilde{P}(\mathbf{k}) e^{j\langle \mathbf{k}, \boldsymbol{r}\rangle} d^3\boldsymbol{r}, \quad \mathcal{D} = \left\{\mathbf{k} \in \mathbb{R}^3 : ||\mathbf{k}|| = \frac{\omega}{c}\right\} \tag{3.35}$$

where $\tilde{P}(\mathbf{k})$ denotes the coefficients of plane waves at different spatial frequencies $\mathbf{k}$.

The coefficients $\tilde{P}(\mathbf{k})$ can be derived from the multi-dimensional Fourier transform of the sound field $P(\boldsymbol{r}, \omega)$ with respect to the spatial variable $\boldsymbol{r}$

$$\tilde{P}(\overline{\mathbf{k}}) = \iiint_{\mathbb{R}^3} P(\boldsymbol{r}, \omega) e^{-j\langle \overline{\mathbf{k}}, \boldsymbol{r}\rangle} d^3\overline{\mathbf{k}} \tag{3.36}$$

and constraining $\overline{\mathbf{k}}$ to satisfy the dispersion relation

$$\tilde{P}(\mathbf{k}) = \tilde{P}(\overline{\mathbf{k}})|_{||\overline{\mathbf{k}}||=||\mathbf{k}||=\frac{\omega}{c}}. \tag{3.37}$$

An intuitive geometric interpretation of the Whittaker's expansion is obtained factorizing the wavenumber as

$$\mathbf{k} = \frac{\omega}{c}\left[\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)\right]^T, \tag{3.38}$$

whose norm is ω/c and the vector is pointing in the direction given by (ϕ, θ), the azimuth and inclination angles, respectively. A different form of the Whittaker's expansion is then obtained by performing a proper change of variable in (3.35) and adopting (3.38)

$$\begin{aligned}&\tilde{P}(\overline{\mathbf{k}}) = \\ &\left(\frac{\omega}{c}\right)^3 \iint_{\mathcal{S}} \tilde{P}(\theta, \phi, \omega) e^{j\frac{\omega}{c}(x\sin(\theta)\cos(\phi)+y\sin(\theta)\sin(\phi)+z\cos(\theta))}\sin(\theta)d\theta d\phi,\end{aligned} \tag{3.39}$$

where the domain of integration is defined as $\mathcal{S} = \{\theta \in [0, \phi], \phi \in [0, 2\pi)\}$ In practice, we can interpret (3.39) as the superposition of propagating plane waves whose magnitudes and phases are encoded in the coefficients $\tilde{P}(\theta, \phi, \omega)$, while the directions are dictated by θ and ϕ,. It is worth mentioning that the values, $\tilde{P}(\theta, \phi, \omega)$, also known in the literature as *Herglotz density* are not dependent on the location of observation $\boldsymbol{r}$.

3.6.2 Spherical Waves Representation

The review of the spherical waves-based acoustic field representation is provided in this section. The solution of the homogeneous Helmholtz equation using spherical waves was shown in Section 3.3.2 as

$$P(\boldsymbol{r}, \omega) = R(\rho)Y_{lm}(\theta, \phi). \tag{3.40}$$

This formulation is characterized by the separation of the angular and radial distance contributions. In particular, the dependence on the angles is contained in the frequency independent spherical harmonics $Y_{lm}(\theta, \phi)$, while the dependency on the distance is given in $R(\rho)$ with forms

$$R(\rho) = R_1 j_l((\omega/c)\rho) + R_2 y_l((\omega/c)\rho) \tag{3.41}$$

or

$$R(\rho) = R_3 h_l^{(1)}((\omega/c)\rho) + R_4 h_l^{(2)}((\omega/c)\rho) \tag{3.42}$$

as discussed in Section 3.3.2.

In [286] the description of any solution to the homogeneous Helmholtz equation in terms of infinite sum of spherical waves is given as

$$\begin{aligned}P(\boldsymbol{r}, \omega) &= \sum_{l=0}^{+\infty}\sum_{m=-l}^{+l}\left(A_{lm}(\omega)j_l\left(\frac{\omega}{c}\rho\right) + B_{lm}(\omega)y_l\left(\frac{\omega}{c}\rho\right)\right)Y_{lm}(\theta, \phi) \\ &= \sum_{l=0}^{+\infty}\sum_{m=-l}^{+l}\left(C_{lm}(\omega)h_l^{(1)}\left(\frac{\omega}{c}\rho\right) + D_{lm}(\omega)h_l^{(2)}\left(\frac{\omega}{c}\rho\right)\right)Y_{lm}(\theta, \phi).\end{aligned} \tag{3.43}$$

In Section 3.3.2 we discussed the role of the terms $j_l(\cdot)$ and $h_l^{(1)}(\cdot)$ with the former suitable for describing the sound field internal to a source distribution and the latter adopted in the case of external acoustic fields.

Therefore, two different spherical waves representations arise according to the spatial arrangement of the sound sources and the volume under analysis. On the one side, we talk about interior problem, when the acoustic field is generated far outside of the analyzed region, namely we are interested in a volume free of sources or scatterers. On the other side, the exterior problem is defined for regions that includes the acoustic sources and scatterers. In the following we provide a discussion of the two representations.

Exterior Problem

In the case of the exterior problem, the dependency on the radial distance is described by means of the spherical Hankel function $h_l(\cdot)$. Hence, the external sound field is given as the expansion [286]

$$P(\boldsymbol{r},\omega) = \sum_{l=0}^{+\infty} \sum_{m=-l}^{l} C_{lm}(\omega) h_l^{(1)} \left(\frac{\omega}{c}\rho\right) Y_{lm}(\theta,\phi) \tag{3.44}$$

where the acoustic field is completely defined by the coefficients $C_{lm}(\omega)$. Under the assumption that the pressure $P(\boldsymbol{r},\omega)$ is known over the surface of a sphere of radius a, the whole external volume $\mathcal{V}$ can be completely determined by the knowledge of the coefficients of the expansion $C_{lm}(\omega)$. Let us recall the orthonormal property of the spherical harmonics

$$\int_0^{2\pi} \int_0^{\pi} Y_{lm}(\theta,\phi) Y^*_{l'm'}(\theta,\phi) \sin\theta d\phi d\theta = \delta_{ll'}\delta_{mm'}, \tag{3.45}$$

where the Kronecker delta is

$$\delta_{ab} \begin{cases} 0 & a \neq b \\ 1 & a = b \end{cases}, \tag{3.46}$$

exploiting (3.45) and applying a multiplication by $Y_{pq}(\theta,\phi)$ on each side of (3.44) after the integration over the surface of a unit sphere, we can determine the coefficients as

$$C_{nm}(\omega) = \frac{1}{h_l^{(1)}\left(\frac{\omega}{c}a\right)} \int_0^{2\pi} \int_0^{\pi} P(a,\theta,\phi,\omega) Y^*_{lm}(\theta,\phi) \sin\theta d\theta d\phi. \tag{3.47}$$

Finally, by inserting (3.47) into (3.44) we obtain the definition of the external pressure in all the locations with $\|\boldsymbol{r}\| > a$ as

$$\begin{aligned} &P(\boldsymbol{r},\omega) \\ &= \textstyle\sum_{l=0}^{+\infty} \frac{h_l^{(1)}\left(\frac{\omega}{c}\rho\right)}{h_l^{(1)}\left(\frac{\omega}{c}a\right)} \sum_{m=-l}^{l} Y_{lm}(\theta,\phi) \int_0^{2\pi} \int_0^{\pi} P\left(a,\theta',\phi',\omega\right) Y_{lm}\left(\theta',\phi'\right)^* d\Gamma' \end{aligned} \tag{3.48}$$

where $d\Gamma' = \sin\theta' d\theta' d\phi'$.

Interior Problem

The interior problem concerns the case when acoustic sources are located outside a sphere with radius b where the origin lies. Hence, the spherical Bessel function $j_l(\cdot)$

that is finite at the origin, is adopted in the expansion of the internal sound field

$$P(\boldsymbol{r}, \omega) = \sum_{l=0}^{+\infty} \sum_{m=-l}^{l} A_{lm}(\omega) j_l \left(\frac{\omega}{c}\rho\right) Y_{lm}(\theta, \phi). \tag{3.49}$$

Again, the coefficients $A_{lm}(\omega)$ of the expansion can be obtained exploiting the knowledge of the pressure of the sphere of radius b as

$$A_{nm}(\omega) = \frac{1}{j_l\left(\frac{\omega}{c}b\right)} \int_0^{2\pi} \int_0^{\pi} P(b, \theta, \phi, \omega) Y_{lm}^*(\theta, \phi) \sin\theta d\theta d\phi. \tag{3.50}$$

Similarly to the exterior problem, substituting (3.50) in (3.49) the expression of the pressure field at the points inside a sphere of radius b, i.e., $\|\boldsymbol{r}\| < b$, is given as

$$\begin{aligned} &P(\boldsymbol{r}, \omega) \\ &= \textstyle\sum_{l=0}^{+\infty} \frac{j_l\left(\frac{\omega}{c}\rho\right)}{j_l\left(\frac{\omega}{c}b\right)} \sum_{m=-l}^{l} Y_{lm}(\theta, \phi) \int_0^{2\pi} \int_0^{\pi} P\left(b, \theta', \phi', \omega\right) Y_{lm}^*\left(\theta', \phi'\right) d\Gamma' \end{aligned} \tag{3.51}$$

where $d\Gamma' = \sin\theta' d\theta' d\phi'$.

Bandlimited Spherical Wave Representation

Inspecting both the exterior field expansion (3.44) and the interior acoustic field formulation (3.49), we can note that an infinite number of coefficients is required, since $0 \leq l < +\infty$. Nevertheless, in [132] it was shown that acoustic fields can be expressed truncating the expansion to a maximum order L

$$P(\boldsymbol{r}, \omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} C_{lm}(\omega) h_l^{(1)} \left(\frac{\omega}{c}\rho\right) Y_{lm}(\theta, \phi), \tag{3.52}$$

in the case of the exterior field, while for an interior field we obtain

$$P(\boldsymbol{r}, \omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} A_{lm}(\omega) j_l \left(\frac{\omega}{c}\rho\right) Y_{lm}(\theta, \phi). \tag{3.53}$$

The equations (3.52) and (3.53) are referred to as bandlimited spherical wave expansions. As shown in [132], the bandlimited representation allows us to describe the sound field with $(L+1)^2$ parameters, and with a truncation error that is upper bounded by $0.16127e^{-\Delta}$ if

$$L = \left\lceil \frac{e}{2} \frac{\omega}{c} \|\boldsymbol{r}\| \right\rceil + \Delta, \quad \Delta \in \mathbb{Z}^+. \tag{3.54}$$

Part II

Sound Field Processing for Extended Reality

CHAPTER 4

State of the art in Sound Field Reconstruction

In this chapter, we review the sound field reconstruction problem. In particular, we present the two main approaches to sound field reconstruction that are divided in non-parametric and parametric methods. The reconstruction of a given acoustic field is a fundamental task not limited to the context of EAR. Sound field reconstruction concerns the estimation of the acoustic field in a target location or region exploiting the information acquired by a set of microphone measurements. Clearly, the solutions to sound field reconstruction are driven by the availability of data, in terms of the sampling of the acoustic field and they are strictly linked to the sound field representations introduced in Section 3.6. In practical scenarios, a direct sampling of the sound field, fulfilling the Nyquist-Shannon condition is usually unfeasible due the high requirements in terms of number and position of the sensors. Therefore, different strategies to the reconstruction of the sound field have been proposed. They can mainly classified as non-parametric approaches, where the solutions of the wave equations are employed, and parametric methods that rely on compact models of the acoustic field. In Section 4.1 we review the state of the art of non-parametric techniques, while in Section 4.2 the parametric methods are introduced. The methods presented in this chapter have been selected due to their strong connection with the proposed sound field reconstruction technique in Chapter 5 or up-to-date solutions to the problem.

4.1 Non-Parametric Approaches

This section reviews the state-of-the-art non-parametric models presented in [88, 142]. We choose to review [88] since the processing, based on the spherical harmonic translation, introduced by the authors can be exploited also in the proposed parametric method in Chapter 5. As regards [142], it represents the state-of-the-art sparsity-based solution for non-parametric sound field reconstruction, therefore it is reviewed in order to provide the reader with up-to-date sound field reconstruction solutions. In general, non-parametric methods adopt the decomposition of the sound field into spatial Fourier basis that arises from the wave equation solutions reviewed in Section 3.3 and Section 3.4. The acoustic field is therefore represented, as discussed in Section 3.6, according to the adopted reference frame that is dictated by the microphone array setup, e.g., planar and spherical arrays lead to plane wave functions and spherical harmonics, respectively.

The technique proposed in [88] and reviewed in Section 4.1.1 is based on the spherical harmonics representation. In particular, the sound field reconstruction is accomplished exploiting the spherical harmonics addition theorem, sometimes referred as spherical harmonics translation theorem, which relates the spherical harmonics expansion locally estimated in a given position with the global description of the sound field. The authors in [88], hence exploit the local spherical harmonics decomposition given by Higher Order Microphones (HOM) and the spherical harmonics addition theorem for achieving the reconstruction of the global sound field.

In Section 4.1.2 we present the method developed in [142] that relies on a sparse representation of the acoustic field making distinction between the direct and the reverberant components of the sound field. Leveraging sparsity assumptions of the acoustic field, the direct and the reverberant components can be estimated from a set of microphones retrieving the solution for reconstructing the acoustic field.

4.1.1 Spherical Harmonics Translation Method

Here, we provide a review of the non-parametric technique for the reconstruction of sound fields introduced in [88]. The considered setup is composed of a source region, where the desired sources are located, while the interferers such as reverberation or background noise and additional sources are confined outside the region. By properly defining the origin, we can obtain a “source-less” receiver region that separates the desired sources from the interferers. The acquired signals are then represented as the superposition of the sound field generated by the desired sources (exterior acoustic field) and the interferers (interior field).

In [88] the authors propose an efficient technique in order to estimate the sound field over a large area that is based on the results of [228, 229], where measurements are performed with HOMs. However, the solution of [88] is limited to the exterior sound field only, while [229] includes also the interior component.

Problem Formulation

Let us consider, the scenario in which every acoustic source is located outside a spherical volume of radius R_0. By properly fixing the origin in $\boldsymbol{O}^{(s)}$ we can define the acoustic pressure using the interior formulation of the spherical harmonics expansion (3.49) as

$$S_I(\boldsymbol{r}, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}^{(0)}(k) j_n(k\rho) Y_{nm}(\theta, \phi) \tag{4.1}$$

where $\boldsymbol{r} = [\rho, \theta, \phi]^T$ denotes a point inside the region, $j_n(\cdot)$ is the spherical Bessel function and $\alpha_{nm}^{(0)}(k)$ are the interior sound coefficients of the spherical harmonics $Y_{nm}(\theta, \phi)$ (3.25) related to $\boldsymbol{O}^{(s)}$.

On the other hand, when all the sources are located inside the spherical region of radius R_s with $R_s < R_0$, the acoustic pressure at the locations outside the source region is described with respect to the origin $\mathbf{O}^{(s)}$ using the formulation of (3.44)

$$S_E(\boldsymbol{r}, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \beta_{nm}^{(0)}(k) h_n(k\rho) Y_{nm}(\theta, \phi) \tag{4.2}$$

where $h_n(\cdot)$ denotes the n-order spherical Hankel function and $\beta_{nm}^{(0)}(k)$ indicate the coefficients of the exterior sound field referred to $\mathbf{O}^{(s)}$.

Inspecting both (4.1) and (4.2) we can note that the two acoustic field definitions do not depend on the point location. Therefore, we can accurately obtain the whole sound field from the knowledge of the coefficients $\alpha_{nm}^{(0)}(k)$ and $\beta_{nm}^{(0)}(k)$. Hence, the goal of the methodology is to correctly estimate the interior and exterior coefficients given the signals of distributed HOMs.

Spherical Harmonics Addition Theorem

The spherical harmonics addition theorem represents the fundamental tool for sound field reconstruction in [88, 229].

We define the position vectors $\boldsymbol{r} = [\rho, \theta, \phi]^T$, $\boldsymbol{r}_q = [\rho_q, \theta_q \phi_q]^T$ $\boldsymbol{r}_r = [\rho_r, \theta_r \phi_r]^T$ such that $\boldsymbol{r} = \boldsymbol{r}_q + \boldsymbol{r}_r$. The spherical harmonics addition (or translation) theorem for the interior formulation with $j_n(k\rho)Y_{nm}(\theta, \phi)$ reads as [163]

$$j_n(k\rho)Y_{nm}(\theta, \phi) = \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\nu} \hat{S}_{n\nu}^{m\mu}(\boldsymbol{r}_q, k)\, j_\nu(k\rho_r)\, Y_{\nu\mu}(\theta_r, \phi_r) \tag{4.3}$$

where

$$\begin{aligned}\hat{S}_{n\nu}^{m\mu}(\boldsymbol{r}_q, k) =& 4\pi i^{\nu-n} \sum_{l=0}^{\infty} i^l (-1)^{2m-\mu} j_l(k\rho_q)\, Y_{l(\mu-m)}^*(\theta_q, \phi_q) \\ &\times \sqrt{\frac{(2n+1)(2\nu+1)(2l+1)}{4\pi}} W_1 W_2\end{aligned} \tag{4.4}$$

with W_1 and W_2 the Wigner $3-j$ symbols defined as

$$W_1 = \begin{pmatrix} n & \nu & l \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad W_2 = \begin{pmatrix} n & \nu & l \\ m & -\mu & (\mu - m) \end{pmatrix} \tag{4.5}$$

In the case of the exterior sound field, the addition theorem for $h_n(k\rho)Y_{nm}(\theta, \phi)$ is given by [163]

$$h_n(k\rho)Y_{nm}(\theta, \phi) = \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\nu} S_{n\nu}^{m\mu}(\boldsymbol{r}_q, k)\, j_\nu(k\rho_r)\, Y_{\nu\mu}(\theta_r, \phi_r) \tag{4.6}$$

where

$$\begin{aligned}S_{n\nu}^{m\mu}(\boldsymbol{r}_q, k) =& 4\pi i^{\nu-n} \sum_{l=0}^{\infty} i^l (-1)^{2m-\mu} h_l(k\rho_q)\, Y_{l(\mu-m)}^*(\theta_q, \phi_q) \\ &\times \sqrt{\frac{(2n+1)(2\nu+1)(2l+1)}{4\pi}} W_1 W_2\end{aligned} \tag{4.7}$$

The summations in (4.4) and (4.7) can be truncated to $I = n + \nu + 1$ according to the inherent properties of (4.5).

Sound Field Acquisition with HOMs

We consider a setup in which the recording of the sound field is performed by means of Q V^{th} order HOMs enclosing the desired region. The qth HOM location is indicated as $\boldsymbol{r}_q = [\rho_q, \theta_q, \phi_q]^T$ with $q = 1, \dots, Q$, while each sensor in the HOM is positioned at $\boldsymbol{r}_{q'} = [\rho_{q'}, \theta_{q'}, \phi_{q'}]^T$, with $q' = 1, \dots, Q'$ and coordinates $\boldsymbol{r}_{q'}$ expressed in terms of the HOM location $\boldsymbol{r}_q$. The signals of the qth HOM can be expressed in term of the Vth order spherical harmonics expansion and the coefficients can be obtained adopting (3.50) as

$$\alpha_{\nu\mu}^{(q)} = \frac{1}{b_\nu\left(k\rho_M\right)} \sum_{q'=1}^{Q'} S\left(\boldsymbol{r}_{q'}^{(q)}, k\right) Y_{\nu\mu}^{*}\left(\theta_{q'}^{(q)}, \phi_{q'}^{(q)}\right) \tag{4.8}$$

where $\nu = 0, \dots, V$, $\mu = -\nu, \dots, \nu$ and $b_\nu(k\rho_M)$ is defined as [217]

$$b_\nu\left(k\rho_M\right) = \begin{cases} j_\nu\left(k\rho_M\right) & \text{for open sphere array} \\ j_\nu\left(k\rho_M\right) - \frac{j_\nu'(k\rho_M)}{h_\nu'(k\rho_M)} h_\nu\left(k\rho_M\right) & \text{for rigid sphere array} \end{cases} \tag{4.9}$$

with $j_\nu'(\cdot)$ and $h_\nu'(\cdot)$ denoting the first derivatives of the spherical Bessel and Hankel function, respectively.

Interior Sound Field Estimation

The relation between the overall interior sound field coefficients $\alpha_{nm}^{(0)}$ of (4.1) and the local sound field ones $\alpha_{\nu\mu}^{(q)}$ of (4.8) can be obtained by expressing the sound field at the location $\boldsymbol{r}$ changing the origin with the qth HOM location $\boldsymbol{r}_q$

$$S_I(\boldsymbol{r}, k) = \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{n} \alpha_{\nu\mu}^{(q)}(k) j_\nu(k\rho_r) Y_{\nu\mu}(\theta_r, \phi_r). \tag{4.10}$$

Therefore, if we equate (4.1) and (4.10) inserting (4.3) into (4.1) the relation becomes

$$\begin{aligned} &\sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}^{(0)}(k) j_n(k\rho) Y_{nm}(\theta, \phi) = \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\nu} \alpha_{\nu\mu}^{(q)}(k) j_\nu\left(k\rho_r\right) Y_{\nu\mu}\left(\theta_r, \phi_r\right) \\ &\sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}^{(0)}(k) \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\nu} \hat{S}_{n\nu}^{m\mu}\left(\boldsymbol{r}_q, k\right) j_\nu\left(k\rho_r\right) Y_{\nu\mu}\left(\theta_r, \phi_r\right) = \\ &\sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\nu} \alpha_{\nu\mu}^{(q)}(k) j_\nu\left(k\rho_r\right) Y_{\nu\mu}\left(\theta_r, \phi_r\right), \end{aligned} \tag{4.11}$$

Finally, equating mode by mode (4.11) the relation between the global and the local interior sound coefficients can be derived

$$\alpha_{\nu\mu}^{(q)}(k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}^{(0)}(k) \hat{S}_{n\nu}^{m\mu}\left(\boldsymbol{r}_q, k\right). \tag{4.12}$$

Exploiting the Q HOMs employed in the setup, the relation between the local and the global coefficients can be expressed as the linear system

$$\boldsymbol{\alpha}^{Q}(k) = \mathbf{T}_I(k)\boldsymbol{\alpha}(k), \tag{4.13}$$

where the vector

$$\boldsymbol{\alpha}^{Q}(k) = [\alpha_{00}^{(1)}(k), \dots, \alpha_{VV}^{(1)}(k), \dots, \alpha_{00}^{(Q)}(k), \dots, \alpha_{VV}^{(Q)}(k)]^T \tag{4.14}$$

contains the local coefficients, while global coefficients are in

$$\boldsymbol{\alpha}(k) = [\alpha_{00}^{(0)}(k), \dots, \alpha_{N_I N_I}^{(1)}(k)]^T \tag{4.15}$$

and

$$\mathbf{T}_I(k) = \begin{bmatrix} \hat{S}_{00}^{00}\left(\boldsymbol{r}_1, k\right) & \cdots & \cdots & \hat{S}_{N_I 0}^{N_I 0}\left(\boldsymbol{r}_1, k\right) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{S}_{0V}^{0V}\left(\boldsymbol{r}_q, k\right) & \cdots & \cdots & \hat{S}_{N_I V}^{N_I V}\left(\boldsymbol{r}_q, k\right) \end{bmatrix}. \tag{4.16}$$

with the inifinte summation of (4.1) limited to N_I.

It follows that an estimate of the coefficients of the global interior sound can be obtained by inverting (4.13) as

$$\boldsymbol{\alpha}(k) = (\mathbf{T}_I(k))^{\dagger}\boldsymbol{\alpha}^{Q}(k). \tag{4.17}$$

Exterior Sound Field Estimation

Similarly to the interior sound field case, the relation between the exterior local sound field coefficients $\alpha_{\nu\mu}^{(q)}$ and the global $\beta_{nm}^{(0)}$ of (4.2) is expressed as

$$\alpha_{\nu\mu}^{(q)}(k) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \beta_{nm}^{(0)}(k)\hat{S}_{n\nu}^{m\mu}\left(\boldsymbol{r}_q, k\right). \tag{4.18}$$

Again, a system of linear equation can be set exploiting the Q HOMs as

$$\boldsymbol{\alpha}^{Q}(k) = \mathbf{T}_E(k)\boldsymbol{\beta}(k), \tag{4.19}$$

where the global coefficients are contained in the vector

$$\boldsymbol{\beta}(k) = [\beta_{00}^{0}(k), \dots, \beta_{N_E N_E}^{0}(k)]^T \tag{4.20}$$

and

$$\mathbf{T}_E(k) = \begin{bmatrix} S_{00}^{00}\left(\boldsymbol{r}_1, k\right) & \cdots & \cdots & S_{N_E 0}^{N_E 0}\left(\boldsymbol{r}_1, k\right) \\ \vdots & \vdots & \vdots & \vdots \\ S_{0V}^{0V}\left(\boldsymbol{r}_q, k\right) & \cdots & \cdots & S_{N_E V}^{N_E V}\left(\boldsymbol{r}_q, k\right) \end{bmatrix}, \tag{4.21}$$

with the infinite summation of (4.2) truncated at N_E. Hence, the coefficients of the global exterior sound field can be estimated by inverting (4.19) as

$$\boldsymbol{\beta}(k) = (\mathbf{T}_E(k))^{\dagger}\boldsymbol{\alpha}^{Q}(k). \tag{4.22}$$

Mixed Sound Field Estimation

In actual conditions, usually, both the interior and exterior components of the sound field are captured by the HOMs. Hence, we can easily derive the estimation of the global coefficients for both the sound field components merging (4.13) and (4.19) in the linear system of equations

$$\boldsymbol{\alpha}^Q(k) = \mathbf{T}(k)\mathbf{d}(k), \tag{4.23}$$

where

$$\begin{aligned} \mathbf{T}(k) &= [\mathbf{T}_I(k), \mathbf{T}_E(k)] \text{ and} \\ \mathbf{d}(k) &= \left[\boldsymbol{\alpha}^T(k), \boldsymbol{\beta}^T(k)\right]^T. \end{aligned} \tag{4.24}$$

Finally, an estimation of both the interior and exterior coefficients of the global sound field can be derived as

$$\mathbf{d}(k) = \mathbf{T}(k)^\dagger \boldsymbol{\alpha}^Q(k), \tag{4.25}$$

where the estimate is limited up to the N_Ith and N_Eth order for the interior and exterior coefficients, respectively.

In order to avoid the system in (4.23) to be under-determined, a minimum number of HOMs Q_{min} is required and it is defined according to

$$Q_{min} = \frac{(N_I + 1)^2 + (N_E + 1)^2}{(V + 1)^2}, \tag{4.26}$$

where the truncation limits are given by [229] as $N_E = \lceil keR_s/2 \rceil$ and $N_I = \lceil keR_0/2 \rceil$ with e the Euler's number and $\lceil \cdot \rceil$ the ceiling operator. It is worth noticing that the order required for an accurate estimation of the sound field components is directly related to the radius of the desired region and the frequency. Hence, the maximum frequency for which an aliasing-free reconstruction can be obtained, is related to the radius as $f_E = \frac{c[\sqrt{Q}(V+1)-1]}{\pi e R_s}$ and $f_I = \frac{c[\sqrt{Q}(V+1)-1]}{\pi e R_0}$.

4.1.2 Method Based on Sparse Sound Field Representation

In this section the non-parametric technique presented in [142] is reviewed. Here, the sound field is represented as the superposition of the direct source and the reverberant field. The technique is referred to as sparse, since it assumes a spatially sparse source distribution, while the reverberant field is represented by means of few plane wave components and a low-rank term. The coefficients describing the sound field are estimated by convex optimization using alternating direction method of multipliers [47].

Sound Field Model

Let us consider a region of interest Ω in which acoustic sources are present. The model adopted in [142] defines the inhomogenous sound field (3.12) at the location $\boldsymbol{r}$ as the sum of the source contribution (3.28) described in [142] as the particular solution and the homogeneous (3.10) term

$$P(t, \omega, \boldsymbol{r}) = P_P(t, \omega, \boldsymbol{r}) + P_H(t, \omega, \boldsymbol{r}), \tag{4.27}$$

where $P_P(t, \omega, \boldsymbol{r})$ is the particular solution defined as [286]

$$P_P(t, \omega, \boldsymbol{r}') = \int_{\boldsymbol{r}' \in \Omega} Q(t, \omega, \boldsymbol{r}') G(\boldsymbol{r}|\boldsymbol{r}', \omega) d\boldsymbol{r}', \tag{4.28}$$

with $G(\boldsymbol{r}|\boldsymbol{r}', \omega)$ the Green's function (3.29) and $Q(\cdot)$ the source distribution inside Ω. The homogeneous component $P_H(\cdot)$ can be represented as a linear combination of plane waves as discussed in Section 3.6.1 and in [142] the authors assumes that only a limited number of plane wave components contributes as

$$P_H(t, \omega, \boldsymbol{r}) \approx \sum_{l=1}^{I} \varphi_l(t, \omega) e^{j\mathbf{k}_l^T \boldsymbol{r}} \tag{4.29}$$

where $\mathbf{k}$ indicates the wave vector of the lth plane wave and $\varphi(t, \omega)$ are the coefficients of the Herglotz density.

An alternative model for $P_H(\cdot)$ is presented in [174], where the reverberant component is approximated as the superposition of J source signals

$$P_H(t, \omega, \boldsymbol{r}) \approx \sum_{j=1}^{J} h_j(\omega, \boldsymbol{r}) \varsigma_j(t, \omega), \tag{4.30}$$

where $h_j(\omega, \boldsymbol{r})$ is the direct-path-less transfer function between the jth source and $\boldsymbol{r}$ while $\varsigma_j(t, \omega)$ is the signal emitted by the jth source.

To reconstruct the sound field (4.27) a sparsity constraint in [142] is assumed on the spatial distribution of the sources of (4.28). It follows that the large source region Ω is discretized into a set of N smaller regions Ω_n with the index $n = 1, \dots, N$ denoting the different elements. Hence, only a sparse set of small regions will include the acoustic sources. Then, for each region, a representative point is defined as a grid point $\boldsymbol{r}_n$ and (4.28) is approximated as

$$\begin{aligned} P_P(t, \omega, \boldsymbol{r}) &= \sum_{n=1}^{N} \int_{\boldsymbol{r}' \in \Omega_n} Q(t, \omega, \boldsymbol{r}') G(\boldsymbol{r}|\boldsymbol{r}', \omega) d\boldsymbol{r}' \\ &= \sum_{n=1}^{N} G(\boldsymbol{r}|\boldsymbol{r}_n, \omega) \int_{\boldsymbol{r}' \in \Omega_n} Q(t, \omega, \boldsymbol{r}'). \end{aligned} \tag{4.31}$$

Considering a setup of M microphones, each one located at $\boldsymbol{r}_m$, $m = 1, \dots, M$ and the sound field models (4.31) and (4.29), a linear system of equations can be written as

$$\mathbf{y}(t, \omega) = \mathbf{D}(\omega)\mathbf{x}(t, \omega) + \mathbf{W}(\omega)\mathbf{u}(t, \omega), \tag{4.32}$$

where $\mathbf{D} \in \mathbb{C}^{M \times N}$ and $\mathbf{W} \in \mathbb{C}^{M \times I}$ are matrices with elements

$$\begin{aligned} [\mathbf{D}(\omega)]_{m,n} &= G(\boldsymbol{r}_m|\boldsymbol{r}_n, \omega), \\ [\mathbf{W}(\omega)]_{m,l} &= e^{j\mathbf{k}_l^T \boldsymbol{r}_m}, \end{aligned} \tag{4.33}$$

respectively, while $\mathbf{y}(t, \omega) = [P(t, \omega, \boldsymbol{r}_1), \dots, P(t, \omega, \boldsymbol{r}_M)]^T$ is the vector of signals, $\mathbf{x}(t, \omega) = \left[\int_{\boldsymbol{r}' \in \Omega_1} Q(t, \omega, \boldsymbol{r}'), \dots, \int_{\boldsymbol{r}' \in \Omega_N} Q(t, \omega, \boldsymbol{r}')\right]^T$ represents the source distribution and $\mathbf{u}(t, \omega) = [\varphi_1(t, \omega), \dots, \varphi_I(t, \omega)]^T$ contains the plane wave coefficients.

If (4.30) is adopted, then the system (4.32) becomes

$$\mathbf{y}(t,\omega) = \mathbf{D}(\omega)\mathbf{x}(t,\omega) + \mathbf{H}(\omega)\mathbf{s}(t,\omega), \tag{4.34}$$

where $\mathbf{H} \in \mathbb{C}^{M\times J}$ is the matrix of elements $[\mathbf{H}(\omega)]_{m,j} = h(\omega, \boldsymbol{r}_m)$ and the signal of the sources is contained in $\mathbf{s}(t,\omega) = [\varsigma_1(t,\omega), \dots, \varsigma_J(t,\omega)]^T$.

Sparse Sound Field Decomposition

The systems (4.32) and (4.34) are defined for a fixed time instant t and frequency bin ω. Nevertheless, we can exploit the group sparsity physical property of the sound field when multiple time-frequency bins are considered. In particular, when the sources are static, the vector $\mathbf{x}(t,\omega)$ will present a fixed sparsity pattern both in time and frequency and the same happens for plane waves with a subset of components actually active. Hence, the third-order tensors $\mathbf{X} \in \mathbb{C}^{N\times T\times F}$ and $\mathbf{U} \in \mathbb{C}^{I\times T\times F}$ are defined grouping $\mathbf{x}(t,\omega)$ and $\mathbf{u}(t,\omega)$, respectively and a solution to (4.32) can be found by solving the optimization problem [142]

$$\begin{aligned} &\underset{X,U}{\operatorname{argmin}} \|\mathbf{X}\|_{1,2} + \mu\|\mathbf{U}\|_{1,2}, \\ &\text{s.t. } \mathbf{y}(t,\omega) = \mathbf{D}(\omega)\mathbf{x}(t,\omega) + \mathbf{W}(\omega)\mathbf{u}(t,\omega) \end{aligned} \tag{4.35}$$

where μ is a balancing coefficient and the $\ell_{1,2}$-norm $||\cdot||_{1,2}$ is given as

$$\|\mathbf{U}\|_{1,2} = \sum_{l=1}^{I} \sqrt{\sum_{t=1}^{T}\sum_{\omega=1}^{F} |\varphi_l(t,\omega)|^2} \tag{4.36}$$

with an equivalent expression for $\|\mathbf{X}\|_{1,2}$.

When the model (4.34) is adopted, again under the assumption that sources are static within T time frames, the jth transfer function can be collected in the matrix defined as

$$\mathbf{Z}(\omega) = \mathbf{H}(\omega)\mathbf{S}(\omega) \tag{4.37}$$

where $\mathbf{S}(\omega) = [\mathbf{s}(1,\omega), \dots, \mathbf{s}(T,\omega)]$. Due to the sparsity assumption on the source distribution $Q(\boldsymbol{r})$, the number of the sources J is relatively small compared with M. Therefore, the matrix (4.37) presents a rank that approximately corresponds to J. Organizing the reverberant contributions of the sources for each time frame $\mathbf{v}(t,\omega) = \mathbf{H}(\omega)\mathbf{s}(t,\omega)$ in $\mathbf{V} \in \mathbb{C}^{M\times T\times F}$ the system in (4.34) can be solved through the optimization problem

$$\begin{aligned} &\underset{X,V}{\operatorname{argmin}} \|\mathbf{X}\|_{1,2} + \nu \sum_{\omega=1}^{F} \|\mathbf{V}(\omega)\|_\star \\ &\text{s.t. } \mathbf{y}(t,\omega) = \mathbf{D}(\omega)\mathbf{x}(t,\omega) + \mathbf{v}(t,\omega), \end{aligned} \tag{4.38}$$

with ν a balancing parameter and $\|\cdot\|_\star$ the nuclear norm [221].

It is worth noticing that the two models present some drawbacks. In particular in case of the plane wave model (4.35), an optimal number of components in $\mathbf{U}$ cannot be easily determined since it depends on the size and shape of the source region Ω beside the frequency. Moreover, the balancing parameter μ must be carefully set. As regards the transfer function model (4.38), the separation of both $\mathbf{X}$ and $\mathbf{V}$ is not trivial due to their low rank. In addition, the flexibility in (4.38) is limited since the parameter ν cannot be varied as a function of the frequency.

In [142], the authors proposed a mixed model aiming at reducing the aforementioned drawbacks. This hybrid solution combines both descriptions and through a joint optimization it aims at refining the plane wave representation (4.35) with the transfer function formulation (4.38). The mixed model is then defined as

$$\mathbf{y}(t,\omega) = \mathbf{D}(\omega)\mathbf{x}(t,\omega) + \mathbf{W}(\omega)\mathbf{u}(t,\omega) + \mathbf{v}(t,\omega). \tag{4.39}$$

Given (4.39), a solution can be found by solving the following optimization problem

$$\begin{aligned} &\underset{X,U,V}{\operatorname{argmin}} \|\mathbf{X}\|_{1,2} + \mu\|\mathbf{U}\|_{1,2} + \nu \sum_{\omega=1}^{F} \|\mathbf{V}(\omega)\|_{\star} \\ &\text{s.t. } \mathbf{y}(t,\omega) = \mathbf{D}(\omega)\mathbf{x}(t,\omega) + \mathbf{W}(\omega)\mathbf{u}(t,\omega) + \mathbf{v}(t,\omega). \end{aligned} \tag{4.40}$$

The joint optimization of the sparse matrix $\mathbf{U}$ of the plane waves and the low-rank matrix $\mathbf{V}$ provides a refined model of the reverberant component of the sound field. In [142], the optimization (4.40) is solved using the alternating direction method of multipliers [47].

4.2 Parametric Approaches

In this section, a review of the state-of-the-art parametric techniques for the reconstruction of sound fields is provided. The methods presented here, rely on a parametric model of the acoustic field that allows a compact and general description of the acoustic information. A parametric approach provides a complete and intuitive representation of the sound field that includes both the spatial recording and the manipulation of the sound field. Therefore, parametric approaches are particularly interesting in the context of EAR, where the simple static reconstruction of an acoustic scene is not sufficient. Generally, parametric techniques require two successive steps. First, the sound field is analyzed and the parameters of the model are estimated. Secondly, the desired signal is synthesized reconstructing the sound field through the model and the estimated parameters.

In the literature, a wide range of parametric models have been proposed, among which the directional audio coding (DirAC) [211, 272] and the high resolution plane wave expansion (HARPEX) [33] are well known, with extensions and developments [70, 208, 212, 213]. The final goal of these techniques is to provide the listener a perceptually convincing spatial sound. This can be obtained recreating the spatial sound features, defined as the direct and diffuse components of the sound field plus additional data such as the DoA or the location of the source. As regards DirAC, the direct signal is represented as a single plane wave for each time-frequency bin, while HARPEX employs two plane waves. The diffuse field corresponds to the reverberant part of the sound field, but it is also associated to spatially extended sources and interferers. Techniques that take into account both direct and diffuse components are known in the literature as *geometric-based* parametric model [141]. In their basic formulation, DirAC and HARPEX are able to provide the spatial sound restricted to the acquisition location.

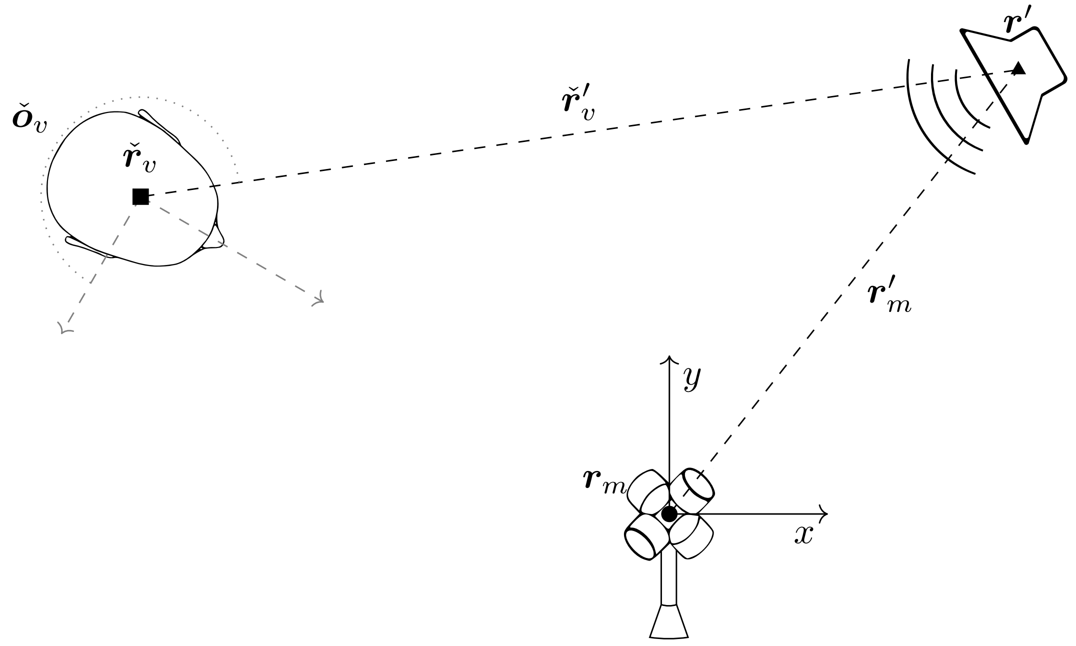

$\check{\boldsymbol{o}}_v$
$\check{\boldsymbol{r}}_v$
$\check{\boldsymbol{r}}'_v$
$\boldsymbol{r}'$
$\boldsymbol{r}'_m$
y
x
$\boldsymbol{r}_m$

Figure 4.1: *The setup adopted in [206]. A first order Ambisonics microphone located in $\boldsymbol{r}_m$ acquires the sound source in $\boldsymbol{r}'$. The listener is at the target position $\check{\boldsymbol{r}}$ with orientation $\check{\boldsymbol{o}}_v$.*

Hence, although allowing a three-degrees-of-freedom interaction, they are not suitable for a full EAR experience. In fact, in the context of EAR we are interested in a six-degrees-of-freedom interaction, which allows the listener to navigate the sound field. With the aim of allowing the navigation of the sound field, different techniques have been proposed, and in this section we review [206] and [260]. Both [206] and [260] are based on the DirAC sound field representation. In particular, [206] exploits the signals of a single first-order Ambisonics microphone and a-priori information on the source distance in order to allow the navigation of the sound field. The technique in [260], instead relies on a spatial distribution of microphone arrays in order to estimate the parameters of the sound field. Both methods are directly related to the proposed parametric sound field reconstruction technique (Chapter 5) which can be considered as an enhanced version of [260].

4.2.1 Parametric Sound Field Reconstruction with Single Array

In [206], the authors propose a parametric technique for six-degree-of-freedom sound field reconstruction, based on the signals of a first-order Ambisonics microphone. The signal at a target position, namely the listener location, is obtained from the data acquired by the microphones placed in a different recording position and the side information of the distance between the source and the sensor. We assume that the physical sources are separable by their angle with respect to the recording position.

The setup is shown in Figure. 4.1, where the location of the microphone is considered as the origin of the reference frame. An acoustic source is located in $\boldsymbol{r}'$ and the distance between the source and the microphone position $\boldsymbol{r}_m$ is given as $r'_m = \|\boldsymbol{r}'_m\| = \|\boldsymbol{r}' - \boldsymbol{r}_m\|$. In [206], r'_m is assumed to be automatically estimated, for instance with a time-of-flight camera. It follows that the DoA with respect to the recording location is

$\boldsymbol{d}'_m = \frac{\boldsymbol{r}'_m}{\|\boldsymbol{r}'_m\|}$. The target location is $\check{\boldsymbol{r}}_v$ while orientation of the listener is denoted as $\check{\boldsymbol{o}}_v$, both quantities are known a priori.

The technique in [206] works in the time-frequency domain, i.e., the microphone signals are transformed by means of the short time Fourier transform (discussed in Section 2.4). For each time-frequency bin, a single direct source is assumed to be dominant resulting in time-frequency varying DoAs $\boldsymbol{d}'_m(\omega, t) = \frac{\boldsymbol{r}'_m(\omega,t)}{\|\boldsymbol{r}'_m(\omega,t)\|}$. Since the method in [206], is based on the the DirAC parametric encoding the estimation of the diffuseness $\psi(\omega, t, \boldsymbol{r}_m)$ (4.44) and the complex spectrum $X(\omega, t, \boldsymbol{r}_m)$ are required during the analysis step. From the synthesis side, the direct and diffuse components of the signal are computed through a set of virtual loudspeakers properly defined according to the listener's location and orientation.

Analsysis and Parameter Estimation

During the analysis phase, the parameters of the DirAC encoding are estimated from the signals of a first-order Ambisonics microphone. In particular, the signals coming from the Ambisonics sensor are converted in the B-format [90] four-channels signal consisting in the omnidirectional pressure, and the three first-order gradients. As explained in [207] B-format signal can be readily derived from the coefficients of the first order spherical harmonics expansion, namely setting $I = 1$ in (3.53). The B-format data is then encoded by a DirAC encoder computing the complex spectrum, the diffuseness and the DoA of the source. While, the complex spectrum $X(\omega, t, \boldsymbol{r}_m)$ is directly derived from the omnidirectional pressure, the DoA is estimated exploiting the so called active sound intensity vector $\mathbf{i}(\omega, t, \boldsymbol{r}_m)$ that is computed as

$$\mathbf{i}(\omega, t, \boldsymbol{r}_m) = \frac{1}{2}\mathrm{Re}\left(X(\omega, t, \boldsymbol{r}_m)\mathbf{b}(\omega, t, \boldsymbol{r}_m)^*\right), \tag{4.41}$$

where Re is the function that takes only the real part of the complex spectrum and

$$\mathbf{b}(\omega, t, \boldsymbol{r}_m) = [U_X(\omega, t, \boldsymbol{r}_m), U_Y(\omega, t, \boldsymbol{r}_m), U_Z(\omega, t, \boldsymbol{r}_m)] \tag{4.42}$$

contains the first-order gradients given by the B-format signal. From (4.41) an estimate of the DoA can be computed as

$$\hat{\boldsymbol{d}}'_n(\omega, t) = -\frac{\mathbf{i}(\omega, t, \boldsymbol{r}_m)}{\|\mathbf{i}(\omega, t, \boldsymbol{r}_m)\|}. \tag{4.43}$$

The active sound intensity vector is employed also in the estimation of the diffuseness $\psi(\omega, t, \boldsymbol{r}_m)$ as [5, 70]

$$\hat{\psi}\left(\omega, t, \boldsymbol{r}_m\right) = \sqrt{1 - \frac{\|E\left\{\mathbf{i}\left(\omega, t, \boldsymbol{r}_m\right)\right\}\|}{E\left\{\|\mathbf{i}\left(\omega, t, \boldsymbol{r}_m\right)\|\right\}}} \tag{4.44}$$

where the expectation operator over time $E\{\cdot\}$ is actually approximated by moving average filtering.

Synthesis

Since we are interested in the synthesis of the signal at a location different from the recording one, a translation transformation has to be applied on the acquired signals.

Exploiting the DoA estimate $\hat{\boldsymbol{d}}'_n(\omega, t)$ and the distance value $r'_m(\omega, t)$ given as a-priori information, the actual location of the acoustic source can be estimated as

$$\hat{\boldsymbol{r}}'(\omega, t) = r'_m(\omega, t)\hat{\boldsymbol{d}}'_n(\omega, t), \tag{4.45}$$

from which the vector between the target location and the source is obtained as

$$\check{\boldsymbol{r}}'_v(\omega, t) = \hat{\boldsymbol{r}}'(\omega, t) - \check{\boldsymbol{r}}_v(t). \tag{4.46}$$

Additionally, a rotation transformation is performed in order to render the signal according to the listener orientation $\check{\boldsymbol{o}}_v(t)$

$$\check{\boldsymbol{d}}'_v(\omega, t) = \mathbf{R}\left(\check{\boldsymbol{o}}_v(t)\right) \frac{\check{\boldsymbol{r}}'_v}{\|\check{\boldsymbol{r}}'_v\|} \tag{4.47}$$

where the rotation is given by the rotation matrix $\mathbf{R}(\check{\boldsymbol{o}}_v(t))$. It is worth noting that both the orientation and location of the listener are time dependent and they have to be tracked. Finally, the DirAC decoder renders the sound field at the target location through the computation of I virtual loudspeaker signals that are located around the target position

$$Y(\omega, t, \bar{\boldsymbol{r}}_i) = Y_{\text{dir}}(\omega, t, \bar{\boldsymbol{r}}_i) + Y_{\text{diff}}(\omega, t, \bar{\boldsymbol{r}}_i) \tag{4.48}$$

where $Y_{\text{dir}}(\omega, t, \bar{\boldsymbol{r}}_i)$ and $Y_{\text{diff}}(\omega, t, \bar{\boldsymbol{r}}_i)$ represent the direct and diffuse sound, respectively, with the location of the ith virtual loudspeaker indicated as $\bar{\boldsymbol{r}}_i$. The direct component si computed as

$$Y_{\text{dir}}\left(\omega, t, \bar{\boldsymbol{r}}_i\right) = \frac{\sqrt{1 - \psi\left(\omega, t, \boldsymbol{r}_m\right)} X\left(\omega, t, \boldsymbol{r}_m\right) G\left(\check{\boldsymbol{d}}'_v(\omega, t), \bar{\boldsymbol{r}}_i\right)}{\left(\|\boldsymbol{r}'_m(\omega, t)\| \,/\, \|\hat{\boldsymbol{r}}'(\omega, t)\|\right)^{-\gamma}}. \tag{4.49}$$

where $G\left(\check{\boldsymbol{d}}'_v(\omega, t), \bar{\boldsymbol{r}}_i\right)$ is a gain function that depends on the DoA and it aims at providing the correct perception of the source DoA through the edge fading amplitude panning technique [45] , while γ is a parameters that governs the distance attenuation. As regards the diffuse component of the sound field, it is defined as [206]

$$Y_{\text{diff}}(\omega, t, \bar{\boldsymbol{r}}_i) = \sqrt{\psi(\omega, t, \boldsymbol{r}_m)} \frac{1}{\sqrt{I}} \tilde{Y}(\omega, t, \bar{\boldsymbol{r}}_i), \tag{4.50}$$

where the signal $\tilde{Y}(\omega, t, \bar{\boldsymbol{r}}_i)$ is obtained from ith decorrelated version of the pressure $X(\omega, t, \mathbf{r}_m)$. It is worth noticing that if the final target signal is reproduced through binaural synthesis, then the signals of the virtual loudspeakers (4.48) are further processed performing a convolution with the HRTF, otherwise, they are all summed in order to obtain the sound field at the listener location.

4.2.2 Parametric Sound Field Reconstruction with Distributed Arrays

In this section we review the parametric sound field reconstruction technique proposed in [260]. Differently from [206], in [260] the authors employ a set of distributed microphone arrays in order to analyze the sound field and estimate the signal of a virtual microphone (VM). The VM is characterized by an arbitrary location in the space. The direct sound field is assumed to be generated by isotropic point-like sources (IPIS), while an isotropic diffuse sound component models reverberation and noise.

Data Model and Problem Formulation

Let us consider the signal of a virtual microphone located in $\check{\boldsymbol{r}}_v$ that is composed by the superposition of a direct and a diffuse component as [260]

$$S(\omega, t, \check{\boldsymbol{r}}_v) = C_v(\omega) S_{\text{dir}}(\omega, t, \check{\boldsymbol{r}}_v) + Q_v(\omega) S_{\text{diff}}(\omega, t, \check{\boldsymbol{r}}_v), \tag{4.51}$$

where $S_{\text{dir}}(\omega, t, \check{\boldsymbol{r}}_v)$ denotes the direct sound field, $S_{\text{diff}}(\omega, t, \check{\boldsymbol{r}}_v)$ indicates the diffuse component, while $C_v(\omega) \in \mathbb{R}$ and $Q_v(\omega) \in \mathbb{R}$ represent the characteristics of the VM, i.e., its pick-up pattern $C_v(\omega)$ and sensitivity to the diffuse sound field $Q_v(\omega)$. In the case of multiple simultaneous sources, (4.51) requires the source signals to be sparse in the time-frequency domain. The direct component of (4.51) is defined as

$$S_{\text{dir}}(\omega, t, \check{\boldsymbol{r}}_v) = H_{\text{dir}}(\omega, t, \check{\boldsymbol{r}}_v, \boldsymbol{r}') S_{\text{dir}}(\omega, t, \boldsymbol{r}') \tag{4.52}$$

where $S_{\text{dir}}(\omega, t, \boldsymbol{r}')$ is the direct sound of the IPIS located in $\boldsymbol{r}'$, while the propagation from $\boldsymbol{r}'$ to $\check{\boldsymbol{r}}_v$ is given by the transfer function $H_{\text{dir}}(\omega, t, \check{\boldsymbol{r}}_v, \boldsymbol{r}')$. As regards the diffuse part of (4.51), it is defined as

$$S_{\text{diff}}(\omega, t, \check{\boldsymbol{r}}_v) = H_{\text{diff}}(\omega, t, \check{\boldsymbol{r}}_v, \boldsymbol{r}') S_{\text{diff}}(\omega, t, \boldsymbol{r}') \tag{4.53}$$

with $S_{\text{diff}}(\omega, t, \boldsymbol{r}')$ the diffuse components at the source position $\boldsymbol{r}'$ and $H_{\text{diff}}(\omega, t, \check{\boldsymbol{r}}_v, \boldsymbol{r}')$ the diffuse transfer function. Tipically, a diffuse sound field is not dependent on the location, hence $H_{\text{diff}}(\cdot)$ is not deterministic [87,260]. Let us denote with $\boldsymbol{r}_i$, $i = 1, \ldots, I$ the location of the ith microphone recording the sound field, the model for its signal follows (4.51) as

$$X(\omega, t, \boldsymbol{r}_i) = X_{\text{dir}}(\omega, t, \boldsymbol{r}_i) + X_{\text{diff}}(\omega, t, \boldsymbol{r}_v), \tag{4.54}$$

where

$$X_{\text{dir}}(\omega, t, \boldsymbol{r}_i) = H_{\text{dir}}(\omega, t, \boldsymbol{r}_i, \boldsymbol{r}') S_{\text{dir}}(\omega, t, \boldsymbol{r}'), \tag{4.55}$$

$$X_{\text{diff}}(\omega, t, \boldsymbol{r}_i) = H_{\text{diff}}(\omega, t, \boldsymbol{r}_i, \boldsymbol{r}') S_{\text{diff}}(\omega, t, \boldsymbol{r}'), \tag{4.56}$$

and the pick-up pattern and the sensitivity of the sensor are omitted since omnidirectional microphones are assumed (i.e., $C_i(\cdot) = 1$ and $Q_i(\cdot) = 1$).

Direct Signal Estimation

In order to estimate the direct component at the IPLS location, equation (4.55) can be inverted as

$$S_{\text{dir}}(\omega, t, \boldsymbol{r}') = H_{\text{dir}}(\omega, t, \boldsymbol{r}_i, \boldsymbol{r}')^{-1} X_{\text{dir}}(\omega, t, \boldsymbol{r}_i), \tag{4.57}$$

where the direct sound component $X_{\text{dir}}(\omega, t, \boldsymbol{r}_i)$ is unknown and have to be estimated. In [260], the authors adopted a Wiener-filtering approach such that

$$\hat{X}_{\text{dir}}(\omega, t, \boldsymbol{r}_i) = G_{\text{dir}}(\omega, t, \boldsymbol{r}_i) X(\omega, t, \boldsymbol{r}_i), \tag{4.58}$$

where $G_{\text{dir}}(\cdot)$ is defined as the square-root Wiener filter [241,269]

$$G_{\text{dir}}(\omega, t, \boldsymbol{r}_i) = \sqrt{1 - \frac{1}{\text{CDR}(\omega, t, \boldsymbol{r}_i) + 1}} = \sqrt{1 - \psi(\omega, t, \boldsymbol{r}_i)} \tag{4.59}$$

with $\psi(\cdot)$ the diffuseness as in (4.44). The authors in [260] adopt the square-root Wiener filter due to its property of preserving the correct direct sound power in the estimate $\hat{X}_{\text{dir}}(\cdot)$. With $\text{CDR}(\omega, t, \boldsymbol{r}_i)$ the authors refer to the coherence-to-diffuse power ratio, defined as

$$\text{CDR}\left(\omega, t, \boldsymbol{r}_i\right) = \frac{E\left\{X_{\text{dir}}\left(\omega, t, \boldsymbol{r}_i\right) X_{\text{dir}}^{*}\left(\omega, t, \boldsymbol{r}_i\right)\right\}}{E\left\{X_{\text{diff}}\left(\omega, t, \boldsymbol{r}_i\right) X_{\text{diff}}^{*}\left(\omega, t, \boldsymbol{r}_i\right)\right\}} = \frac{\Phi_{\text{dir},ii}(t, \omega)}{\Phi_{\text{diff},ii}(t, \omega)}. \tag{4.60}$$

In the literature, the quantity in (4.60) is also known as signal-to-diffuse power ratio.

Diffuse Signal Estimation

As far as the diffuse signal estimation at the IPIS location is concerned, a complementary approach with respect to the direct component estimate is adopted.

In particular, the square-root Wiener filter for the estimation of the ith microphone diffuse component is defined as

$$G_{\text{diff}}\left(\omega, t, \boldsymbol{r}_i\right) = \sqrt{1 - \left|G_{\text{dir}}\left(\omega, t, \boldsymbol{r}_i\right)\right|^2} \tag{4.61}$$

with $G_{\text{dir}}(\cdot)$ given by (4.59). It follows that an estimate of the diffuse component at the ith microphone can be obtain applying (4.61) to (4.56) as

$$\hat{X}_{\text{diff}}(\omega, t, \boldsymbol{r}_i) = G_{\text{diff}}(\omega, t, \boldsymbol{r}_i) X(\omega, t, \boldsymbol{r}_i). \tag{4.62}$$

Finally, the diffuse signal estimation at the IPIS can be derived by inverting (4.56)

$$S_{\text{diff}}(\omega, t, \boldsymbol{r}') = H_{\text{diff}}(\omega, t, \boldsymbol{r}_i, \boldsymbol{r}')^{-1} X_{\text{diff}}(\omega, t, \boldsymbol{r}_i). \tag{4.63}$$

Analysis and Sound Field Parameters Estimation

Adopting the direct and diffuse estimation strategies of Section 4.2.2 we are required to determine the source position $\boldsymbol{r}'(\omega, t)$ and the CDR at microphones for each time-frequency bin.

Source Position Estimation Thanks to the adoption of more than one microphone array distributed in space, the location $\boldsymbol{r}'(\omega, t)$ of the source can be obtained triangulating the different DoAs. In particular, in [260], the authors consider a setup of distributed circular microphone arrays, for which the DoAs are obtained using the technique introduced in Section 4.2.1. Therefore, for each ath array, $a = 1, \dots, A$, the DoA is computed using the active sound intensity vector (4.41)

$$\hat{\boldsymbol{d}}^{(a)}(\omega, t) = -\frac{\mathbf{i}(\omega, t, \boldsymbol{r}_a)}{\|\mathbf{i}(\omega, t, \boldsymbol{r}_a)\|}. \tag{4.64}$$

where $\boldsymbol{r}_a$ is array center. The source location is then estimated from the A DoAs minimizing

$$\hat{\boldsymbol{r}}'(\omega, t) = \text{argmin}_{\boldsymbol{c}} \sum_{a=1}^{A} \left\|J\left(\omega, t, \boldsymbol{r}_a, \boldsymbol{c}\right)\right\|^2 \tag{4.65}$$

where

$$J(\cdot) = \left[\mathbf{I} - \left(\hat{\boldsymbol{d}}^{(a)}(\omega, t)\right)^T \hat{\boldsymbol{d}}^{(a)}(\omega, t)\right] (\boldsymbol{c} - \boldsymbol{r}_a) \tag{4.66}$$

with $\mathbf{I}$ the identity matrix and $\boldsymbol{c}$ the line defined by the ath array location and its relative DoA. In practice, the intersection is found minimizing the squared distance between the lines.

Coherence-to-Diffuse Ratio Estimation The CDR can be readily derived from the estimated diffuseness (4.44) as [71]

$$\begin{aligned}\widehat{\mathrm{CDR}}\left(\omega, t, \boldsymbol{r}_a\right) &= \frac{1}{\hat{\psi}\left(\omega, t, \boldsymbol{r}_a\right)} - 1 \\ &= \frac{1}{\sqrt{1 - \frac{|E\{\mathbf{i}(\omega,t,\boldsymbol{v}_a)\}\|}{E\{\|\mathbf{i}(\omega,t,\boldsymbol{r})\|\}}} - 1} \\ &= \sqrt{\frac{E\left\{\|\mathbf{i}\left(\omega, t, \boldsymbol{r}_a\right)\|\right\}}{E\left\{\|\mathbf{i}\left(\omega, t, \boldsymbol{r}_a\right)\|\right\} - \left\|E\left\{\mathbf{i}\left(\omega, t, \boldsymbol{r}_a\right)\right\}\right\|} - 1}.\end{aligned} \tag{4.67}$$

Once the CDR have been estimated, the square-root Wiener filters (4.59) and (4.61) can be computed and the direct and diffuse components at the microphone array are obtained as

$$\hat{X}_{\text{dir}}(\omega, t, \boldsymbol{r}_a) = G_{\text{dir}}(\omega, t, \boldsymbol{r}_a) X(\omega, t, \boldsymbol{r}_a), \tag{4.68}$$

$$\hat{X}_{\text{diff}}(\omega, t, \boldsymbol{r}_a) = G_{\text{diff}}(\omega, t, \boldsymbol{r}_a) X(\omega, t, \boldsymbol{r}_a), \tag{4.69}$$

where the input $X(\omega, t, \boldsymbol{r}_a)$ is the given by the omnidirectional signal of the B-format encoding (see Section 4.2.1).

Synthesis

The signal of the VM is synthesized applying the model (4.51) by superimposing the direct sound component (4.52) and the diffuse signal component (4.53) computed at the target location. It follows that appropriate models for the direct and diffuse transfer functions are required.

As regards the direct signal S_{dir}, in [260] the authors adopt the Green's function (3.29) as transfer function

$$H_{\text{dir}}\left(\omega, t, \boldsymbol{r}_a, \boldsymbol{r}'\right) = \frac{e^{j\frac{\omega}{c}\|\boldsymbol{r}_a - \boldsymbol{r}'(\omega,t)\|}}{\|\boldsymbol{r}_a - \boldsymbol{r}'(\omega, t)\|}, \tag{4.70}$$

from which the direct component at the source location is obtained as

$$\hat{S}_{\text{dir}}(\omega, t, \boldsymbol{r}') = H_{\text{dir}}(\omega, t, \boldsymbol{r}_a, \boldsymbol{r}')^{-1} \hat{X}_{\text{dir}}(\omega, t, \boldsymbol{r}_a). \tag{4.71}$$

In [260], the homogeneous assumption of the diffuse sound field is exploited in order to derive the transfer function of the diffuse component as $H_{\text{diff}}(\omega, t, \boldsymbol{r}_a, \boldsymbol{r}') = 1$. Hence, the diffuse sound at the IPSI simplifies to

$$\hat{S}_{\text{diff}}(\omega, t, \boldsymbol{r}') = \hat{S}_{\text{diff}}(\omega, t, \boldsymbol{r}_a). \tag{4.72}$$

Finally, the signal is synthesized exploiting the estimated quantities in the model (4.51), where the pick-up pattern $C_v(\cdot)$ and the sensitivity $Q(\cdot)$ of the VM can be arbitrarily designed by the final user, reproducing physical or even non-physical microphone characteristics.

CHAPTER 5

Parametric Sound Field Reconstruction with Directional Sources

In this chapter we introduce two techniques for the reconstruction of the sound field adopting a parametric model. The main characteristic of these novel methods lies in the possibility of modeling the acoustic source directivity. In fact, the state-of-the-art parametric techniques, reviewed in Section 4.2, do not consider the directional behavior of sound sources. Nevertheless, it is known that in general acoustic emitters present a directional emission of sound energy that characterize their radiation and interaction with the environment. Therefore, in order to properly reconstruct the sound field with directional sources, extended parametric models are required. The techniques presented in this chapter explicitly model the source directivity and allow an improved reconstruction in terms of the spatial cues of sound fields. The possibility of modeling source directivity is particularly appealing in the context of EAR, where not only an accurate reconstruction of actual acoustic field is desired, but also the integration of virtual sources, that could have directional characteristics, is required. The developed techniques follow a *divide et impera* approach, thus in Section 5.2, we first introduce a methodology for sound field reconstruction with directional sources in free-field and in Section 5.3 the approach is extended to reverberant environments. For both the techniques, the overall system is a combination of individual sub-systems. This structure brings us two main advantages; on the one hand we can think of distributing part of the computational load locally to each array (e.g., the estimation of the DoAs, the estimation of the direct and diffuse components); on the other hand such a structure allows us to possibly substitute sub-systems depending on the application scenario with the only constraint of maintaining the same input/output relationship. Furthermore, this structure gives us a wider insight into the system behaviour and promotes the model interpretability.

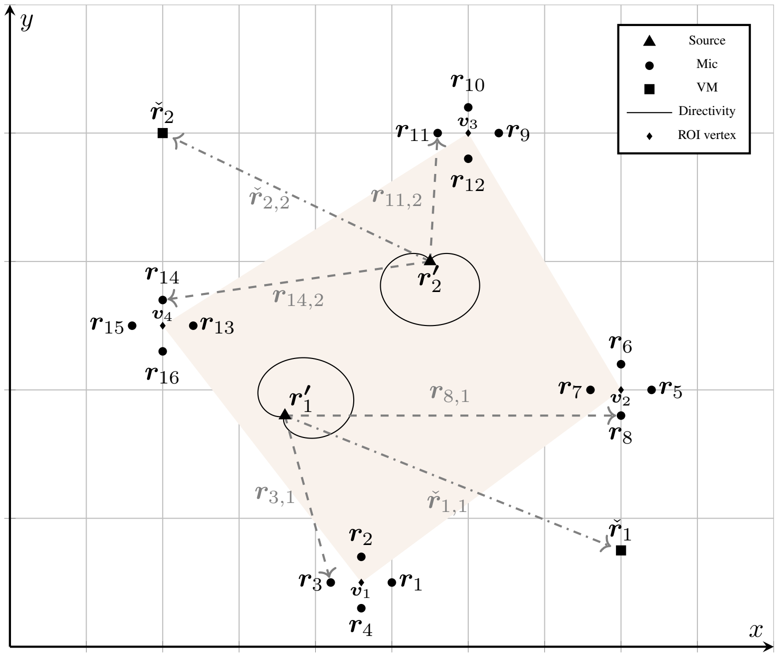


Figure 5.1: *2D graphical representation of the model. The setup is presented with $A = 4$ circular microphone arrays of $M = 4$ microphones each. Two sources ($N = 2$) and two VMs ($V = 2$) are present in the scene. The directivity function of the source is superimposed on the plot of the scene.*

5.1 Parametric Sound Field Reconstruction Model

In this section we introduce the general data model of the proposed parametric sound field reconstruction techniques. This model can be adapted according to the environment under analysis. In particular, Section 5.2 considers a controlled acoustic environment, while Section 5.3 exploits the general model considering also reverberation. We tackle the sound field reconstruction as a virtual miking problem. Hence, our goal is the estimation of the signal of a virtual microphone (VM).

5.1.1 Data Model

Given a Cartesian coordinate system, let us consider N acoustic sources, placed in arbitrary locations $\boldsymbol{r}'_n = [x'_n, y'_n, z'_n]^T$, $n = 1, \dots, N$; a network of $A \geq 2$ distributed compact microphone arrays with M microphones each, located at $\boldsymbol{r}_i = [x_i, y_i, z_i]^T$, $i = 1, \dots, M \times A$; and a set of V VMs positioned in $\check{\boldsymbol{r}}_v = [\check{x}_v, \check{y}_v, \check{z}_v]^T$, $v = 1, \dots, V$, as shown in Figure 5.1.

We assume that sources, microphones and VMs lie on the same plane. Moreover, we define the Region Of Interest (ROI) where the sources lie as the polygonal region $\mathcal{R} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_A)$, where $\boldsymbol{v}_a$ is the ath vertex of the polygon. The vertices are defined as the centroids of each array, i.e.,

$$\boldsymbol{v}_a = \frac{1}{M} \sum_{i=(a-1)M+1}^{aM} \boldsymbol{r}_i \ , \qquad a = 1, \dots, A. \tag{5.1}$$

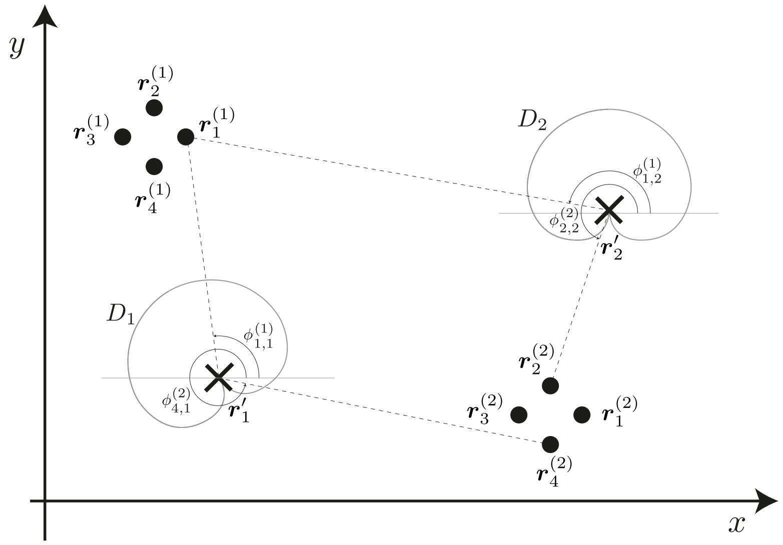


Figure 5.2: *Graphical representation of the analysis model with two acoustic sources placed in $\boldsymbol{r}_1'$ and $\boldsymbol{r}_2'$, respectively. The sound field is captured by $A = 2$ microphone array each of them constituted by $M = 4$ omnidirectional microphones.*

When only two arrays are present in the acoustic scene, the definition of the ROI degenerates since we have a polygonal region $\mathcal{R}$ with only two vertices. In this case, we consider the ROI as the whole plane where sources, microphones and VMs lie. In the most general scenario, we model the signal of the vth directional VM in $\check{\boldsymbol{r}}_v$ in the time-frequency domain as the linear combination of a direct sound component and a diffuse sound component [260], i.e.,

$$\begin{aligned} S(t,\omega,\check{\boldsymbol{r}}_v) = C_v(\omega) S_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v) \\ + Q_v(\omega) S_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v),\, n \in \{1,\dots,N\}, \end{aligned} \tag{5.2}$$

where t is the time-frame index, $\omega = 2\pi f$ the radial frequency with $f > 0$ the temporal frequency, $C_v(\omega) \in \mathbb{R}$ models the VM microphone pick-up pattern and $Q_v(\omega) \in \mathbb{R}$ its sensitivity to the diffuse field. The model in (5.2) is valid under the assumption that the N source signals are sufficiently sparse in the time-frequency domain [260, 272]. More precisely, when multiple sources are simultaneously active, their signal content in the frequency domain must not overlap significantly, i.e. one source is dominant at each time-frequency bin. In the next sections, we specify the VM signal model according to the acoustic characteristics of the environment describing the parameters of the signal. First Section 5.2 focuses on free field scenarios, while Section 5.3 adopts the full model of (5.2).

5.2 Parametric Sound Field Reconstruction in the Free Field

5.2.1 Signal Model and Problem Formulation

Here, we specify the signal model (5.2) in a free field scenario. In this scenario, we can express the signal at the vth VM signal located at $\check{\boldsymbol{r}}_v$ as

$$S\left(\omega, t, \check{\boldsymbol{r}}_v\right) = S_{\text{dir}}\left(\omega, t, \check{\boldsymbol{r}}_v\right) = \sum_{n=1}^{N} C_v(\omega) D_n\left(\omega, t, \check{\phi}_{v,n}\right) H\left(\omega, \check{\boldsymbol{r}}_v, \boldsymbol{r}'_n\right) S\left(\omega, t, \boldsymbol{r}'_n\right) \tag{5.3}$$

where N sound sources are present in the scene with signal $S\left(\omega, t, \boldsymbol{r}'_n\right)$ and location $\boldsymbol{r}'_n$. The directivity (or radiation) pattern of the nth acoustic source is denoted by $D_n\left(\omega, t, \check{\phi}_{v,n}\right) : \mathbb{R}^3 \to (0, 1)$. This function defines the directional energy emission pattern of the source as a function of the angle $\check{\phi}_{v,n} = \angle(\check{\boldsymbol{r}}_v - \boldsymbol{r}')$ (see Figure 5.2) and frequency ω. Note that since in Section 5.1.1 we assume that sources and arrays lie on the same plane, $D_n(\cdot)$ in (5.3) -depends on the azimuth angle $\check{\phi}_{v,n}$ only. With $C_v(\omega)$ we denote the VM pick-up pattern, while the free-field transfer function $H\left(\omega, \check{r}_v, \boldsymbol{r}'_n\right)$ is defined as the Green's function (3.29)

$$H\left(\omega, \check{\boldsymbol{r}}_v, \boldsymbol{r}'_n\right) = \frac{e^{j\frac{\omega}{c}\|\check{\boldsymbol{r}}_v - \boldsymbol{r}'_n\|}}{\|\check{\boldsymbol{r}}_v - \boldsymbol{r}'_n\|}. \tag{5.4}$$

The setup adopted in order to analyze the sound field is composed of a set of A distributed compact arrays with M microphone each, as shown in Figure 5.2. The signal at the mth omnidirectional microphone of the ath array can be written as

$$\begin{aligned} X\left(\omega, t, \boldsymbol{r}_m^{(a)}\right) = &\sum_{n=1}^{N} D_n\left(\omega, t, \phi_{m,n}^{(a)}\right) H\left(\omega, \boldsymbol{r}_m^{(a)}, \boldsymbol{r}'_n\right) S\left(\omega, t, \boldsymbol{r}'_n\right) \\ &+ N\left(\omega, t, \boldsymbol{r}_m^{(a)}\right) \end{aligned} \tag{5.5}$$

where $\phi_{m,n}^{(a)} = \angle(\boldsymbol{r}_m^{(a)} - \boldsymbol{r}')$ is the angle between the nth source and the microphone and $N\left(\omega, t, \boldsymbol{r}_m^{(a)}\right)$ models the sensor self-noise. We remark that for omnidirectional microphones the pick-up pattern $C(\cdot)$ is equal to 1 and it is omitted in (5.5). Let us describe in matrix form the signal model (5.5) for each array

$$\mathbf{x}^{(a)} = \left[\mathbf{D}^{(a)} \otimes \mathbf{H}^{(a)}\right] \mathbf{s} + \mathbf{n}^{(a)}, \quad \forall a = 1, \dots, A, \tag{5.6}$$

where

$$\begin{aligned} \mathbf{x}^{(a)} &= \left[X\left(\boldsymbol{r}_1^{(a)}\right), \dots, X\left(\boldsymbol{r}_M^{(a)}\right)\right]^T, \\ \mathbf{s} &= \left[S\left(\boldsymbol{r}'_1\right), \dots, S\left(\boldsymbol{r}'_N\right)\right]^T, \\ \mathbf{n}^{(a)} &= \left[N\left(\boldsymbol{r}_1^{(a)}\right), \dots, N\left(\boldsymbol{r}_M^{(a)}\right)\right]^T, \\ \left[\mathbf{D}^{(a)}\right]_{m,n} &= D_n\left(\phi_{m,n}^{(a)}\right), \\ \left[\mathbf{H}^{(a)}\right]_{m,n} &= H\left(\boldsymbol{r}_m^{(a)}, \boldsymbol{r}'_n\right), \end{aligned} \tag{5.7}$$

with $\otimes$ the Hadamard product and a the array index. Note that, for the sake of compactness, we omitted the dependency on frequency ω and time t.

5.2.2 Sound Field Analysis and Parameter Estimation

According to the assumed parametric model (5.3), in order to compute the VM signal, we need to estimate three parameters for each source: its location $\boldsymbol{r}'_n$, its directivity pattern $D_n(\cdot)$ and the emitted signal $S(\omega, t, \boldsymbol{r}'_n)$.

Source Localization

Assuming that the microphones locations are known, the transfer function model (5.4) requires the localization of multiple sources in the acoustic scene. In this paper we introduce the Distributed Ray Space Transform (DRST) as a tool for source localization. Recently, the Ray Space Transform (RST) [37], reviewed in Section 6.1.1, has proved to be a convenient tool for mapping the acoustic information in the ray space. In this domain, the main acoustic primitives (sources, array and reflectors) are mapped onto lines due to the fact that each point in the ray space corresponds to an acoustic ray in the geometric space, and the line parameters are uniquely related to the source location (see Section 6.1.1). As a consequence, we can easily localize multiple sources in the ray space through a pattern analysis approach.

The DRST inherits the core idea of the RST, but differently from the RST, which is based on multiple beamforming operations performed on sub-arrays of a single extended linear array, the DRST carries out the beamforming independently for each array within a network of distributed compact arrays. Morever, DRST maps acoustic information in the Projective Ray Space (PRS) [161]. Similarly to the parametric technique in [260] and discussed in Section 4.2.2, the localization is performed with two consecutive steps: the estimation of the DoA at each array and the triangulation of the directions in order to find the source position. In this case, we adopt a spatial filtering approach for the DoA identification, employing a Delay and Sum beamforming operation [268]. Successively, the DoAs are triangulated in the PRS domain. The beamforming operation estimates the directional energy distribution at the center of mass $\boldsymbol{v}_a = [x_a, y_a]^T$ of each microphone array. The pseudospectrum $\lambda^{(a)}(\alpha, t, \omega)$ of the ath array can be computed as the absolute value of the beamformer output for all the possible directions $\alpha \in (0, 2\pi]$. As we are interested only in localizing the acoustic sources, similarly to [27, 37, 160] we compute a wideband extension of the pseudospectra by averaging $\lambda^{(a)}(\alpha, t, \omega)$ through

$$\bar{\lambda}^{(a)}(t, \alpha) = \left\{ \prod_{w=1}^{W/2} \lambda^{(a)}(\alpha, t, \omega_w) \right\}^{\frac{2}{W}}, \tag{5.8}$$

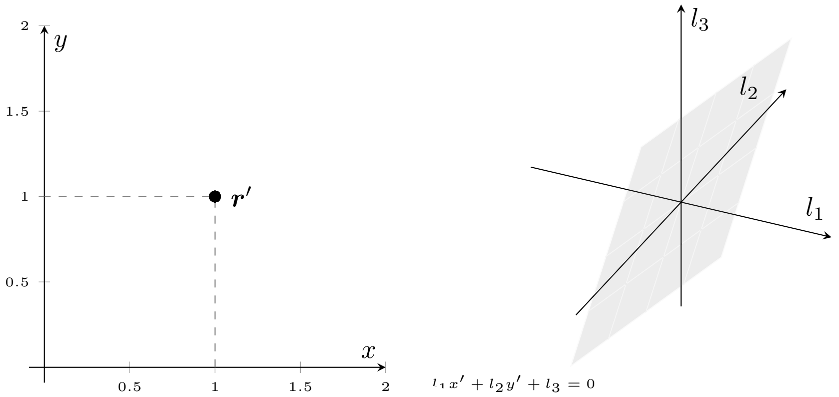
where W is the number of points on the discretized frequency axis, gaining robustness against spatial aliasing and frequency bands with low SNR. It is worth noticing that (5.8) is not the only possible choice to obtain a wideband pseudospectrum. However, as investigated in [27], this choice provides a wideband pseudospectrum with greater resolution, narrower main-lobe and side-lobes attenuation with respect to the one obtained with the arithmetic mean. The DoAs of the sources can be measured as the directions $\bar{\alpha}_n^{(a)}(t)$, $n = 1, \ldots, N$ corresponding to the N highest peaks of $\bar{\lambda}^{(a)}(t, \alpha)$, i.e., that maximizes the pseudospectrum $\lambda^{(a)}(\alpha, t, \omega)$

$$\bar{\boldsymbol{\alpha}}^{(a)}(t) = \mathcal{D}(\bar{\lambda}^{(a)}(\alpha, t), N), \tag{5.9}$$

where $\mathcal{D}(\cdot, N)$ is the operator that returns the N highest peaks and

$$\bar{\boldsymbol{\alpha}}^{(a)}(t) = \left[\bar{\alpha}_1^{(a)}(t), \bar{\alpha}_2^{(a)}(t), \ldots, \bar{\alpha}_N^{(a)}(t) \right]^T \tag{5.10}$$

is the $N \times 1$ vector of the estimated DoAs for the ath array.



(a)

(b)

Figure 5.3: *A point-like source in the geometric space (a) is mapped, in the projective ray space (b), onto a plane with normal direction* $\bar{\boldsymbol{r}}' = [x', y', 1]^T$

Once the DoAs for each array have been estimated, the source locations in Cartesian coordinates are estimated through triangulation. In a multi-source scenario the problem of DoA disambiguation arises, i.e. the matching of the DoAs measured from different arrays corresponding to the same source. Here, we tackle the disambiguation and triangulation problems employing a localization method based on the Distributed Ray Space Transform (DRST).

We introduce the DRST, as a tool devoted to the mapping of the signals of distributed arrays onto the PRS domain [161]. The PRS, derived as a generalization of the ray space [160], is the domain of representation of the sound field in the scenarios where the same acoustic scene is observed by multiple viewpoints. This parameterization is based on a generalization of the ray space reviewed in Section 6.1.1. In fact, the PRS is defined by the parameters of the implicit equation $l_1x + l_2y + l_3 = 0$ that identifies an acoustic ray in 2D, rather than the euclidean equation of the RST (6.1). The distinctive characteristics of such parameterization, is that the acoustic primitives, such as sources and reflectors, are mapped in the PRS onto linear subspaces or combinations thereof.

Let us consider a generic point-like acoustic source at a given time instant t placed in $\boldsymbol{r}'(t) = [x'(t), y'(t)]^T$ as in Figure 5.3(a). This can be seen as the point of origin of acoustic rays and, therefore it is represented in the ray space by all the rays crossing it. It is worth to underline that since sources, microphones and VMs are assumed to be lying on the same plane (see Section 5.1.1), we can reduce the location of the elements to a point in 2D as depicted in Figure 5.3(a). It follows that given the source location in homogeneous coordinates $\bar{\boldsymbol{r}}'(t) = [x'(t), y'(t), 1]^T$, a ray emitted by the source satisfies

$$\mathbf{l}^T \bar{\boldsymbol{r}}'(t) = 0, \tag{5.11}$$

where $\mathbf{l} = \varepsilon[l_1, l_2, l_3]^T, \varepsilon \neq 0$ are the parameters of the projective line describing the ray. As described in [161], the representation of $\boldsymbol{r}'(t)$ is given by the set of rays passing through it, and in the PRS corresponds to a plane (see Figure 5.3(b)). DoAs in $\bar{\boldsymbol{\alpha}}^{(a)}(t)$

in (5.9) are converted in acoustic rays in the PRS through

$$\begin{aligned} l_{1,n}^{(a)}(t) &= \varepsilon \sin\left(\bar{\alpha}_n^{(a)}(t)\right);\\ l_{2,n}^{(a)}(t) &= \varepsilon \cos\left(\bar{\alpha}_n^{(a)}(t)\right);\\ l_{3,n}^{(a)}(t) &= \varepsilon[y_a \cos\left(\bar{\alpha}_n^{(a)}(t)\right) - x_a \sin\left(\bar{\alpha}_n^{(a)}(t)\right)], \varepsilon > 0. \end{aligned} \tag{5.12}$$

These points will form clusters in the PRS on the planes representing the acoustic sources.

In order to associate DoAs to sources and then proceed to the localization task, we adopt techniques of pattern analysis. More precisely, we use a RANSAC algorithm [99] (RANdom SAmple Consensus) over the set of $N \times A$ points in the PRS. Let us indicate with the subscript $\hat{n}$ the points identified by RANSAC as pertaining to the nth source

$$\hat{\mathbf{l}}_{\hat{n}}(t) = [l_{1,\hat{n}}(t), l_{2,\hat{n}}(t), l_{3,\hat{n}}(t)]^T. \tag{5.13}$$

Note that $\hat{\mathbf{l}}_{\hat{n}}$ no longer depends on the array index a. The points in (5.13) are then re-arranged in matrix form such that the condition of (5.11) becomes

$$\mathbf{L}_n(t)\bar{\boldsymbol{r}}'_n(t) = 0, \tag{5.14}$$

where $\mathbf{L}_n(t) = [\hat{\mathbf{l}}_1(t), \dots, \hat{\mathbf{l}}_{\hat{N}}(t)]^T$ and $\hat{N}$ is the number of rays related to the nth source. From (5.14), we can infer that in order to cross the source location point $\bar{\boldsymbol{r}}'_n$, the rays should belong to the null space of $\mathbf{L}_n$. Therefore, we compute the singular value decomposition of the matrix $\mathbf{L}_n$

$$\mathbf{L}_n(t)\left(\mathbf{L}_n(t)\right)^T = \mathbf{U}_{\mathbf{L},n}(t)\boldsymbol{\Lambda}_{\mathbf{L},n}(t)\mathbf{V}_{\mathbf{L},n}^T(t) \tag{5.15}$$

where $\mathbf{U}_{\mathbf{L},n}(t)$ and $\mathbf{V}_{\mathbf{L},n}^T(t)$ are the singular vectors matrices of the decomposition and $\boldsymbol{\Lambda}_{\mathbf{L},n}(t)$ is the diagonal matrix of the singular values. Finally, the estimate of the source location $\hat{\boldsymbol{r}}'_n(t)$ is obtained as [161]

$$\hat{\boldsymbol{r}}'_n(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{v}_n(t), \tag{5.16}$$

where $\mathbf{v}_n(t)$ is the singular vector of $\mathbf{V}_{\mathbf{L},n}^T(t)$ associated to the smallest singular value in $\boldsymbol{\Lambda}_{\mathbf{L},n}(t)$ given by the singular value decomposition (5.15). It is worth noticing that, in the present section, we made explicit the dependence of the nth source position from the time instant t in order to show that the presented method can account for moving sources. However, for the sake of readability, we omit such an explicit dependence in the following sections.

Directivity Pattern Estimation

Let us assume the distance between sources and arrays being much greater than the arrays size. Therefore, in (5.6) we can assume $\phi_{m,n}^{(a)} \approx \phi_n^{(a)} = \angle(\boldsymbol{r}^{(a)} - \boldsymbol{r}'_n)$ for each microphone m in the ath array. Similarly to estimate of the source location, we will omit the time dependence of the directivity pattern $\mathbf{D}(\cdot)$. In addition, also the frequency dependence is omitted from hereinafter for the sake of readability. With this assumption, the matrix $\mathbf{D}^{(a)}$ becomes $\left[\mathbf{D}^{(a)}\right]_{m,n} = D_n(\phi_n^{(a)})$. Hence, defining $\mathbf{P}^{(a)} = \text{diag}\left(D_0(\phi_0^{(a)}), \ldots, D_{N-1}(\phi_{N-1}^{(a)})\right)$, the model in (5.6) reduces to

$$\mathbf{x}^{(a)} = \mathbf{H}^{(a)}(\mathbf{P}^{(a)}\mathbf{s}) + \mathbf{e}^{(a)}, \quad \forall a = 0, \ldots, A-1. \tag{5.17}$$

After source localization, we have an estimate $\widehat{\mathbf{H}}^{(a)}$ of $\mathbf{H}^{(a)}$, thus we can estimate the vector $\mathbf{b}^{(a)} = \mathbf{P}^{(a)}\mathbf{s}$ through an LCMV beamformer [268] defined by the optimization problem

$$\mathbf{g}_n^{(a)} = \underset{\mathbf{g}}{\arg\min}\ \mathbf{g}^H\mathbf{g} \qquad \text{s.t. } \mathbf{g}^H\widehat{\mathbf{H}}^{(a)} = \mathbf{c}_n, \tag{5.18}$$

where $\mathbf{c}_n \in \mathbb{C}^{1\times N}$ with $[\mathbf{c}_n]_i = 1$ for $i = n$ and zero otherwise. The solution to the optimization problem given by [252] is

$$\mathbf{g}_n^{(a)} = \widehat{\mathbf{H}}^{(a)}\left(\widehat{\mathbf{H}}^{(a)^H}\widehat{\mathbf{H}}^{(a)}\right)^{-1}\mathbf{c}_n^H. \tag{5.19}$$

Therefore, an estimate of $\mathbf{b}^{(a)}$ is computed as

$$\hat{\mathbf{b}}^{(a)} = \mathbf{G}^{(a)^H}\mathbf{x}^{(a)} \quad \forall a = 1, \ldots, A, \tag{5.20}$$

where $\mathbf{G}^{(a)} = \left[\mathbf{g}_1^{(a)}, \ldots, \mathbf{g}_N^{(a)}\right]$. Let us define the vector $\mathbf{q}_n$

$$\begin{aligned}\mathbf{q}_n &= \left[|\left[\hat{\mathbf{b}}^{(1)}\right]_n|, \ldots, |\left[\hat{\mathbf{b}}^{(A)}\right]_n|\right]^T \\ &= |\hat{S}(\boldsymbol{r}'_n)|\left[\hat{D}_n(\phi_n^{(1)}), \ldots, \hat{D}_n(\phi_n^{(A)})\right]^T,\end{aligned} \tag{5.21}$$

that represents an estimate of the nth source directivity pattern for the directions $\phi_n^{(a)}$ $\forall a = 1, \ldots, A$ scaled by the factor $|\hat{S}(\boldsymbol{r}'_n)|$. Since VMs can be arbitrarily placed in space, we need to reconstruct $D_n(\cdot)$ under any possibile direction $\check{\phi}_{v,n}$. Hence, a circular harmonics model for the interpolation of the directivity pattern $D_n(\cdot)$ is adopted

$$D_n(\phi_n^{(a)}) = \sum_{l=0}^{I-1} w_{n,l}\cos(l\phi_n^{(a)}) + r_{n,l}\sin(l\phi_n^{(a)}). \tag{5.22}$$

where the coefficients $w_{n,l}$ and $r_{n,l}$ parametrize the directivity pattern enabling the computation for arbitrary angles. In order to estimate the coefficients in (5.22), we define the matrix

$$\mathbf{A}_n = [\mathbf{A}_{n,1}, \mathbf{A}_{n,2}] \tag{5.23}$$

where $[\mathbf{A}_{n,1}]_{a,l} = \cos\left(l\phi_n^{(a)}\right)$ and $[\mathbf{A}_{n,2}]_{a,l} = \sin\left(l\phi_n^{(a)}\right)$ and the vector of coefficients

$$\mathbf{y}_n = [w_{n,0}, \ldots, w_{n,I-1}, r_{n,0}, \ldots, r_{n,I-1}]^T. \tag{5.24}$$

An estimate of the coefficients $\mathbf{y}_n$ can be found by solving the optimization problem [48]

$$\hat{\mathbf{y}}_n = \underset{\mathbf{y}_n}{\arg\min}\ \|\mathbf{q}_n - \mathbf{A}_n\mathbf{y}_n\|^2 \qquad \text{s.t. } \mathbf{F}\mathbf{y}_n \geq \mathbf{0}, \tag{5.25}$$

where $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2] \in \mathbb{D}^{I\times 2I}$, $[\,\mathbf{F}_1]_{i,l} = \cos(l\phi_i)$, $[\,\mathbf{F}_2]_{i,l} = \sin(l\phi_i)$ with $\phi_i \in [0, 2\pi)$ $i = 0, \ldots, I-1$. It is worth noting that using the result of (5.25) in (5.22) we will obtain a scaled version of the radiance pattern $\hat{\mathbf{D}}_n$ due to the presence of the source signal magnitude spectrum as shown in (5.21). In practice, the scaling is removed by normalizing the estimates $\hat{\mathbf{D}}_n$ in the range $(0, 1)$.

Source Signal Estimation

The estimation of the source signal $S(\omega, t, \boldsymbol{r}'_n)$ takes advantage of the previously estimated parameters. Specifically, an informed spatial filter is designed for the nth source as

$$\mathbf{p}_n^\star = \arg\min_{\mathbf{p}} \ \mathbf{p}^H\mathbf{p} \qquad \text{s.t. } \mathbf{p}^H\mathbf{Y} = \mathbf{d}_n, \tag{5.26}$$

where

$$\mathbf{Y} = \left[\left(\widehat{\mathbf{H}}^{(1)} \otimes \hat{\mathbf{D}}^{(1)}\right)^T, \ldots, \left(\widehat{\mathbf{H}}^{(A)} \otimes \hat{\mathbf{D}}^{(A)}\right)^T\right]^T \in \mathbb{C}^{AM\times N} \tag{5.27}$$

and $\mathbf{d} \in \mathbb{C}^{1\times N}$ with $[\mathbf{d}]_i = 1$ if $i = n$ and zero otherwise. Note that $\hat{\mathbf{D}}^{(a)}$ contains the directivity values obtained using (5.22) with the estimated coefficients (ref. (5.25)). Finally, we filter the microphones signals with $\mathbf{p}_n^\star$ in order to obtain the estimate $\hat{S}(\boldsymbol{r}'_n)$ of $S(\boldsymbol{r}'_n)$

$$\hat{S}\left(\boldsymbol{r}'_n\right) = \left(\mathbf{p}_n^\star\right)^H \mathbf{x} \tag{5.28}$$

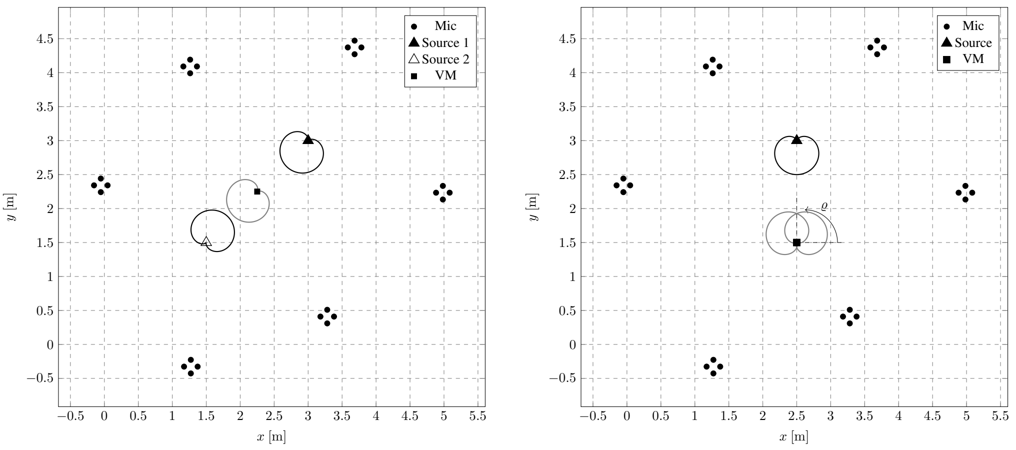
where $\mathbf{x} = \left[\left(\mathbf{x}^{(1)}\right)^T, \ldots, \left(\mathbf{x}^{(A)}\right)^T\right]^T$. We remark that in the current and previous sections we omitted the dependency on frequency ω time t of the different quantities for the sake of readability.

5.2.3 Synthesis

Once the sound field analysis is completed as described in Section 5.2.2, the estimated parameters are used to compute the signal of the vth virtual microphone as

$$\hat{S}\left(\omega, t, \check{\boldsymbol{r}}_v\right) = \sum_{n=1}^{N} C_v(\omega)\hat{D}_n\left(\omega, t, \hat{\phi}_{v,n}\right) H\left(\omega, \check{\boldsymbol{r}}_v, \hat{\boldsymbol{r}}'_n\right) \hat{S}\left(\omega, t, \hat{\boldsymbol{r}}'_n\right) \tag{5.29}$$

Note that $H(\check{\boldsymbol{r}}_v, \hat{\boldsymbol{r}}'_n, \omega)$ is proportional to $1/\|\check{\boldsymbol{r}}_v - \hat{\boldsymbol{r}}'_n\|$, thus in practice, it is limited to the maximum value $H_{\max}$ to avoid an excessive amplification of the signal when the distance between the vth VM and the nth source becomes close to zero. Similarly to [260], one can assign an arbitrary pick-up pattern to the vth VM defining the values of $C_v(\cdot)$. This enables the simulation of different microphone models. The pick-up pattern of the common directional microphones e.g. cardioid, supercardioid and hypercardioid, can be easily modelled by a circular harmonics expansion. As an alternative, the nominal pick-up pattern of a real microphone can be adopted. It is worth noting that in general, any function can be assigned to $C_v(\cdot)$, defining even non-physical characteristics to the VM.



(a)

(b)

Figure 5.4: *Simulation setup. (a) Two sources scenario with a single VM with cardioid pattern. (b) A stereo recording scenario with two cardioid VMs in X-Y configuration.*

5.2.4 Simulations

In order to validate the VM technique, our procedure is tested through software simulations. We run the simulation using the setup shown in Figure 5.4. Both the sources and the virtual microphones present a cardioid directivity pattern for all the frequency range. We employ six circular microphone arrays having a radius of $10\,\mathrm{cm}$, accommodating $M = 4$ omnidirectional microphones. The additive noise at the sensor is simulated using a random white Gaussian noise, whose variance is set in such a way to obtain a desired SNR w.r.t each microphone in the acoustic scene. The microphone signals are processed after performing a STFT with a $20\,\mathrm{ms}$ Hann window and $75\,\%$ overlap. The estimated signal $\hat{S}(\omega, t, \check{\boldsymbol{r}}_v)$ of the vth VM is obtained using the estimated values (Section 5.2.2), while the reference signal $S(\omega, t, \check{\boldsymbol{r}}_v)$ is computed adopting the actual theoretical values. For the tests we employ speech signals taken from [264] in order to simulate an actual scenario with one or more talkers.

We devise five different metrics as a means to evaluate the proposed technique: the first three are devoted to the evaluation of the analysis stage while the last two to the synthesis stage. Here below we report a list of the aforementioned metrics.

1. **Localization Error:** In order to analyze the performance of the localization step, the Mean Squared Error (MSE) between estimated and actual positions of the sources has been adopted:

$$\mathrm{LE}\left(\mathbf{r}'\right) = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\mathbf{r}}'_n - \mathbf{r}'_n\| \tag{5.30}$$

2. **Directivity Error:** The source radiance pattern has been evaluated with an ad hoc metrics called Directivity Error (DE). DE is defined for each source n as follows:

$$\mathrm{DE}_n = \frac{\sum_\omega \sum_\phi \left[\hat{D}_n(\omega, \phi) - D_n(\omega, \phi)\right]^2}{IW} \tag{5.31}$$

where I is the number of considered angles and W the number of frequency bins.

3. **Source to Distortion Ratio:** The Source to Distortion Ratio (SDR) is defined as the energy ratio between the desired signal and sources of distortion (i.e. interferers, noise and artifacts) and it has been used to evaluate the signal extraction performance using the methodology described in [275].

4. **Synthesized Signal Error:** The Synthesized Signal Error (SSE) is defined as the Normalized Mean Squared Error (NMSE) between the vth VM estimated signal and the reference one:

$$\mathrm{SSE}_v = 10\log_{10}\frac{\sum_\omega \sum_t \left|\hat{S}\left(\omega, t, \check{\boldsymbol{r}}_v\right) - S\left(\omega, t, \check{\boldsymbol{r}}_v\right)\right|^2}{\sum_\omega \sum_t \left|S(\omega, t, \check{\boldsymbol{r}}_v)\right|^2} \tag{5.32}$$

5. **Intensity Ievel Difference:** The ILD is defined as:

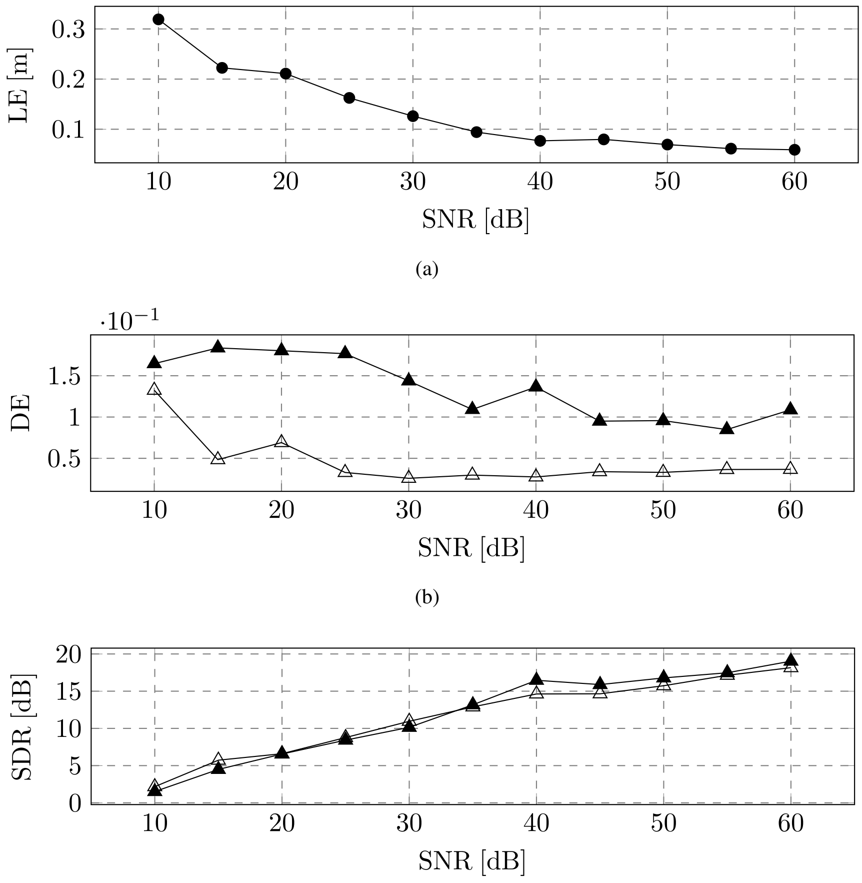
$$\mathrm{ILD}(\varrho) = 10\log_{10}\frac{\sum_\omega \sum_t \left|\hat{S}\left(\omega, t, \check{\boldsymbol{r}}_I, \check{\boldsymbol{o}}_I(\varrho)\right)\right|^2}{\sum_\omega \sum_t \left|\hat{S}\left(\omega, t, \check{\boldsymbol{r}}_R, \check{\boldsymbol{o}}_R(\varrho)\right)\right|^2} \tag{5.33}$$

where $\hat{S}\left(\omega, t, \check{\boldsymbol{r}}_I, \check{\boldsymbol{o}}_I(\varrho)\right)$ and $\hat{S}\left(\omega, t, \check{\boldsymbol{r}}_R, \check{\boldsymbol{o}}_R(\varrho)\right)$ are the signals of two coincident cardioid virtual microphones. The microphones angled 90° in a XY stereo configuration directed toward ϱ (see Figure 5.4(b)), hence the cardioid patterns of the two sensor are oriented toward $\varrho - \pi/4$ and $\varrho + \pi/4$.

The setup in Figure 5.4(b) is devoted for the evaluation of the ILD, while the other metrics are computed with the setup of Figure 5.4. Concerning the metrics related to the analysis stage, we report the results obtained by varying the SNR in Figure 5.5. As we can see both the LE and the DE decrease as the SNR increases, while the SDR increases monotonically. It is worth noting that, although at very low SNR we have a localization error in the order of $30\,\mathrm{cm}$, we are still able to maintain a low DE and a positive SDR.

As far as the metrics related to the synthesis stage are concerned, we can note from Figure 5.6 that the SSE monotonically decreases as the SNR increases, but the error is in the order of $-3\,\mathrm{dB}$ even when the SNR is very low. Moreover, the behaviour of this metric directly reflects that of the three analysis metrics. As an example, we report in Figure 5.7 the comparison of the spectrograms of the estimated and reference virtual microphone signals for a $\mathrm{SNR} = 40\,\mathrm{dB}$. We notice that the two spectrograms are very similar especially in the frequency range below $2\,\mathrm{kHz}$. Above that frequency the spatial aliasing dominates affecting in particular the frequency dependent operations i.e., the directivity pattern estimation (Section 5.2.2) and the signal estimation (Section 5.2.2).

Concerning the ILD, we consider a stereo recording scenario using two directional VMs with cardioid pattern and an angular difference of 90°, as shown in Figure 5.4(b). The difference in sound pressure level between the two microphones is an important spatial cue to be reproduced. In Figure 5.8 the ILD of the two cardioid VMs is illustrated as a function of the stereo VMs look direction ϱ, where the microphone pair is rotated from 0° to 360°. As expected, the behavior of the ILD curve reflects the orientation of the microphones, with the maximum and minimum in correspondence of a zero of the cardioid pattern pointing towards the source (225° and 315°). The zero crossings



(a)

(b)

(c)

Figure 5.5: *Analysis metrics. The markers in* DE *and* SDR *curves correspond to the ones in setup Figure 5.4(a).*

of the curve occur at 90° and 270° when the signal of the source is sensed with the same intensity from the two VMs. The curve of Figure 5.8 corresponds to the expected behavior of a of a physical stereo pair microphones with the same characteristics.

5.3 Parametric Sound Field Reconstruction in Reverberant Environments

In this section we propose an enhanced version of the parametric method introduced in Section 5.2, that extends the analysis to directional sources in reverberant environments. The first step of the sound field analysis concerns the separation of the direct and diffuse components from the measured signals and the localization of the sources. The direct component is processed assuming a spherical harmonic representation of the emitted sound field that inherently describes the directional behavior of the sources. The diffuse component is assumed to be isotropic and homogeneous. The synthesis of the VM signal is accomplished by properly mixing the estimated direct and diffuse components at the desired location.

Results show that the proposed technique is able to reconstruct the main cues of the VM signal, for instance, the ones related to reverberation (e.g. Direct to Reverberant Ratio) and spatial recording (Interchannel Ievel Difference). Moreover, by analyzing

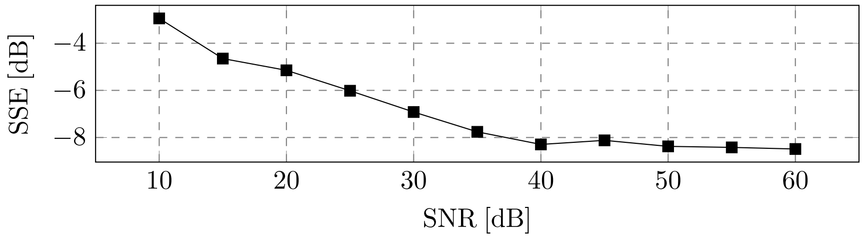
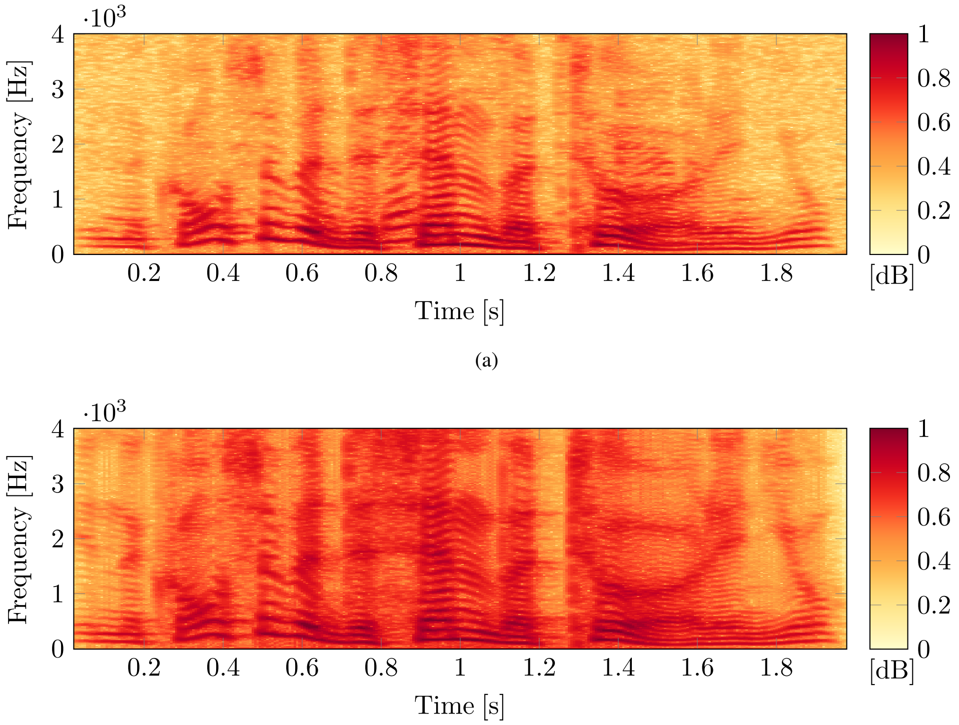


Figure 5.6: SSE *of the VM in Figure 5.4(a) for different level of* SNR *at the microphones.*



(a)

(b)

Figure 5.7: *Spectrograms of the estimated (a) and reference (b) virtual microphone signals.*

the metrics at different locations in the space, we show that the proposed approach is able to capture the spatial characteristic of the recorded acoustic scene.

5.3.1 Data Model and Problem Formulation

Here we characterize the data model in a reverberant environment and the virtual miking problem. We specify the general data model of Section 5.1 and we also introduce the parameters that need to be estimated. We then define the model of the signals acquired by the microphones recording the acoustic scene. Finally, in Section 5.3.1 we describe the virtual miking problem with the help of a block diagram underlying the needs and requirements of the proposed approach. We report here the VM signal model (5.2) that is composed of two contribution: the direct sound and the diffuse components

$$\begin{aligned} S(t,\omega,\check{\boldsymbol{r}}_v) = {} & C_v(\omega) S_{n,\text{dir}}(t,\omega,\check{\boldsymbol{r}}_v) \\ & + Q_v(\omega) S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v),\; n \in \{1,\dots,N\}, \end{aligned}$$

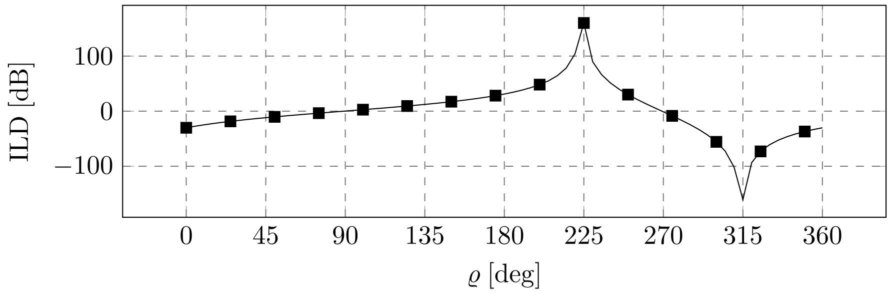


Figure 5.8: ILD *as a function of the rotation of the two cardioid virtual microphones in the stereo recording scenario of Figure 5.4(b).*

Let us denote with $\check{\boldsymbol{r}}_{v,n} = \check{\boldsymbol{r}}_v - \boldsymbol{r}'_n = [\check{x}_{v,n}, \check{y}_{v,n}, \check{z}_{v,n}]$ the vector pointing from the source position to the VM position (see Figure 5.1) and with $\check{\rho}_{v,n}$, $\check{\theta}_{v,n}$ and $\check{\phi}_{v,n}$ the coordinates of $\check{\boldsymbol{r}}_{v,n}$ in a spherical coordinate system, i.e.,

$$\begin{aligned}\check{\rho}_{v,n} &= \sqrt{\check{x}_{v,n}^2 + \check{y}_{v,n}^2 + \check{z}_{v,n}^2},\\ \check{\theta}_{v,n} &= \arccos \frac{\check{z}_{v,n}}{\check{\rho}_{v,n}},\\ \check{\phi}_{v,n} &= \arctan \frac{\check{y}_{v,n}}{\check{x}_{v,n}}.\end{aligned} \tag{5.34}$$

The term $S_{n,\mathrm{dir}}(t, \omega, \check{\boldsymbol{r}}_v)$ represents the direct sound emitted by the nth source and received by the vth VM and it is modelled as the exterior field [229] (see Section 3.6.2)

$$S_{n,\mathrm{dir}}(t, \omega, \check{\boldsymbol{r}}_v) = \sum_{\ell=0}^{I} \sum_{\mu=-\ell}^{\ell} \beta_{\ell\mu}^{n}(t,\omega) h_\ell\left(k\check{\rho}_{v,n}\right) Y_{\ell\mu}\left(\check{\theta}_{v,n}, \check{\phi}_{v,n}\right), \tag{5.35}$$

where $k = 2\pi f/c$, c is the speed of sound, $\beta_{\ell\mu}^{n}(\omega)$ are the exterior sound field coefficients of the nth source, $h_\ell\left(\cdot\right)$ is the ℓth order spherical Hankel function and $Y_{\ell\mu}\left(\check{\theta}_{v,n}, \check{\phi}_{v,n}\right)$ is the spherical harmonic of order ℓ and degree μ, defined as

$$Y_{\ell\mu}\left(\check{\theta}_{v,n}, \check{\phi}_{v,n}\right) = K_{\ell\mu} \mathcal{P}_{\ell\mu}\left(\cos(\check{\theta}_{v,n})\right) e^{j\mu\check{\phi}_{v,n}}, \tag{5.36}$$

with

$$K_{\ell\mu} = (-1)^{\mu} \sqrt{\frac{(2\ell+1)}{4\pi} \frac{(\ell-\mu)!}{(\ell+\mu)!}} \tag{5.37}$$

and $\mathcal{P}_{\ell\mu}(\cdot)$ the normalized associated Legendre polynomial (eq. (3.22) and (3.23)). It is worth noting that even though sources, arrays, and VMs are assumed to be lying on the same plane as in Section 5.2, we consider them as placed in a 3D environment. Hence, in (5.35) we adopt a 3D propagation model. It is worth noticing that the direct signal model in (5.35) differs from (5.3). In this case, we adopt a more efficient description of the direct sound field. In fact, through (5.35) both the signal and the directivity of the sources are encoded by means of the spherical harmonics expansion. This provides a compact description that automatically model the directivity of the sources, removing the requirement of directly estimating this component as in (5.3).

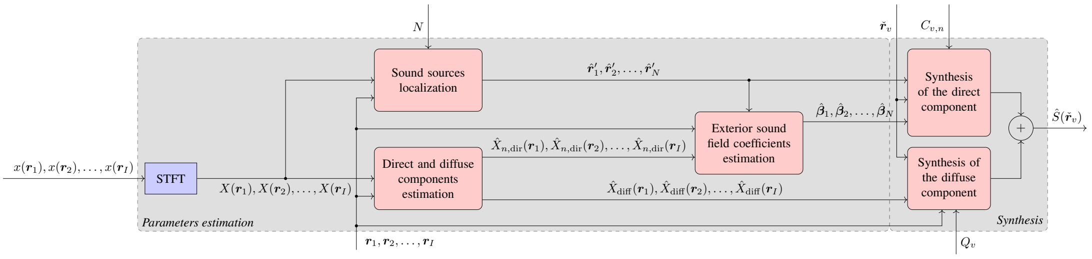

N
$\check{\boldsymbol{r}}_v$
$C_{v,n}$
$x(\boldsymbol{r}_1), x(\boldsymbol{r}_2), \ldots, x(\boldsymbol{r}_I)$
STFT
$X(\boldsymbol{r}_1), X(\boldsymbol{r}_2), \ldots, X(\boldsymbol{r}_I)$
Sound sources localization
Direct and diffuse components estimation
$\hat{\boldsymbol{r}}'_1, \hat{\boldsymbol{r}}'_2, \ldots, \hat{\boldsymbol{r}}'_N$
$\hat{X}_{n,\text{dir}}(\boldsymbol{r}_1), \hat{X}_{n,\text{dir}}(\boldsymbol{r}_2), \ldots, \hat{X}_{n,\text{dir}}(\boldsymbol{r}_I)$
$\hat{X}_{\text{diff}}(\boldsymbol{r}_1), \hat{X}_{\text{diff}}(\boldsymbol{r}_2), \ldots, \hat{X}_{\text{diff}}(\boldsymbol{r}_I)$
Exterior sound field coefficients estimation
$\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_N$
Synthesis of the direct component
Synthesis of the diffuse component
+
$\hat{S}(\check{\boldsymbol{r}}_v)$
Parameters estimation
Synthesis
$\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_I$
Q_v

Figure 5.9: *The virtual miking technique block diagram. The microphone signals $x(\boldsymbol{r}_i)\, i = 1, \ldots, I$ are first transformed using the STFT into $X(\boldsymbol{r}_i)\, i = 1, \ldots, I$ which are used as input for the Sound sources localization (Section 5.3.2) and Direct and diffuse components estimation (Section 5.3.3) blocks along with the location of the microphones $\boldsymbol{r}_i$, $i = 1, \ldots, I$. The number of sources N is provided as input to the Sound sources localization block. The estimated location of the sources $\hat{\boldsymbol{r}}'_n$, $n = 1, \ldots, N$ and the direct component estimates $\hat{X}_{n,\text{dir}}$ are used as input for the Exterior sound field coefficients estimation block (Section 5.3.4). As regards the Synthesis phase (Section 5.3.5), the position of the vth VM $\check{\boldsymbol{r}}_v$ is shared by both the Synthesis of the direct component (Section 5.3.5) and Synthesis of the diffuse component (Section 5.3.5) blocks. In addition, the Synthesis of the direct component block requires the vth VM pick-up pattern $C_{v,n}$, $n = 1, \ldots N$, the estimated location of the sources $\hat{\boldsymbol{r}}'_n$, $n = 1, \ldots, N$ and the estimates of the exterior sound field coefficients $\hat{\boldsymbol{\beta}}_n$, $n = 1, \ldots, N$, while the sensitivity of the vth VM Q_v and the estimated diffuse components $\hat{X}_{\text{diff}}$ are given as input to the Synthesis of the diffuse component block. Finally, the signal of the vth VM $\hat{S}\,(\check{\boldsymbol{r}}_v)$ is obtained as the sum of the synthesized direct and diffuse components.*

The term $S_{\text{diff}}(t, \omega, \check{\boldsymbol{r}}_v)$ represents the diffuse sound field component and it is assumed as spatially isotropic and homogeneous, i.e., it arrives with equal strength from all the directions and its mean power does not vary with the position [241, 260]. It is worth noticing that in (5.2) we implicitly assume that the VM is noiseless.

The signal acquired by the ith omnidirectional microphone placed in $\boldsymbol{r}_i$ is modeled as

$$X\left(t, \omega, \boldsymbol{r}_i\right) = X_{n,\text{dir}}(t, \omega, \boldsymbol{r}_i) + X_{\text{diff}}(t, \omega, \boldsymbol{r}_i) + N(t, \omega, \boldsymbol{r}_i). \tag{5.38}$$

The term $X_{n,\text{dir}}(t, \omega, \boldsymbol{r}_i)$ is the direct sound emitted by the nth source and received by the ith microphone and, similarly to (5.35), is modeled as

$$X_{n,\text{dir}}(t, \omega, \boldsymbol{r}_i) = \sum_{\ell=0}^{I} \sum_{\mu=-\ell}^{\ell} \beta_{\ell\mu}^{n}(t, \omega) h_\ell\left(k\rho_{i,n}\right) Y_{\ell\mu}\left(\theta_{i,n}, \phi_{i,n}\right), \tag{5.39}$$

where $\rho_{i,n}$, $\theta_{i,n}$ and $\phi_{i,n}$ are the spherical coordinates of the vector $\boldsymbol{r}_{i,n} = \boldsymbol{r}_i - \boldsymbol{r}'_n$ (see (5.34)). The terms $X_{\text{diff}}(t, \omega, \boldsymbol{r}_i)$ and $N(t, \omega, \boldsymbol{r}_i)$ are the spatially isotropic and homogeneous diffuse sound component and the ith sensor self-noise, respectively. The microphone self-noise $N(t, \omega, \boldsymbol{r}_i)$ is modelled as an uncorrelated zero-mean complex Gaussian noise with mean power

$$\Phi_{N,ii}(t, \omega) = E\{N(t, \omega, \boldsymbol{r}_i) N^{\star}(t, \omega, \boldsymbol{r}_i)\}, \tag{5.40}$$

where $E\{\cdot\}$ denotes the mathematical expectation and $(\cdot)^{\star}$ refers to the conjugate of a complex number.

Problem Formulation

In this section we approach the sound field reconstruction problem as a virtual miking operation performed in a parametric fashion. In particular, in Section 5.1.1, we developed in (5.2) and (5.38) an extended parametric model (with respect to the one adopted in Section 5.2) for both the VMs and the microphones recording the scene, respectively. The proposed solution can be seen as a system characterized by a set of unknown parameters that need to be estimated. The inputs to the estimation problem are the signals $X(t,\omega,\boldsymbol{r}_i)$, $i=1,\dots,I$ of the microphones, their positions $\boldsymbol{r}_i$, the characteristics of each VM, namely the position $\check{\boldsymbol{r}}_v$, the pick-up pattern $C_v(\omega)$ and the sensitivity to diffuse noise $Q_v(\omega)$ and the number of sources N. In particular, as regards the latter parameter, it can be estimated using other sensors in the room or directly from the signals at the microphones as proposed, for example, in [25, 26, 154, 192, 194, 255, 277, 291]. The output of the algorithm is an estimate $\hat{S}(\check{\boldsymbol{r}}_v)$ of the VM signal $S(\check{\boldsymbol{r}}_v)$.

In Figure 5.9 a graphical representation of the proposed solution is depicted. In the block diagram we can identify the two main phases of the procedure namely the *parameters estimation* and the *synthesis* phase. In the former all the parameters needed for the synthesis of the VM signal are estimated. In particular, as it is clear from (5.2), in order to synthesize the signal at each VM we need to estimate both the direct $S_{n,\text{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ and the diffuse $S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$ components.

VM direct component estimation In (5.35) the model of the direct component $S_{n,\text{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ is described. The parameters characterizing the direct sound component of a VM are the source location $\boldsymbol{r}'_n$, $n=1,\dots,N$ and the exterior sound field coefficients $\beta^n_{\ell,\mu}(t,\omega)$, $n=1,\dots,N$. The positions $\boldsymbol{r}'_n$ of the sources are estimated using the acoustic source localization algorithm described in Section 5.2.2.

The estimation of the exterior sound field coefficients $\beta^n_{\ell,\mu}(t,\omega)$ from the microphone signals requires the knowledge of the direct sound component $X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)$ at each microphone (see (5.39)). However, only the microphone signals $X(t,\omega,\boldsymbol{r}_i)$ are directly available. It follows that a procedure for estimating the direct and the diffuse components from $X(t,\omega,\boldsymbol{r}_i)$, $i=1,\dots,I$ is required. This procedure must be *blind* with respect to the room transfer function between sources and microphones. This, in fact, is a desirable feature, as measuring the transfer functions is not always feasible for all possible source locations and, in addition, transfer functions can be time-varying.

The algorithm for the estimation of the direct components is described in Section 5.3.3. It is worth noticing that the algorithm used for estimating the direct component requires the knowledge of the exterior field spherical harmonic coefficients, which is detailed in the same section. Finally, given the acoustic scene parameters described above, it is possible to synthesize the VM signals. This is detailed in Section 5.3.5.

VM diffuse component estimation Starting from from the microphone diffuse sound components $X_{\text{diff}}(\cdot)$, the diffuse sound component $S_{\text{diff}}(\cdot)$, can be estimated, as detailed in Section 5.3.5. Inputs for the estimation of the diffuse components, as shown in Figure 5.9, are the VM position $\check{\boldsymbol{r}}_v$, the microphone positions $\boldsymbol{r}_i$ and the VM sensitivity to the diffuse field $Q_v(\omega)$.

As defined in (5.2), an estimate of the vth VM signal is obtained as the linear combination of the estimates of the direct component $\hat{S}_{n,\text{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ and the diffuse component $\hat{S}_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$.

5.3.2 Source Localization

The accurate estimation of the source location is a crucial step, as the estimation of all the other parameters depend on that. Furthermore, it is well-known in the literature [178, 244] that accurate source localization in the presence of strong reverberation is a challenging problem. Again, we approach the source localization problem as a two-step procedure: in the first step we estimate a set of source DoAs for each array, while in the second step the locations of the sources are found solving the DoAs association and triangulation problem [62].

DoA Estimation

In the literature different DoA estimation algorithms can be found and they can be mainly divided into two classes: parametric [187, 236, 253] and spatial methods [34, 42, 268]. The former class of techniques leverages on assumptions about the covariance structure of the signals, while the latter concerns the computation of a spatial filter, customarily through beamforming. Here, due to the assumption on the setup (see Section 5.1.1), we are interested in the 2D position of the sources, hence, we adopt an improved version of the localization technique introduced in Section 5.2.2 that is based on spatial filtering. In particular, in this case we are dealing with reverberant environment, therefore, the presence of reflections due to the reverberation is likely to introduce errors in the estimate of the DoAs $\bar{\alpha}_n^{(a)}$ (5.9). It is also worth noticing that, in general, correlation between the source signals can negatively affect the estimation of the DoAs. The use of sufficiently short time windows in the STFT and the assumption of uncorrelation among the time-frequency bins, however, attenuates this problem. With the aim of reducing the impact of reverberation on the location accuracy, we select the DoAs in $\bar{\boldsymbol{\alpha}}^{(a)}(t)$ compatible with source locations inside the ROI described in Section 5.3.1. This is done by intersecting the half-lines with origin $\boldsymbol{v}_a$ and direction $\boldsymbol{D}^{(a)}(t) = \left[\boldsymbol{d}_1^{(a)}(t), \boldsymbol{d}_2^{(a)}(t), \ldots, \boldsymbol{d}_N^{(a)}(t)\right]$, $\boldsymbol{d}_n^{(a)}(t) = [\cos\bar{\alpha}_n^{(a)}(t), \sin\bar{\alpha}_n^{(a)}(t)]^T$ with the polygon $\mathcal{R}$ that defines the ROI

$$\tilde{\boldsymbol{\alpha}}^{(a)}(t) = \mathcal{I}(\boldsymbol{v}_a, \boldsymbol{D}^{(a)}(t), \mathcal{R}), \tag{5.41}$$

where $\mathcal{I}$ is the operator that returns all the DoAs for which at least one intersection with the polygon edges exists and

$$\tilde{\boldsymbol{\alpha}}^{(a)}(t) = \left[\tilde{\alpha}_1^{(a)}(t), \tilde{\alpha}_2^{(a)}(t), \ldots, \tilde{\alpha}_{\tilde{N}}^{(a)}(t)\right]^T \tag{5.42}$$

is the resulting DoA vector with dimensions $\tilde{N} \times 1$, $\tilde{N} \leq N$.

Association of DoAs and triangulation

The DoA association and disambiguation problem appears in the context of multisource scenarios. As described in details in Section 5.2.2, this problem can be solved by mapping the DoAs in the projective ray space, where the information related to the location of the acoustic source is conveniently represented by linear patterns. We exploit this procedure for localizing the acoustic sources in reverberant environments, using $\tilde{\boldsymbol{\alpha}}^{(a)}$ (5.41) in (5.12) instead of $\bar{\boldsymbol{\alpha}}^{(a)}$ (5.9).

5.3.3 Direct and Diffuse Components Estimation

Once the source locations are obtained, we address the problem of estimating the direct and diffuse components of a microphone signal namely $X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)$ and $X_{\text{diff}}(t,\omega,\boldsymbol{r}_i)$ from the recorded microphone signal .This is a crucial step of the process, as the knowledge of $X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)$ is required to estimate the exterior sound field coefficients of the sources (see (5.39)) and $X_{\text{diff}}(t,\omega,\boldsymbol{r}_i)$ is needed for the estimation of the VM diffuse component $S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$.

The estimation of the direct and the diffuse component is also known as the dereverberation problem. The dereverberation algorithms proposed in the literature can be divided in two categories: inverse filtering algorithms [51, 73, 173]; and algorithms that estimate and suppress reverberation with spectral subtraction or Wiener filtering [241, 259]. From an operative standpoint, the two categories differ in the fact that the former requires the knowledge of the room transfer function, while the latter does not need it. In this work, as stated in Section 5.3.1, the second class meets our requirements.

Following [241] and [259] we can obtain an estimate of the direct sound component $X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)$ at the position $\boldsymbol{r}_i$ as the output of a squared root Wiener filter whose coefficients are computed as [241, 269]

$$G_{\text{dir}}(t,\omega,\boldsymbol{r}_i) = \sqrt{1 - \frac{1}{\text{CDR}(t,\omega,\boldsymbol{r}_i) + 1}}, \tag{5.43}$$

where $\text{CDR}(t,\omega,\boldsymbol{r}_i)$ is the time-frequency dependent signal to diffuse ratio at the ith microphone, defined as

$$\text{CDR}(t,\omega,\boldsymbol{r}_i) = \frac{\Phi_{\text{dir},ii}(t,\omega)}{\Phi_{\text{diff},ii}(t,\omega)}. \tag{5.44}$$

Here $\Phi_{\text{dir},ii}$ and $\Phi_{\text{diff},ii}$ are the auto-power spectra of the direct and diffuse component, respectively, and are defined as

$$\begin{aligned} \Phi_{\text{dir},ii}(t,\omega) &= E\{X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)X^{\star}_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)\} \\ \Phi_{\text{diff},ii}(t,\omega) &= E\{X_{\text{diff}}(t,\omega,\boldsymbol{r}_i)X^{\star}_{\text{diff}}(t,\omega,\boldsymbol{r}_i)\}. \end{aligned} \tag{5.45}$$

As shown in [241], an estimate of $\text{CDR}(t,\omega,\boldsymbol{r}_i)$ can be obtained from the knowledge of the microphone signal coherence function and the diffuse noise coherence function. This can be accomplished using the CDR estimator defined in [241] as

$$\begin{aligned} \text{CDR}(\boldsymbol{r}_i) = &\frac{\Gamma_{\text{diff},ii'}\text{Re}\{\hat{\Gamma}_{ii'}\} - |\hat{\Gamma}_{ii'}|^2}{|\hat{\Gamma}_{ii'}|^2 - 1} \\ &- \frac{\sqrt{\left(\Gamma_{\text{diff},ii'}\text{Re}\{\hat{\Gamma}_{ii'}\}\right)^2 - \left(\Gamma_{\text{diff},ii'}|\hat{\Gamma}_{ii'}|\right)^2 + (\Gamma_{\text{diff},ii'})^2 - 2\Gamma_{\text{diff},ii'}\text{Re}\{\hat{\Gamma}_{ii'}\} + |\hat{\Gamma}_{ii'}|^2}}{|\hat{\Gamma}_{ii'}|^2 - 1}, \end{aligned} \tag{5.46}$$

where $\text{Re}\{\cdot\}$ is the operator that retrieves the real part of a complex number. The dependencies on time and frequency have been omitted for the sake of readability. The term $\Gamma_{\text{diff},ii'}(\omega)$ in (5.46) is the diffuse noise coherence function between the ith and i'th microphones. Assuming a spherically isotropic sound field as in (5.38), $\Gamma_{\text{diff},ii'}(\omega)$

can be modelled as [241]

$$\Gamma_{\text{diff},ii'}(\omega) = \frac{\Phi_{\text{diff},ii'}(t,\omega)}{\sqrt{\Phi_{\text{diff},ii}(t,\omega)\Phi_{\text{diff},i'i'}(t,\omega)}} = \frac{\sin(kd_{ii'})}{kd_{ii'}}, \tag{5.47}$$

where

$$\begin{aligned} \Phi_{\text{diff},ii'}(t,\omega) &= E\{X_{\text{diff}}(t,\omega,\boldsymbol{r}_i)X^{\star}_{\text{diff}}(t,\omega,\boldsymbol{r}_{i'})\}, \\ d_{ii'} &= \|\boldsymbol{r}_i - \boldsymbol{r}_{i'}\|_2 \end{aligned} \tag{5.48}$$

with $\|\cdot\|_2$ the ℓ-2 norm of a vector.

The term $\hat{\Gamma}_{ii'}(t,\omega)$ in (5.46) is the estimate of the microphone signal coherence function between the ith and the i'th microphone. If we assume that the sensor noise between microphones i and i' is uncorrelated, the microphone signal coherence function can be estimated as [259]

$$\hat{\Gamma}_{ii'}(t,\omega) = \frac{\hat{\Phi}_{ii'}(t,\omega)}{\sqrt{\hat{\Phi}_{ii}(t,\omega) - \hat{\Phi}_{N,ii}(t,\omega)}\sqrt{\hat{\Phi}_{i'i'}(t,\omega) - \hat{\Phi}_{N,i'i'}(t,\omega)}}, \tag{5.49}$$

where $\Phi_{\text{N},ii}(t,\omega)$ is the noise auto-power spectrum defined in (5.40) and

$$\Phi_{ii'}(t,\omega) = E\{X\left(t,\omega,\boldsymbol{r}_i\right)X^{\star}\left(t,\omega,\boldsymbol{r}_{i'}\right)\}. \tag{5.50}$$

The auto and cross spectra can be obtained from the microphone signals by recursive averaging [241]

$$\hat{\Phi}_{ii'}(t,\omega) = \lambda\hat{\Phi}_{ii'}(t-1,\omega) + (1-\lambda)X\left(t,\omega,\boldsymbol{r}_i\right)X^{\star}\left(t,\omega,\boldsymbol{r}_{i'}\right) \tag{5.51}$$

where λ is a constant in the range $[0,1)$. In our scenario, the microphone pairs are chosen as belonging to the same array. The sensor noise auto-spectra $\hat{\Phi}_{N,ii}(t,\omega)$ and $\hat{\Phi}_{N,i'i'}(t,\omega)$ can be obtained applying recursive averaging on the microphone signals as in (5.51) when neither acoustic sources nor diffuse noise are present (i.e., only the sensor noise component is active). In order to determine the activity or inactivity of the sources we use the voice activity detector in [247]. It is worth noting that [247] assumes that all sources emit speech signals that are sufficiently sparse in the time-frequency domain thus agreeing with the assumption stated in Section 5.1.1.

Once an estimate of the CDR at the ith microphone is obtained using (5.46) we can use (5.43) to obtain the Wiener filter coefficients that allows to extract the direct component of a the ith microphone signal. However, as highlighted in [241], a more practical implementation of (5.43) is given by [269]

$$G_{\text{dir}}(t,\omega,\boldsymbol{r}_i) = \max\left\{G_{\min}, \sqrt{1 - \frac{\nu}{\text{CDR}(t,\omega,\boldsymbol{r}_i)+1}}\right\}, \tag{5.52}$$

where ν is the oversubtraction factor and $G_{\min}$ the gain floor. The term ν controls the amount of noise subtracted from the noisy signal. For full noise subtraction, $\nu = 1$ and for over-subtraction $\nu > 1$. The term $G_{\min}$ acts as a lower bound for the filter coefficients weights. This is useful in order to reduce artefacts in the output signal. Inspecting (5.52), it is clear that high values of CDR leads to low filter gain and vice versa.

Finally, the filter in (5.52) is used to compute the direct signal component at the ith microphone through [241]

$$\hat{X}_{\mathrm{dir}}(t,\omega,\boldsymbol{r}_i) = G_{\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)U(t,\omega,\boldsymbol{r}_i), \tag{5.53}$$

where

$$U(t,\omega,\boldsymbol{r}_i) = \sqrt{\frac{Z\left(t,\omega,\boldsymbol{r}_i\right) + Z\left(t,\omega,\boldsymbol{r}_{i'}\right)}{2}} e^{j\arg\{X(t,\omega,\boldsymbol{r}_i)\}}, \tag{5.54}$$

with

$$\begin{aligned} Z\left(t,\omega,\boldsymbol{r}_i\right) &= \left|X\left(t,\omega,\boldsymbol{r}_i\right)\right|^2 - \hat{\Phi}_{N,ii}(t,\omega) \\ Z\left(t,\omega,\boldsymbol{r}_{i'}\right) &= \left|X\left(t,\omega,\boldsymbol{r}_{i'}\right)\right|^2 - \hat{\Phi}_{N,i'i'}(t,\omega) \end{aligned} \tag{5.55}$$

and $\arg\{\cdot\}$ the operator that takes the argument of a complex number. The spatial magnitude averaging performed in (5.54) is typically used in order to reduce the variance of the estimates for microphone array post-filters [166, 294].

The diffuse component of the microphone signal can be obtained using the filter [259]

$$G_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i) = \sqrt{1 - \left[G_{\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)\right]^2}, \tag{5.56}$$

where $G_{\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)$ is defined in (4.59). It follows that an estimate of $X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)$ can be obtained as

$$\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i) = G_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)U(t,\omega,\boldsymbol{r}_i), \tag{5.57}$$

where $U(t,\omega,\boldsymbol{r}_i)$ is defined in (5.54). As demonstrated in Appendix A, using the filters in (5.52) and (5.56) and assuming that $\nu = 1$, $G_{\min} = 0$ and that the auto-spectra of the direct, diffuse and noise components at the i microphone and at the i' microphone are the same, the power of the estimated sound field components corresponds to the actual sound power (i.e., $E\{|\hat{X}_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)|^2\} = E\{|X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)|^2\}$ and $E\{|\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)|^2\} = E\{|X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)|^2\}$) [269].

5.3.4 Exterior Sound Field Coefficients Estimation

Once the direct signal components of the microphone signals have been estimated using (5.53), and the sources have been localized, we can exploit the model of the direct sound component in (5.39) in order to estimate the set of spherical harmonics coefficients related to each source in the acoustic scene. Let us define the vector $\hat{\mathbf{x}}_{\mathrm{dir}}(t,\omega)$ containing the estimates of the direct component for all the microphones, i.e.,

$$\left[\hat{\mathbf{x}}_{\mathrm{dir}}(t,\omega)\right]_i = \hat{X}_{\mathrm{dir}}(t,\omega,\boldsymbol{r}_i) \qquad i = 1,\ldots,I, \tag{5.58}$$

where $[\cdot]_i$ is the ith element of the vector.

We denote the vector of the coefficients of the spherical harmonic for the nth source as

$$\boldsymbol{\beta}_n(t,\omega) = \left[\beta^n_{00}(t,\omega), \beta^n_{0-1}(t,\omega), \ldots, \beta^n_{LL}(t,\omega)\right]^T, \tag{5.59}$$

where $(\cdot)^T$ is the transpose operator.

Let us define the matrix $\widehat{\mathbf{Y}}_n(k)$ containing the spherical harmonics as

$$\widehat{\mathbf{Y}}_n(k) = \begin{bmatrix} h_0\left(k\hat{\rho}_{1,n}\right)Y_{00}\left(\hat{\theta}_{1,n},\hat{\phi}_{1,n}\right) h_1\left(k\hat{\rho}_{1,n}\right)Y_{1-1}\left(\hat{\theta}_{1,n},\hat{\phi}_{1,n}\right) \cdots h_L\left(k\hat{\rho}_{1,n}\right)Y_{LL}\left(\hat{\theta}_{1,n},\hat{\phi}_{1,n}\right) \\ h_0\left(k\hat{\rho}_{2,n}\right)Y_{00}\left(\hat{\theta}_{2,n},\hat{\phi}_{2,n}\right) h_1\left(k\hat{\rho}_{2,n}\right)Y_{1-1}\left(\hat{\theta}_{2,n},\hat{\phi}_{2,n}\right) \cdots h_L\left(k\hat{\rho}_{2,n}\right)Y_{LL}\left(\hat{\theta}_{2,n},\hat{\phi}_{2,n}\right) \\ \vdots \qquad\qquad \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\ h_0\left(k\hat{\rho}_{I,n}\right)Y_{00}\left(\hat{\theta}_{I,n},\hat{\phi}_{I,n}\right) h_1\left(k\hat{\rho}_{I,n}\right)Y_{1-1}\left(\hat{\theta}_{I,n},\hat{\phi}_{I,n}\right) \cdots h_L\left(k\hat{\rho}_{I,n}\right)Y_{LL}\left(\hat{\theta}_{I,n},\hat{\phi}_{I,n}\right) \end{bmatrix} \tag{5.60}$$

where $\hat{\rho}_{i,n}$, $\hat{\theta}_{i,n}$ and $\hat{\phi}_{i,n}$. are the estimates of $\rho_{i,n}$, $\theta_{i,n}$ and $\phi_{i,n}$ defined in (5.39) obtained using the estimate of the n source position $\hat{\boldsymbol{r}}'_n$. In the light of the definitions in (5.58) and (5.60), the direct sound components acquired by the microphones are given by

$$\begin{aligned} \hat{\mathbf{x}}_{\text{dir}}(t,\omega) &= \left[\widehat{\mathbf{Y}}_1(k)\widehat{\mathbf{Y}}_2(k)\cdots\widehat{\mathbf{Y}}_N(k)\right]\begin{bmatrix} \boldsymbol{\beta}_1(t,\omega) \\ \vdots \\ \boldsymbol{\beta}_N(t,\omega) \end{bmatrix}, \\ &= \widehat{\mathbf{Y}}(k)\boldsymbol{\beta}(t,\omega). \end{aligned} \tag{5.61}$$

An estimate $\hat{\boldsymbol{\beta}}(t,\omega)$ of $\boldsymbol{\beta}(t,\omega)$ can be obtained as

$$\hat{\boldsymbol{\beta}}(t,\omega) = \widehat{\mathbf{Y}}^{\dagger}(k)\hat{\mathbf{x}}_{\text{dir}}(t,\omega), \tag{5.62}$$

where † denotes the matrix pseudo-inverse. However, under the assumption that only one source is dominant in each time-frequency bin, we can solve (5.61) by enforcing the sparsity of the resulting coefficients vector. In particular, we obtain $\hat{\boldsymbol{\beta}}(t,\omega)$ as the result of a group lasso optimization problem [293], i.e.,

$$\hat{\boldsymbol{\beta}}(t,\omega) = \underset{\boldsymbol{\beta}(t,\omega)}{\text{argmin}}\ \frac{1}{2}\|\widehat{\mathbf{Y}}(k)\boldsymbol{\beta}(t,\omega) - \hat{\mathbf{x}}_{\text{dir}}(t,\omega)\|_2^2 + \kappa\sum_{n=1}^{N}\|\boldsymbol{\beta}_n(t,\omega)\|_2. \tag{5.63}$$

As shown in [47], this problem can be solved using the alternating direction method of multipliers (ADMM).

Discussion It is worth noticing that, since sources and microphones are assumed to lie on the same plane (i.e., $\theta = \pi/2$), the columns of $\widehat{\mathbf{Y}}_n(k)$ for which $\ell + |\mu|$ is even have been removed. As shown in [2, 230], in fact, when $\theta = \pi/2$ the summation in (5.35) goes to zero since $Y_{\ell\mu}(\pi/2, \phi_{v,n}) = 0$. The truncation order I is usually set as $I = \lceil keR_s/2\rceil$ where e is the Euler's number, $\lceil\cdot\rceil$ is the ceiling operator and R_s is the radius of the region surrounding a source [229]. Hence, the truncation order should be a function of the source but, in order to simplify the notation, in this manuscript we assumed that the truncation order is the same for all the sources. As stated in [216], the radius R_s and, as a consequence, the value of I can be reduced with a suitable choice of the origin of the reference frame. In this manuscript, the origin coincides with the coordinates of the sources location estimates. Moreover, as it is clear from (5.62), in order for the system to be over-determined, the following condition should be satisfied: $\left[(I+1)^2 - T_I\right]N < I$, where $T_I = I(I+1)/2$. However, considering the assumption that only one source is dominant in each time-frequency bin, the above-mentioned condition can be relaxed to $[(I+1)^2 - T_I] < I$. It follows that $I = \min\left(\lceil keR_s/2\rceil, (\sqrt{8I+1}-3)/2\right)$.

5.3.5 Synthesis

Synthesis of the Direct Component

An estimate $\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ of the direct sound component at the vth VM due to the nth source can be obtained by exploiting the model in (5.35). More precisely, given the estimate $\hat{\boldsymbol{r}}'_n$ of the nth source position obtained in (5.16), and the set of exterior field coefficients $\hat{\boldsymbol{\beta}}_n(t,\omega)$ obtained in (5.63), $\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ is obtained through

$$\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v) = \sum_{\ell=0}^{I}\sum_{\mu=-\ell}^{\ell} \hat{\beta}^n_{\ell\mu}(t,\omega) h_\ell\left(k\hat{\check{\rho}}_{v,n}\right) Y_{\ell\mu}\left(\hat{\check{\theta}}_{v,n},\hat{\check{\phi}}_{v,n}\right), \tag{5.64}$$

where $\hat{\check{\rho}}_{v,n}$, $\hat{\check{\theta}}_{v,n}$ and $\hat{\check{\phi}}_{v,n}$ are the estimates of $\check{\rho}_{v,n}$, $\check{\theta}_{v,n}$ and $\check{\phi}_{v,n}$ in (5.35) and can be computed by inserting in (5.34) the estimate $\hat{\boldsymbol{r}}'_n$ of the nth source location.

The term $C_v(\omega)$ in (5.2) models the VMs pick-up pattern. Usually this term can be expressed as a function $f(\cdot)$ that depends on the frequency ω, the angle between the vth VM and the nth source $[\iota_{n,v},\zeta_{n,v}]^T = \angle\boldsymbol{r}'_n - \check{\boldsymbol{r}}_v$, with ι the azimuth and ζ the elevation, the orientation $\check{\boldsymbol{o}}_v$ of the vth VM,

$$C_v(\iota_{n,v},\zeta_{n,v},\check{\boldsymbol{o}}_v,\omega) = f\left(\iota_{n,v},\zeta_{n,v},\check{\boldsymbol{o}}_v,\omega\right). \tag{5.65}$$

The proposed framework enables also the estimation of the direct signal component acquired by a higher order VM located at $\check{\boldsymbol{r}}_v$. In particular, given the estimate $\hat{\boldsymbol{\beta}}_n(t,\omega)$ of the exterior sound field produced by the nth source, the spherical harmonics coefficients acquired by a Z-order VM can be directly obtained through the spherical harmonics addition theorem [163] through

$$\begin{gathered}\gamma^{v,n}_{zb}(t,\omega) = \sum_{l=0}^{I}\sum_{\mu=-l}^{l} \hat{\beta}^n_{l\mu}(t,\omega) T^{\mu b}_{lz}(k,\hat{\check{\rho}}_{v,n},\hat{\check{\theta}}_{v,n},\hat{\check{\phi}}_{v,n}),\\ z=0,\dots,Z, \quad b=-z,\dots,z\end{gathered} \tag{5.66}$$

where $T^{\mu b}_{lz}(k,\hat{\check{\rho}}_{v,n},\hat{\check{\theta}}_{v,n},\hat{\check{\phi}}_{v,n})$ is defined according to (4.4). The knowledge of the spherical harmonics coefficients acquired by the Z-order microphone enables applications like, for example, modal beamforming [292], binaural synthesis [36, 218] etc. In all these applications the spherical harmonics coefficients of the higher order microphone are filtered through

$$C_v(\omega)S_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v) = \sum_{z=0}^{Z}\sum_{b=-z}^{z} [\psi^{v,n}_{zb}(\omega)]^* \, \gamma^{v,n}_{zb}(t,\omega), \tag{5.67}$$

where $\psi^{v,n}_{zb}(\omega)$ are the spherical harmonics filter coefficients [217].

Synthesis of the Diffuse Component

In order to synthesize the diffuse component $S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$ of the VM, we use the estimates of the diffuse signal at each microphone obtained in (5.57). More precisely, given the estimates $\hat{X}_{\text{diff}}(t,\omega,\boldsymbol{r}_i)$, we compute the power of the diffuse signal component in $\check{\boldsymbol{r}}_v$ as [260]

$$E\{|S_{\text{diff}}(t,\omega,\check{\boldsymbol{r}}_v)|^2\} = \sum_{i=1}^{I} \varpi_i(\omega) E\{|\hat{X}_{\text{diff}}(t,\omega,\boldsymbol{r}_i)|^2\}, \tag{5.68}$$

where $\sum_{i=1}^{I} \varpi_i(\omega) = 1$. As stated in [260], $\varpi_i(\omega)$ are real valued weights that depend on the estimation variance of the power of the signal at the ith microphone (i.e., $E\{|X(t,\omega,\boldsymbol{r}_i)|^2\}$), $i = 1,\ldots,I$. Since these estimates are usually unavailable, we choose the weights to be inversely proportional with respect to the distance between the vth VM and the ith microphone, i.e.,

$$\varpi_i(\omega) = \frac{1}{\|\boldsymbol{r}_i - \check{\boldsymbol{r}}_v\|_2} \left(\sum_{i'=1}^{I} \frac{1}{\|\boldsymbol{r}_{i'} - \check{\boldsymbol{r}}_v\|_2} \right)^{-1}. \tag{5.69}$$

For what concerns the estimation of the phase of the diffuse signal component, we experimentally verify that plausible results can be achieved by using as an estimate the phase of the nearest microphone. It is worth noting that such a solution does not provide the desired spatial coherence among closely spaced VMs. It follows that, if one want to further process the signal coming from different VMs, this fact must be taken into account. Nevertheless, as we will see in the following section, this solution is applicable to coincident VMs for simulating, for example, a stereo recording scenario.

Finally, as defined in (5.2), the function $Q_v(\omega)$ controls the sensitivity of the VM to the diffuse field. In the most general case, this function can be arbitrarily designed. However, in most cases the relationship between $Q_v(\omega)$ and the pick-up pattern of the VM directivity is well approximated by

$$Q_v(\omega) = \sqrt{\frac{1}{4\pi} \int_0^{\pi} \int_0^{2\pi} C_v(\iota,\zeta,\check{\boldsymbol{o}}_v,\omega)^2 \sin\zeta \, d\iota \, d\zeta}. \tag{5.70}$$

The expression in (5.70) is valid under the assumption of a spherically isotropic sound field and accounts for the fact that the diffuse component of the VM can be attenuated depending on the pick-up pattern $C_v(\iota,\zeta,\check{\boldsymbol{o}}_v,\omega)$.

5.3.6 Validation and Results

The proposed virtual miking technique is suitable for many applications thanks to its flexibility in terms of setup configuration. In addition, the possibility of completely characterizing the VM and the intrinsic source model paves the way to the use in advanced spatial audio applications. In order to validate the virtual miking performance, we tested the system through an extensive software simulation campaign. This allows us to control both the characteristics of the sources, such as their directivity pattern and the number and the location of the virtual microphones. Moreover, this eases the development of test cases that require a significant amount of reference VMs, barely

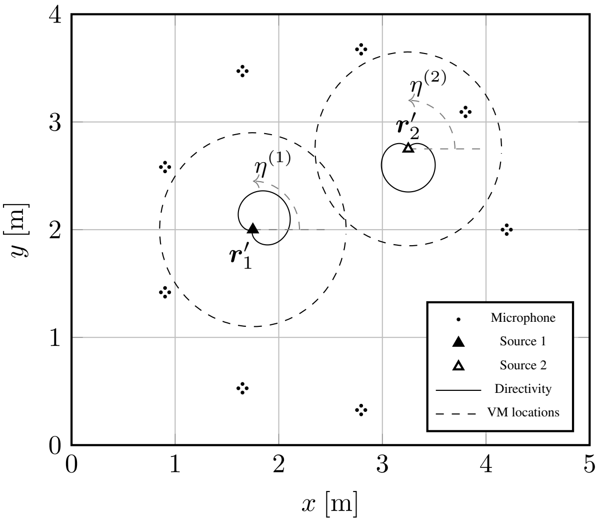


Figure 5.10: *2D graphical representation of the first simulation setup. The two sources have a first-order cardioid directivity and they are located in* $\boldsymbol{r}_1' = [1.75, 2]^T\mathrm{m}$ *and* $\boldsymbol{r}_2' = [3.25, 2.75]^T\mathrm{m}$

deployable in practice. The simulation setup is introduced in Section 5.3.6, while in Section 5.3.6 the different metrics used for the assessment of the VM performance are defined. The simulation results are discussed in Sec 5.3.6.

Simulation Setup and Parameters

The simulation setup is illustrated in Figure 5.10. It consists of $A = 9$ circular microphone arrays with radius $0.04\,\mathrm{m}$, accommodating $M = 4$ omnidirectional microphones each. Therefore, the total number of microphones is $I = A \times M = 36$. The sources emit two speech signals simultaneously (female and male taken from [264]), at $\boldsymbol{r}_1' = [1.75, 2]^T\mathrm{m}$ and $\boldsymbol{r}_2' = [3.25, 2.75]^T\mathrm{m}$, respectively. When a single source setup is considered, only the source in $\boldsymbol{r}_1'$ is active in the scene at any time. The two sources present a first-order cardioid directivity with looking direction (i.e., direction of maximum energy emission) equal to $45°$ and $270°$, respectively. A set of $V = V^{(1)} + V^{(2)}$ omnidirectional VMs (i.e., $C(\cdot) = 1$) are placed on two circumferences of radius R centered around the two sources (see Figure 5.10). In detail, the positions of the VMs are defined as

$$\begin{aligned}\check{\boldsymbol{r}}_v^{(1)} &= R[\cos\eta_v^{(1)}, \sin\eta_v^{(1)}]^T + \boldsymbol{r}_1', \quad v = 1, \dots, V^{(1)} \\ \check{\boldsymbol{r}}_v^{(2)} &= R[\cos\eta_v^{(2)}, \sin\eta_v^{(2)}]^T + \boldsymbol{r}_2', \quad v = 1, \dots, V^{(2)},\end{aligned} \tag{5.71}$$

where $\eta_v^{(1)} = 2\pi/V^{(1)}v$, $\eta_v^{(2)} = 2\pi/V^{(2)}v$ and $\check{\boldsymbol{r}}_v^{(1)}$, $\check{\boldsymbol{r}}_v^{(2)}$ are the positions of VMs surrounding the first and the second source, respectively. The actual number of VMs, V, depends on the simulation setup. This arrangement of the VMs allows us to capture the directional properties of the sources, testing the capability of the proposed virtual miking technique in rendering a spatial sound perception coherent with the VM position in the scene.

The signal at the microphones (see (5.38)) is simulated as the convolution between the signal of the sources and the room impulse response (RIR) computed through the image source method [11] implemented in [121]. The room is $5\,\mathrm{m} \times 4\,\mathrm{m} \times 3\,\mathrm{m}$ with a

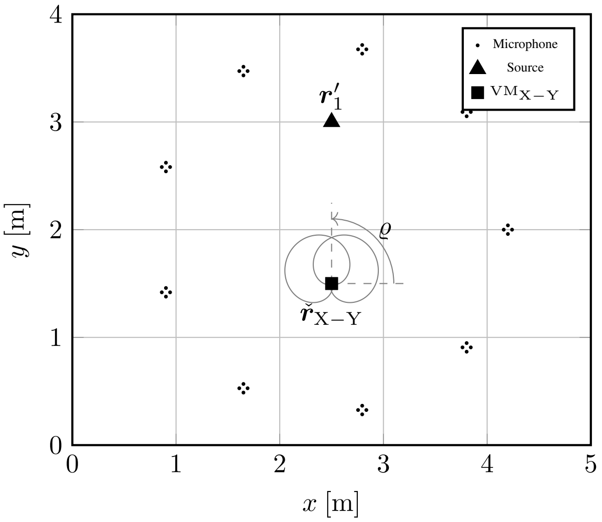


Figure 5.11: *2D graphical representation of the second simulation setup. The source is located at* $\boldsymbol{r}'_1 = [2.5, 3]^T\mathrm{m}$ *while the VMs in X-Y configuration are placed at* $\check{\boldsymbol{r}}_{\mathrm{X-Y}} = [2.5, 1.5]^T\mathrm{m}$.

reverberation time of $T_{60} = 0.4\,\mathrm{s}$. As done in [241] and [212], the diffuse component in (5.38) is computed from the RIR late reverberation part. This can be accomplished suppressing the direct path and the early reflections using a cutoff time T_e set to a typical value of $0.05\,\mathrm{s}$ [145]. The additive noise component in (5.38) is simulated using a random white Gaussian noise, whose variance is set so that the desired signal to noise ratio at each sensor is $60\,\mathrm{dB}$. The signals are processed at a sampling rate of $16\,\mathrm{kHz}$ and their time-frequency representation is obtained through a 4096 points Short Time Fourier Transform (STFT) with a $0.256\,\mathrm{s}$ Hamming window adopted both in the analysis and synthesis phase and $87.5\,\%$ overlap.

The localization is performed as described in Section 5.3.2, where, for the computation of the pseudospectrum $\lambda^{(a)}(\alpha, \omega), \forall a = 1, \dots, A$, we adopted the super directive beamformer [268]. The averaged pseudospectum in (5.8) is compute with $W = 4096$. The actual estimate of the source location $\hat{\boldsymbol{r}}'_n$ is obtained as the median value of 1000 RANSAC executions with different algorithm initializations at each time frame t. An estimate of the direct and diffuse components are obtained as presented in Section 5.3.3 with $\lambda = 0.68$, $\nu = 1.3$ and $G_{\min} = -30\,\mathrm{dB}$. Given the described setup, the pairs of microphones for the estimation of the cross spectra $\hat{\Phi}_{ii'}(t, \omega)$ are chosen as belonging to the same array by following their order in terms of azimuth with respect to each array center. For what concerns the source parameter estimation, we set the spherical harmonics expansion order in (5.35) according to the discussion in Section 5.3.4. Moreover, an applicative scenario regarding a spatial acquisition has been simulated.

We test the virtual miking technique in the context of a stereo recording, simulating a X-Y stereo miking setup. In Figure 5.11 one omnidirectional source, aligned with the X-Y VM along the y axis, is present. The source emits a female speech signal similarly to the previous scenario. X-Y stereo recording requires the employment of two directional microphone ($\mathrm{VM_I}$ and $\mathrm{VM_R}$) with first-order cardioid pick-up pattern. The microphones are characterized by coincident location $\check{\boldsymbol{r}}_{\mathrm{I}} = \check{\boldsymbol{r}}_{\mathrm{R}} = \check{\boldsymbol{r}}_{\mathrm{X-Y}}$ and different pick-up pattern orientation (5.65), such that $|\check{\boldsymbol{o}}_{\mathrm{I}} - \check{\boldsymbol{o}}_{\mathrm{R}}| = [\pi/2, 0]^T\mathrm{rad}$.

Hence, given the X-Y looking direction $\check{\boldsymbol{o}}_{\mathrm{X-Y}} = [\varrho, 0]^T$, namely, the direction corresponding to the center of the stereo plane, the orientation of the VMs are defined as $\check{\boldsymbol{o}}_{\mathrm{I}}(\varrho) = [\varrho - \pi/4, 0]^T \mathrm{rad}$ and $\check{\boldsymbol{o}}_{\mathrm{R}}(\varrho) = [\varrho + \pi/4, 0]^T \mathrm{rad}$. Notice that when it is not explicitly stated, we assume the same setup parameters of the previous scenario.

Metrics

Accordingly to the different simulation setups, we evaluated the virtual miking performance in terms of a set of metrics that are generalized power directivity (GPD), signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), direct-to-reverberant ratio (DRR) and interchannel level difference (ILD). The mathematical expression of the metrics is given in the following.

Generalized Power Directivity (GPD) The GPD of a source is defined as

$$\widehat{\mathrm{GPD}}(\eta_v^{(n)}) = \frac{\sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t, \omega_w, \check{\boldsymbol{r}}_v^{(n)})|^2}{\max\limits_v \sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t, \omega_w, \check{\boldsymbol{r}}_v^{(n)})|^2}, \tag{5.72}$$

and it measures, for each source, the normalized power of the estimated VM signals surrounding the given source as a function of the VMs angle $\eta_v^{(n)}$ in (5.71). In the following we indicate with $\mathrm{GPD}(\eta_v^{(n)})$ the same metric computed with reference signals.

Signal-to-Distortion Ratio (SDR) The SDR is defined as the ratio between the desired reference signal and the distortion that affects the estimation (i.e. interference, noise and artifacts). We adopt the SDR estimator of [275] for the evaluation of the estimated VM signals with respect to the reference VM signals.

Signal-to-Interference Ratio (SIR) The SIR is defined as the ratio of the power of direct component of a desired signal and the sum of the interfering signals. More precisely,

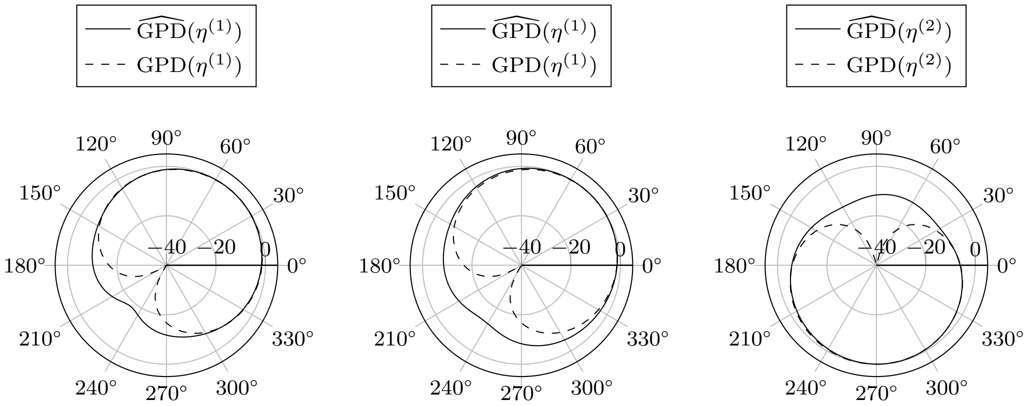
$$\widehat{\mathrm{SIR}}(\eta_v^{(n)}) = \frac{\sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t, \omega_w, \check{\boldsymbol{r}}_v^{(n)})|^2}{\sum_{\bar{n} \neq n} \sum_w \sum_t |\hat{S}_{\bar{n},\mathrm{dir}}(t, \omega_w, \check{\boldsymbol{r}}_v^{(n)})|^2}. \tag{5.73}$$

In the following we indicate with $\mathrm{SIR}(\eta_v^{(n)})$ the same metric computed with reference signals.

Direct-to-Reverberant Ratio (DRR) The DRR represents the ratio between the power of the direct and reverberant components

$$\widehat{\mathrm{DRR}}(\eta_v^{(n)}) = \frac{\sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t, \omega_w, \check{\boldsymbol{r}}_v^{(n)})|^2}{\sum_w \sum_t |\hat{S}_{\mathrm{diff}}(t, \omega_w, \check{\boldsymbol{r}}_v^{(n)})|^2}. \tag{5.74}$$

This metrics allows us to evaluate the VM in terms of spatial sound characteristics, since the DRR reflects the spatial properties of the signal. In the following we indicate with $\mathrm{DRR}(\eta_v^{(n)})$ the same metric computed with reference signals.



(a) (b) (c)

Figure 5.12: *(a)* GPD *of the VMs and their references in a single source scenario. (b)* GPD *of the first source when both are active. (c)* GPD *of the second source when both are active.*

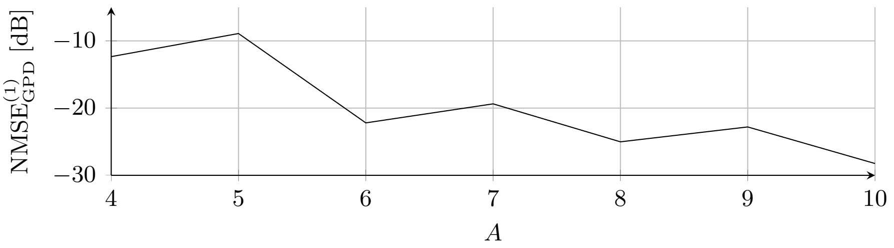


Figure 5.13: *The* $\mathrm{NMSE}_{\mathrm{GPD}}$ *as a function of the number of arrays* A.

Interchannel Ievel Difference (ILD) The ILD is defined as the ratio between two VMs in a stereo configuration

$$\widehat{\mathrm{ILD}}(\varrho) = \frac{\sum_w \sum_t |\hat{S}(t, \omega_w, \check{\boldsymbol{r}}_\mathrm{I}, \check{\boldsymbol{o}}_\mathrm{I}(\varrho))|^2}{\sum_w \sum_t |\hat{S}(t, \omega_w, \check{\boldsymbol{r}}_\mathrm{R}, \check{\boldsymbol{o}}_\mathrm{R}(\varrho))|^2}. \tag{5.75}$$

where ϱ is the $\mathrm{VM}_{\mathrm{X-Y}}$ azimuth orientation. In (5.75) the dependence on the VM orientation is made explicit. In the following we indicate with $\mathrm{ILD}(\varrho)$ the same metric computed with reference signals.

As regards the GPD, SIR and DRR, we also evaluate the estimation in terms of the normalized mean squared error (NMSE). For instance concerning the GPD it is defined as

$$\mathrm{NMSE}^{(n)}_{\mathrm{GPD}} = 10\log_{10} \frac{\sum_v |\widehat{\mathrm{GPD}}(\eta^{(n)}_v) - \mathrm{GPD}(\eta^{(n)}_v)|^2}{\sum_v |\,\mathrm{GPD}(\eta^{(n)}_v)|^2}. \tag{5.76}$$

Moreover, for all the defined metrics we show the results in a decibel scale defined as $10\log_{10}(\cdot)$ of the relative metrics.

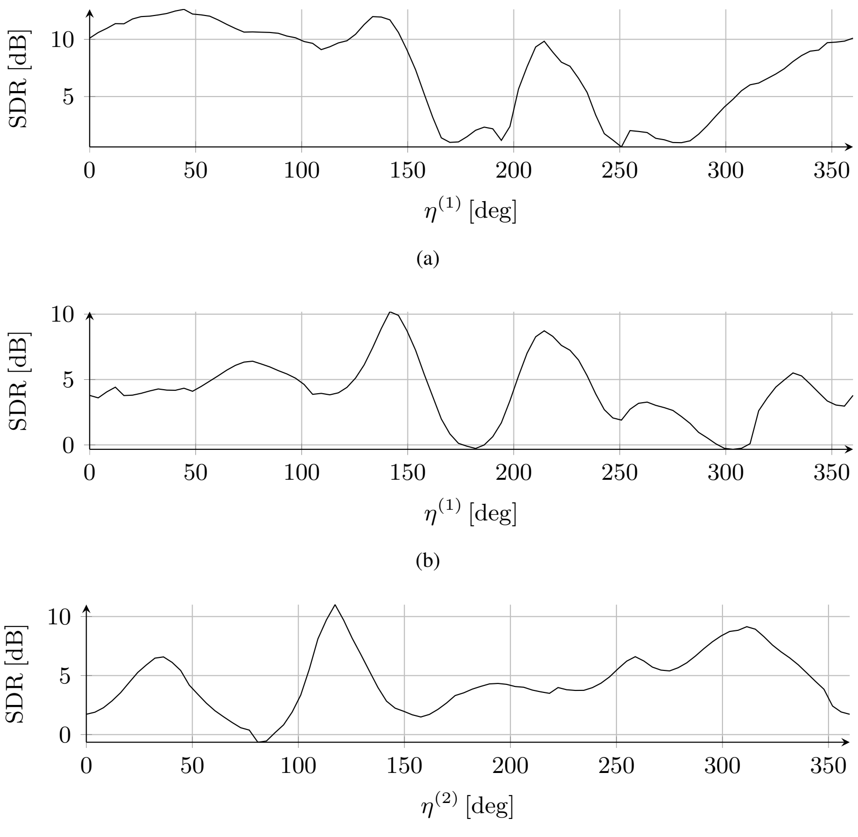
Results

Single-source scenario The single source scenario is simulated using the setup of Figure 5.10, when only the first source in $\boldsymbol{r}'_1$ is active. The estimated source position is $\hat{\boldsymbol{r}}'_1 = [1.7372, 2.0152]^T\mathrm{m}$ giving a localization error of $\|\hat{\boldsymbol{r}}'_1 - \boldsymbol{r}'_1\| = 0.0198\,\mathrm{m}$. The number of VMs employed is $V = V^{(1)} = 90$, equally spaced around the source at distance from the source of $R = 0.9\,\mathrm{m}$. The $\widehat{\mathrm{GPD}}$, computed using the synthesized VM signals is reported in Fig 5.12, compared to the GPD computed with the reference signals. More specifically, Figure 5.12 plots $\widehat{\mathrm{GPD}}$ and GPD in three different cases: a) the GDP of the source in $\boldsymbol{r}'_1$ when it is the only active source; b) the GDP of the source in $\boldsymbol{r}'_1$ when both sources are active; c) the GDP of the source in $\boldsymbol{r}'_2$ when both sources are active. In the context of the single source scenario we focus on Figure 5.12(a). Notice that the $\widehat{\mathrm{GPD}}\left(\eta^{(1)}\right)$ fits the general trend of GPD with a $\mathrm{NMSE}^{(1)}_{\mathrm{GPD}}$ with respect to the reference of $-22.8\,\mathrm{dB}$. In order to evaluate the influence of the number of arrays A on the accuracy of the estimated source directivity, we performed a set of simulations varying $A \in \{4, 10\}$ with respect to the single-source scenario setup. In particular, the arrays are equally distributed on a circumference of radius $1.7\,\mathrm{m}$ centered in the room so that the setup with $A = 9$ corresponds to the one reported in Figure 5.10. In Figure 5.13 the $\mathrm{NMSE}^{(1)}_{\mathrm{GPD}}$ is reported as a function of the number of the arrays A. As expected, the $\mathrm{NMSE}^{(1)}_{\mathrm{GPD}}$ decreases as A increases since a greater number of arrays guarantees a wider angle coverage. Additionally, from the inspection of Figure 5.13 we can notice that between $A = 6$ and $A = 10$ the difference is around $6\,\mathrm{dB}$ in response to an increase of 16 microphones. The SDR of the VMs is reported in Figure 5.14 for the same three cases adopted for GDP. The single source scenario is shown in Figure 5.14(a). We can notice here that for the locations where the VM mostly picks up the diffuse component, the SDR estimation is less consistent. Hence, the performance is affected by the VM location and by the averaging of the diffuse estimates of the arrays (Section 5.3.5).

Figure 5.15(a) reports the comparison between $\widehat{\mathrm{DRR}}$ and DRR. Note that $\widehat{\mathrm{DRR}}$ is comparable with the DRR profile of an ideal source with first-order cardioid directivity pattern. The ideal DRR reaches $-\infty$ in correspondence of the zero in the source cardioid directivity pattern since no direct component is propagated in this direction. The $\widehat{\mathrm{DRR}}$ does not follow this behavior due to the fact that the GPD is not exactly zero in this direction (see Figure 5.12(a)). Moreover, we provide as a comparison the estimated $\widehat{\mathrm{DRR}}_{\mathrm{omni}}$ obtained assuming an omnidirectional sound source. More specifically, we estimated the exterior sound field coefficients (Section 5.3.4) by setting the spherical harmonic order $L = 0$ in (33) and consequently, the synthesized direct signal in (38) presents an omnidirectional characteristic. Inspecting Figure 5.15(a), we can notice that the addition of the source directivity greatly enhances the estimate of $\widehat{\mathrm{DRR}}$ giving a $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}}$ of $-14.3\,\mathrm{dB}$, while the $\widehat{\mathrm{DRR}}_{\mathrm{omni}}$ does not follow the actual trend of the reference giving a $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}_{\mathrm{omni}}}$ of $-1.14\,\mathrm{dB}$.

Two-sources scenario We evaluate the virtual miking technique in a double talk scenario, where both acoustic sources of Figure 5.10 are simultaneously active. The estimated position of the sources are $\hat{\boldsymbol{r}}'_1 = [1.8054, 2.0518]^T\mathrm{m}$ and $\hat{\boldsymbol{r}}'_2 = [3.3556, 2.7087]^T\mathrm{m}$ giving a localization error of $0.0759\,\mathrm{m}$ and $0.1134\,\mathrm{m}$, respectively. We simulate $V = 180$ omnidirectional VMs, equally distributed around the sources with $V^{(1)} = V^{(2)} = 90$ VMs for each source placed accordingly to Figure 5.10.

Both $\widehat{\mathrm{GDP}}$ and GDP are reported in Figure 5.12(b) and Figure 5.12(c). For both sources we can notice that $\widehat{\mathrm{GDP}}$ agrees with GDP suggesting that the spatial radiation



(a)

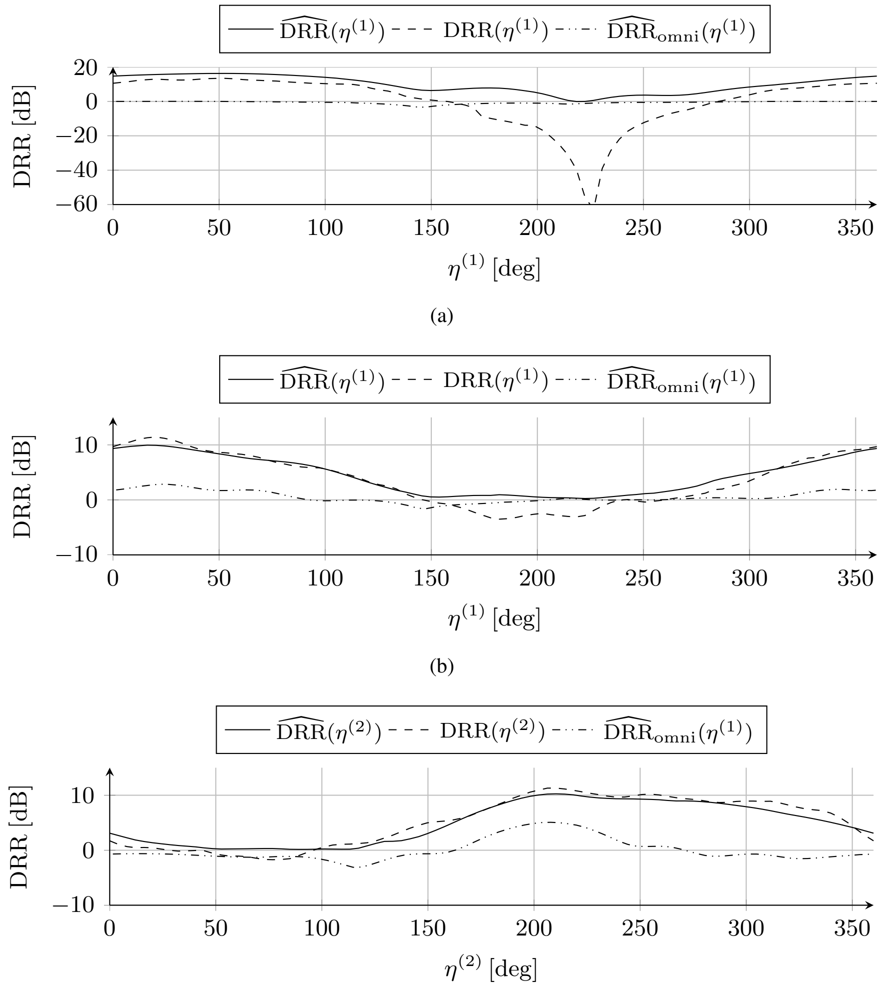
(b)

(c)

Figure 5.14: *(a)* SDR *value referred to a single source scenario. (b)* SDR *value of the first source when both are active. (c)* SDR *value of the second source when both are active. Note that* $\eta^{(n)}$ *refers to the angle with respect to the* n*th source and the location of the VMs according to* (5.71).

characteristics of the sources are correctly reconstructed at the VMs. Another important measure of the sound field spatial characteristics is the SIR. As shown in Figure 5.16, the $\widehat{\mathrm{SIR}}$ follows the behaviour of the actual SIR for a wide range of directions. However, it is worth noting that the SIR goes to $-\infty$ for the directions in correspondence of the zeros in the directivity pattern of the sources. Since the behavior of $\widehat{\mathrm{SIR}}$ is related to the $\widehat{\mathrm{GPD}}$s of the two sources shown in Figure 5.12(b) and Figure 5.12(c), respectively it cannot reach $-\infty$. In fact, the $\widehat{\mathrm{GPD}}$ is not exactly zero for such directions. Similarly to what is done for the *single-source scenario* in Figure 5.15(a), we include in Figure 5.16 the $\widehat{\mathrm{SIR}}_{\mathrm{omni}}$ estimated when the two sources are incorrectly assumed as omnidirectional, i.e. setting $L = 0$ in (33). As expected, the estimation of the SIR effectively improves by modelling the source directivity explicitly. In particular, the $\mathrm{NMSE}^{(1)}_{\mathrm{SIR}_{\mathrm{omni}}}$ is equal to $5.3\,\mathrm{dB}$ for the first source (Figure 5.16(a)) and the $\mathrm{NMSE}^{(2)}_{\mathrm{SIR}_{\mathrm{omni}}}$ is equal to $-2.4\,\mathrm{dB}$ for the second source (Figure 5.16(b)). The $\mathrm{NMSE}^{(1)}_{\mathrm{SIR}}$ and $\mathrm{NMSE}^{(2)}_{\mathrm{SIR}}$ drops to $-12.4\,\mathrm{dB}$ and $-12.6\,\mathrm{dB}$, respectively when the source directivity is taken into account.

As far as the SDR is concerned, results are reported in Figure 5.14(b) and Figure 5.14(c). Similarly to the single source scenario, the VM performance is influenced by the approximation of the diffuse sound field, with lower SDR values when the dif-



(a)

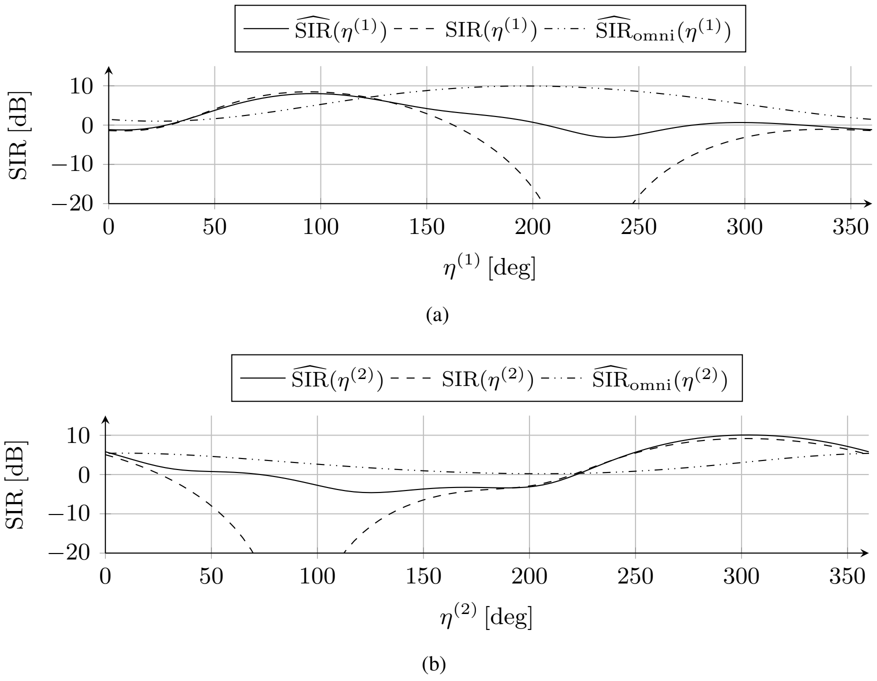
(b)

(c)

Figure 5.15: *(a) Estimated* DRR *and its reference in a single source scenario. (b) Estimated* DRR *and its reference of the first source when both sources are active. (c) Estimated* DRR *and its reference of the second source when both sources are active. The subscript* omni *refers to an estimate obtained assuming an omnidirectional source directivity (i.e.,* $L = 0$*). Note that* $\eta^{(n)}$ *refers to the angle with respect to the* n*th source and the location of the VMs according to* (5.71).

fuse component is mainly present and higher values when the VM is close to an array used for analyzing the sound scene.

Finally, the DRR of the VMs is provided in Figure 5.15(b) and Figure5.15(c). Analogously to the single source scenario, the $\widehat{\mathrm{DDR}}$ follows the reference value achieving a $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}}$ and a $\mathrm{NMSE}^{(2)}_{\mathrm{DRR}}$ of $-14.3\,\mathrm{dB}$ and $-14.7\,\mathrm{dB}$, respectively. As a comparison, also in Figure 9(b)and Figure 9(c) the $\widehat{\mathrm{DDR}}_{\mathrm{omni}}$, estimated assuming two omnidirectional sources, are reported. Both the $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}_{\mathrm{omni}}}$ and the $\mathrm{NMSE}^{(2)}_{\mathrm{DRR}_{\mathrm{omni}}}$ are equal to $-1.7\,\mathrm{dB}$. Such values are approximately $13\,\mathrm{dB}$ higher than the the ones obtained by considering the directivity of the sources. Generally, we can notice that the behavior of the DRR is mainly determined by the directivity of the sound source. Iow DRR values in Figure 5.15(b) and Figure 5.15(c) are associated to locations where



(a)

(b)

Figure 5.16: *(a)* SIR *value referred to the first source compared to its reference. (b)* SIR *value referred to the second source compared to its reference.The subscript* omni *refers to an estimate obtained assuming an omnidirectional source directivity (i.e., $L = 0$). Note that $\eta^{(n)}$ refers to the angle with respect to the nth source and the location of the VMs according to* (5.71)*.)*

the direct component energy is lower than the diffuse component due to the directivity pattern of each source.

Stereo recording scenario The X-Y stereo microphone is commonly adopted for spatial sound acquisition. The spatial characteristics of the sound field are rendered in the stereo recording through the sound pressure level difference between the two directional microphones. Therefore, we adopted the ILD (5.75) as a metric to evaluate the ability of the VM in reproducing the spatial features of the stereo setup. The ILD is computed in the scenario of Figure 5.11, as a function of the $\mathrm{VM}_{\mathrm{X-Y}}$ azimuth orientation ϱ varying from $0°$ to $360°$. The omnidirectional source is localized in $\hat{\boldsymbol{r}}' = [2.5316, 2.9707]^T\mathrm{m}$ with a localization error of $0.0431\,\mathrm{m}$. Inspecting the curves of Figure 5.17, we can notice an high agreement between the $\widehat{\mathrm{ILD}}$ computed with the estimated signals and the

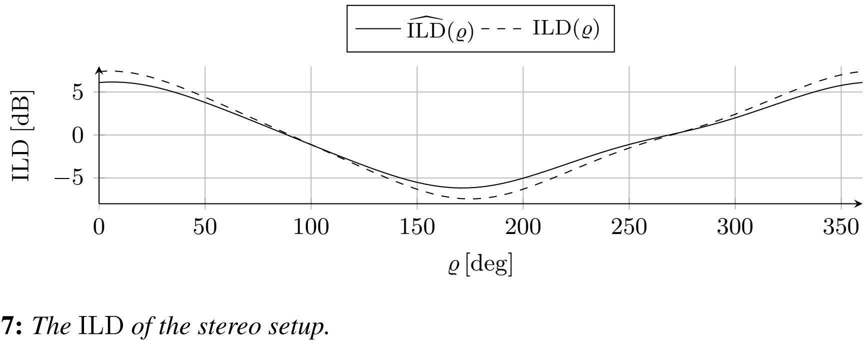


Figure 5.17: *The* ILD *of the stereo setup.*

ILD computed with the reference signals. In fact the overall $\mathrm{NMSE}_{\mathrm{ILD}} = -15.4\,\mathrm{dB}$. The $0\,\mathrm{dB}$ value is crossed around $90°$ and $270°$, when the sound pressure level at VM_{L} and VM_{R} is equivalent.

This coincides with the source located at the center of the stereo plane. In an anechoic scenario, the maximum of the ILD function would occur around $315°$, in correspondence of the zero in the VM_{R} cardioid pattern and the minimum around $225°$, when the VM_{L} amplitude is attenuated. However in Figure 5.17, we cannot observe this behavior due to the presence of the diffuse field. Indeed this does not let the signal power of the two microphones in the X-Y configuration go exactly to zero.

The proposed virtual miking procedure aims therefore to provide a promising tool for a wide variety of spatial sound applications. In particular, we showed that the spatial sound characteristics e.g. the DRR, the SIR or the ILD can be effectively approximated by the synthesized VM signal. Moreover, the explicit modeling of the sound source directional characteristics improves the VM estimation especially in terms of its spatial cues with respect to the omnidirectional source model commonly adopted in the literature. Thanks to the possibility of synthesizing the VMs signals in arbitrary locations, the proposed techniques potentially enables a listener to *virtually navigate* a recorded sound field with *six-degree of freedom*. Therefore, the procedure is particularly interesting for EAR framework, where capturing the sound field spatial features is a relevant aspect in order to provide an immersive user experience.

CHAPTER 6

Multichannel Blind Source Separation

In this chapter, we introduce a blind sound source separation (BSS) technique based on a non-parametric sound field representation. We enable the manipulation of the acquired sound field, by performing the separation of the target source from a mixture of acoustic sources.

Differently from the parametric techniques introduced in the previous chapter, here we aim at processing the sound field with a more flexible setup. As a matter of fact, we introduce a blind source separation technique, for extended linear arrays. Being blind, the separation process does not require a priori information on the source location. Additionally, the setup is limited and it is composed of a single extended uniform linear microphone array.

We adopt a particular non-parametric sound field representation in order to improve the separation, exploiting the information related to the different locations of the sources that are inherently represented in the data.

The BSS technique concerns two phases.

First, the multichannel signal, acquired by the microphone array, is mapped onto a domain known as ray space [37]. This representation is obtained through a beamforming-like operation that project the signals in a domain where each point represents a ray. Thanks to a proper parametrization of the ray space data, the location of acoustic sources is reflected in the data enabling to tackle problems such as data-association using pattern analysis techniques [37, 159, 160].

In Section 6.1, we review the linear operator devoted to the projection of the multichannel signal of a microphone array onto the ray space known as ray space transform (RST) [37]. Moreover, we introduce a fast implementation of the RST, based on the formulation of the transformation as a non uniform Fourier transform.

The second phase of the BSS technique performs the separation of the source signals through the processing of the ray space data.

A popular approach to BSS is the nonnegative matrix factorization (NMF) [148, 276], which consists of a data decomposition technique factorizing a nonnegative input matrix into a sum of rank-1 components. Traditionally, the input matrix is the magnitude or power spectrogram of a sound signal, so that each of the components corresponds to elementary spectral patterns with a given temporal amplitude modulation.

While standard NMF was only suitable for separating single-channel mixtures, multichannel extensions can be obtained by stacking channels into one matrix structure [191] or by considering a parallel factor (PARAFAC) model [100].

One of the well-known problems of these extensions is that they only take advantage of the amplitude information within the mixing system. Consequently, spatial information is not exploited when the observed mixtures come from a microphone array recording. To overcome this limitation, Lee et. al. [150] proposed a beamspace data model (BS-NMF) that considers the projection of the input signal onto a set of steered directions while accounting for the inherent phase-difference information that is present in this type of recordings. While providing good results, the beamspace model relies on the assumption that the incident wavefronts will be planar at the array (far field), leaving room for improvement in more general scenarios.

Due to its close relation to the proposed ray space solution, in Section 6.2 in addition to the customary multichannel NMF (MNMF) formulation, we provide a review of the BS-MNMF [150].

In the literature, other approaches to the modeling of the spatial information have been presented. They rely on the signal representation of the spatial covariance matrix (SCM) [189, 232]. For each time-frequency (TF) point in the Short-Time Fourier Transform (STFT), the SCM represents the mixing of the sources by magnitude correlations and phase differences between channels. Authors in [232] proposed to estimate unconstrained SCM mixing filters together with a NMF magnitude model to identify and separate repetitive frequency patterns corresponding to a single spatial location.

To mitigate spatial aliasing effects and under far-field assumption, [180] proposed a SCM model based on DoA kernels to estimate the inter-microphone time delay given a looking direction. The method proposed in [57], here referred to DoA-MNMF, uses a SCM kernel-based model where the mixing filter is decomposed into two direction-dependent SCMs to represent and estimate both time and level differences between array channels disjointly. However, these full-rank spatial models suffers from both high computational cost and strong sensitivity to parameter initialization. Under moderate echoic conditions, SCMs can be restricted to be rank-1 in a determined scenario, merging independent vector analysis (IVA) and NMF within a framework called independent low-rank matrix analysis (ILRMA) [135].

Several studies have recently proposed to restrict the SCMs of sources to jointly diagonalize the full-rank matrices for multichannel blind source separation [127, 242]. The technique in [242] is commonly referred to as FastMNMF. While FastMNMF [242] projects the signals with an optimizable transform matrix, the authors in [172] adopt a fixed projection, namely, a discrete Fourier transform (DFT) matrix. This transformation acts as the diagonalizer under the DoA kernel based model from [180] projecting the signals into the wave number domain. Thus [172] is here referred to WN-MNMF.

In Section 6.3, we propose to adopt a fixed transformation matrix, the RST [37],

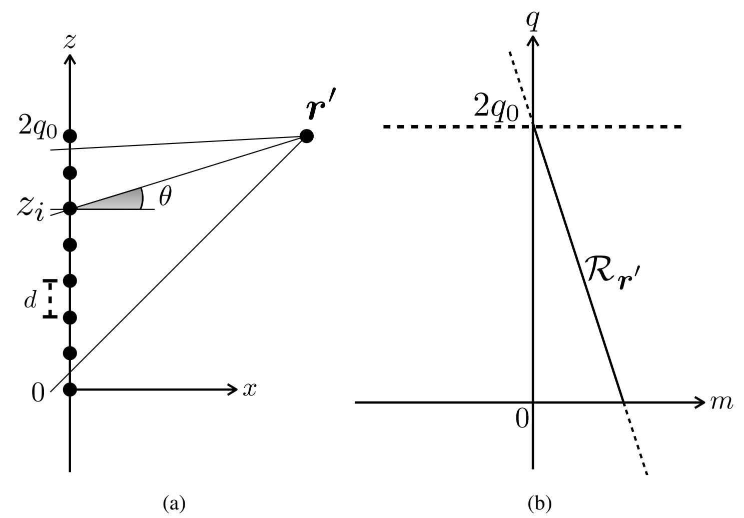


Figure 6.1: *(a) The uniform linear array setup.(b) The representation $\mathcal{R}_{\boldsymbol{r}'}$ of the point source located at $\boldsymbol{r}'$ in the ray space.*

in order to project the signals of a uniform linear array (ULA) onto the ray space domain [160]. In the ray space domain, the geometric location of the sources is directly encoded in the magnitude of the ray-space-transformed signals [37, 159]. Therefore, this allows us to overcome the limited location description given by DoA-based representations. Although the proposed scheme can be combined with other MNMF approaches from the literature modeling the phase information, in this thesis, we used the amplitude-based multiplicate update (MU) algorithm from [189] for an investigation into the effectiveness of our scheme and to limit the computational costs. The ray-space-transformed data leads to more efficient exploitation of the spatial information contained in the recordings while enabling a direct application of the conventional multichannel NMF algorithm.

6.1 Mapping in the Ray Space

6.1.1 Review of the Ray Space Transform

This section reviews the ray space transform (RST) presented in [37]. The RST is defined in order to map the multichannel signal acquired by a linear array of microphones onto the ray space domain. The main advantage of this domain is that acoustic primitives, such as sound sources and reflectors, are mapped to linear patterns (see Figure 6.1). Therefore, pattern analysis algorithms have been used in order to solve problems like source localization, and geometry inference.

The RST can be interpreted as a local Fourier transform, in which a spatial window function is employed and the similarity between the multichannel signal and modulated and shifted copies of the window is computed. In practice, this is implemented as a beamforming operated on a portion (windowed) of the array signals through which we extract the directional components of the sound field on a set of directions. The selection of the analyzed directions characterizes the final representation of the data. In [37], the set of directions is chosen according to the ray space parametrization. Therefore, a direction is described by the concept of ray that is defined by the parameters of the line on which the ray lies on. The Euclidean line equation was adopted in [160] in order to provide the definition of any line in the x-z plane, with the exception of the ones parallel to the z axis as

$$z = mx + q, \tag{6.1}$$

where $m = \tan\theta$ and q are the parameters that describe the lines and they form the ray space. In particular, here we assume that rays are coming from the half-space given by the positive x-axis.

Let us consider a uniform linear array of I microphones lying on the z-axis between 0 and $2q_0$ as shown in Figure 6.1(a). The locations of the microphones are defined as

$$\boldsymbol{r}_i = [x_i, z_i]^T = \left[0, \left(i - \frac{I-1}{2}d\right)\right], \quad i = 0, \dots, I-1, \tag{6.2}$$

where d is the distance between two adjacent microphones. The RST of the multichannel signal of the array is defined as [37]

$$[\mathbf{Z}]_{l,w}(\omega) = d \sum_{i=0}^{I-1} P(z_i, \omega)\, e^{-\frac{jkz_i m_w}{\sqrt{1+m_w^2}}} \psi_{i,l}^* \tag{6.3}$$

where $P(z_i, \omega)$ is the signal of the ith sensor, $k = \frac{\omega}{c}$ and

$$\begin{aligned} m_w &= \left(w - \frac{W-1}{2}\right)\bar{m}, \quad w = 0, \dots, W-1, \\ q_l &= \left(l - \frac{L-1}{2}\right)\bar{q}, \quad l = 0, \dots, L-1, \end{aligned} \tag{6.4}$$

with L and W the number of points on the m and q axis, respectively that are sampled with the corresponding $\bar{m}$ and $\bar{q}$ steps. The term $\psi_{i,l}$ in (6.3) represents a local spatial window which in [37] is defined as the Gaussian window

$$\psi_{i,l} = e^{-\frac{\pi(z_l - q_i)^2}{\sigma^2}}. \tag{6.5}$$

where the term σ controls the width of the window centered in q_l. The RST can be intuitively interpreted as a beamforming operation, weighted by (6.5), considering a single spatial window, i.e. at a fixed l. Actually, the multiplication of the signals $P(z_i, \omega)$ by the complex exponential $e^{-\frac{jkz_i m_w}{\sqrt{1+m_w^2}}}$ represents a beamforming operation where plane waves are characterized by DoA defined according to $\frac{m_w}{\sqrt{1+m_w^2}}$. The choice of this parameterization of the directions follows the ray space definition. More precisely, the DoA is determined by the angle θ, and considering the ray space parameterization we obtain that

$$\sin(\theta) = \sin(\arctan(m)) = \frac{m}{\sqrt{1+m^2}}. \tag{6.6}$$

The RST can be compactly expressed in matrix form. Let us introduce the RST linear operator $\boldsymbol{\Psi} \in \mathbb{C}^{I \times LW}$ whose (i, ι)th element is defined as

$$[\boldsymbol{\Psi}]_{i,\iota} = e^{j\frac{kz_i m_w}{\sqrt{1+m_w^2}}} e^{-\frac{\pi(z_i - q_l)^2}{\sigma^2}} d \tag{6.7}$$

where $\iota = (l + wi + 1)$ is the index of the sampled point in the ray space. Hence, the RST can be computed as a matrix vector multiplication

$$\mathbf{z} = \mathbf{\Psi}^H \mathbf{p} \tag{6.8}$$

where, $\mathbf{p} = \left[P\left(z_0, \omega\right), \ldots, P\left(z_{I-1}, \omega\right)\right]^T$ is the vector of the array signals, and $\mathbf{z} \in \mathbb{C}^{LW \times 1}$ is the ray space data vector obtained rearranging $\mathbf{Z}$

$$[\mathbf{z}]_\iota = [\mathbf{Z}]_{l,w}, \quad \iota = l + wi + 1. \tag{6.9}$$

In [37] the inverse operator of the RST is introduced as the canonical dual matrix $\tilde{\mathbf{\Psi}} \in \mathbb{C}^{LW \times I}$ given by the pseudoinverse of (6.7)

$$\tilde{\mathbf{\Psi}} = \left(\mathbf{\Psi}\mathbf{\Psi}^H\right)^{-1} \mathbf{\Psi}. \tag{6.10}$$

The inverse RST (6.10) is used in order to reconstruct the microphone signals from the ray space data as

$$\tilde{\mathbf{p}} = \tilde{\mathbf{\Psi}}^H \mathbf{z}. \tag{6.11}$$

6.1.2 Fast implementation of the Ray Space Transform

In this section we introduce a computationally efficient implementation of the RST, reviewed in Section 6.1.1, that we refer as the fast ray space transform (FRST).

The fast implementation of the RST represents a potentially appealing tool in the context of BSS, since it allows the reduction of the computational cost of the overall processing. In fact, the use of extended uniform linear microphone arrays implies the processing of a high number of channels. In general, any sound field processing technique in the ray space can benefit from the fast implementation provided by the FRST.

The customary implementation of the RST given in (6.7) employs a matrix-vector multiplication which presents a computational cost that increases linearly with the number of microphones. In the last years, the advantages of low cost digital acoustic sensors, such as MEMS microphones, increased the availability of multichannel systems accommodating a great number of sensors. An example of such technologies is represented by the *eSticks* modular linear microphone array [202]. When it comes to processing a relevant number of sensor signals, however, the computational load of space-time processing algorithms easily becomes a limitation, especially in contexts like extended audio reality. Hence, a less computational demanding implementation of the RST is desirable.

In particular, we show that the RST can be interpreted as a nonuniform Fourier transform and this fact can be exploited for developing a highly-efficient implementation of the RST. This implementation is based on the theory of Nonuniform Fast Fourier Transform (NUFFT) [77, 82, 83, 151, 209, 250, 280]. Similarly to NUFFT algorithms, the FRST requires two consecutive steps. First, a uniform discrete Fourier transform is computed adopting the Fast Fourier Transform (FFT) algorithm on an oversampled range of spatial frequencies. In the second step, the data is interpolated in order to obtain the required RST samples. Clearly, the NUFFT algorithms show a trade-off between accuracy and computational complexity given by the oversampling factor and the interpolation process.

Let us define the nonuniform discrete Fourier transform of a signal $\bar{P}_{i,l}$ as follows [95]

$$[\mathbf{F}]_{l,w} = \sum_{i=0}^{I-1} \bar{P}_{l,i} e^{-j\gamma_w i}, \tag{6.12}$$

where γ_w, $w = 0, \ldots, W-1$ is a set of nonuniformly spaced spacial frequency locations (plane wave directions). In order to show that the RST can be interpreted as a nonuniform discrete Fourier transform, let us equate (6.3) to (6.12) i.e,

$$d \sum_{i=0}^{I-1} P(z_i, \omega) e^{\frac{-jkz_i m_w}{\sqrt{1+m_w^2}}} \psi^*_{l,i} = \sum_{i=0}^{I-1} \bar{P}_{l,i} e^{-j\gamma_w i}. \tag{6.13}$$

The equality in (6.13) is verified under the following conditions

$$\bar{P}_{l,i} = dP(z_i, \omega)\psi^*_{l,i}, \tag{6.14}$$

and

$$\gamma_w = k \frac{m_w d}{\sqrt{1 + m_w^2}}. \tag{6.15}$$

Therefore, given (6.13) and the conditions in (6.14) and (6.15), for each frame $l = 0, \ldots, L-1$, the RST can be rewritten in the form of a nonuniform discrete Fourier transform as

$$[\mathbf{Z}]_{l,w}(\omega) = \sum_{i=0}^{I-1} \bar{p}_{l,i} e^{-j\gamma_w i}. \tag{6.16}$$

A direct evaluation of (6.16) would require $\mathcal{O}(IW)$ operations for each frame $l = 0, \ldots, L-1$. However, in the literature many algorithms have been proposed for a fast computation of the nonuniform discrete Fourier transform [77, 82, 83, 151, 209, 250, 280]. Most of them are based on a two-step approach.

The first step consists of the computation of an oversampled discrete Fourier transform by means of a $N-$point Fast Fourier Transform:

$$[\mathbf{Y}]_{l,n} = \sum_{i=0}^{I-1} s_i \bar{P}_{l,i} e^{-j\frac{2\pi}{N} ni}, \tag{6.17}$$

where $n = 0, \ldots, N-1$ with $N > L$ and s_i are known as scaling factors and are usually designed to reduce the error that is introduced by the subsequent interpolation step [179]. The computation of (6.17) requires $\mathcal{O}(N \log N)$ operations if implemented with an efficient algorithm [65, 76, 267]. As shown in [95], this cost can be further reduced to $\mathcal{O}(N \log I)$ if the ratio N/I is an integer. In fact, in this case, one needs to perform N/I I-point FFTs [248].

The second step consists in an interpolation of the uniformly spaced frequency samples $[\mathbf{Y}]_{l,n}$ in order to obtain an approximation $[\widehat{\mathbf{Z}}]_{l,w}$ of $[\mathbf{Z}]_{l,w}$, i.e.

$$[\widehat{\mathbf{Z}}]_{l,w} = \sum_{n=0}^{N-1} v^*_{wn} [\mathbf{Y}]_{l,n}, \tag{6.18}$$

where v_{wn}^* are the interpolation coefficients. A direct evaluation of (6.18) would require $\mathcal{O}(WN)$ operations. However, usually the interpolation is accomplished by using only the B nearest neighbors to γ_w with $B \ll N$. More precisely, this can be written as

$$[\widehat{\mathbf{Z}}]_{l,w} = \sum_{b=1}^{B} u_{wb}^* \, [\mathbf{Y}]_{l,\{n_w+b\}_N} \,, \tag{6.19}$$

where $\{\cdot\}_N$ is the modulo-N operator and n_w is defined as

$$n_w = \begin{cases} \arg\min_n |\gamma_w - \frac{2\pi}{N}n| - \frac{B+1}{2}, & B \quad \text{odd} \\ \max\left(n : \gamma_w \geq \frac{2\pi}{N}n\right) - \frac{B}{2}, & B \quad \text{even.} \end{cases} \tag{6.20}$$

Various algorithms can be used to find both the interpolation coefficients u_{wb}^* and the scaling factors s_i as discussed in [82, 83, 156, 179]. We consider here two possible approaches for the design of these parameters. The first one is the most straightforward and consists in approximating the value $[\widehat{\mathbf{Z}}]_{i,w}$ with the nearest neighbor. This means setting $B = 1$ and $u_{wb}^* = 1$ in (6.19) and (6.20) and $s_l = 1$ in (6.17). The second approach is the one presented in [95], where the authors adopt a min-max criterion for optimizing the parameters of a Kaiser-Bessel interpolation kernel [184]. As discussed by the authors in [95], since in our case we need to compute multiple NUFFTs of the same size (one for each frame l), using the min-max approach with a Kaiser-Bessel interpolation kernel provides the highest accuracy among all the methods investigated in [95]. Moreover, such approach allows the reduction of the neighborhood size B and, hence, the minimization of the computational complexity. With FRST we refer to the transformation in (6.19).

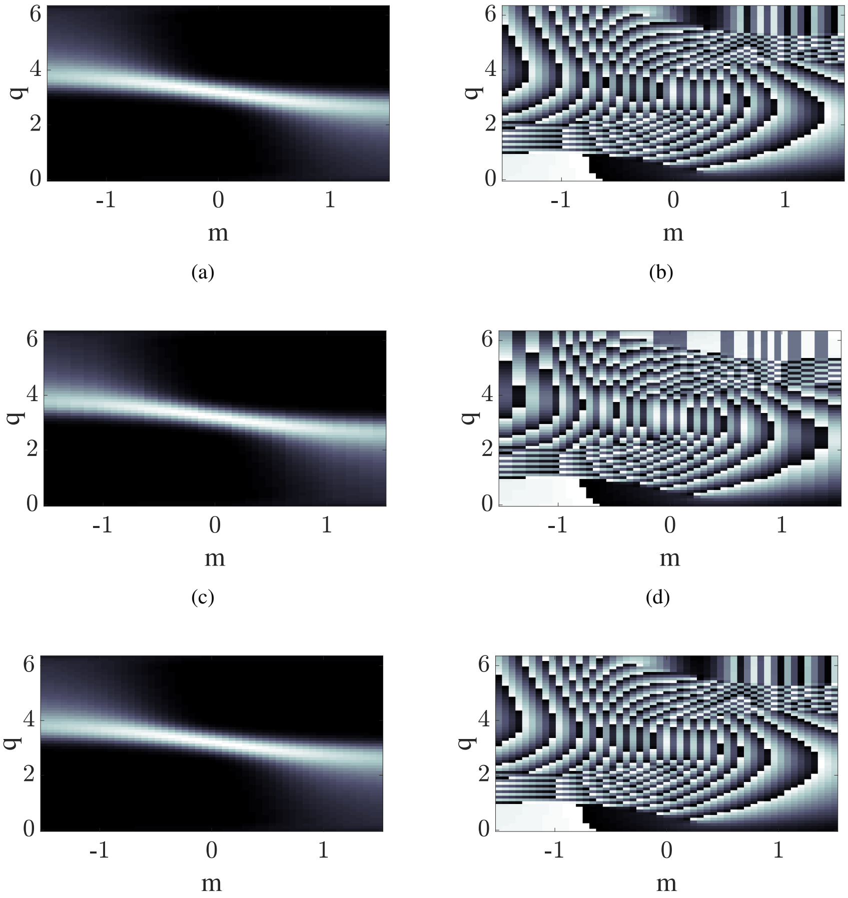
Results

In this section we present some simulation results in order to validate both the efficiency and the accuracy of the FRST. In the following, whenever $B = 1$ we refer to the nearest neighbor criterion, in all other cases we refer to the min-max one. An implementation of both the RST and the FRST is provided online[1].

Simulations The setup adopted for the simulations consists of an array of $I = 64$ microphones, comprised between $z_1 = 0\,\mathrm{m}$ and $z_I = 6.3\,\mathrm{m}$, thus the spacing between consecutive microphones is $d = 0.1\,\mathrm{m}$. The ray space is sampled using $W = 51$ and $L = 64$ points on the m axis and the q axis, respectively, with sampling intervals $\bar{m} = 0.06$ and $\bar{q} = 0.1$.

Given the previously described setup, let us assume that the source in $\mathbf{r}' = [0.5\,\mathrm{m}, 3.2\,\mathrm{m}]^T$ is emitting a sinusoidal signal with temporal frequency $f = 1000\,\mathrm{Hz}$. Figure 6.2 shows both the magnitude and the phase of the output obtained from the RST and FRST. Figure 6.2(a) and Figure 6.2(b) refer to the RST, Figure 6.2(c) and Figure 6.2(d) to the FRST in the case where $B = 1$, while Figure 6.2(e) and Figure 6.2(f) in the case where $B = 5$. Clearly, the results obtained with an higher interpolation order B are characterized by an higher accuracy. This fact is particularly evident when one considers the phase behavior.

[1] https://github.com/polimi-ispl/ray-space-transform



(a) (b)

(c) (d)

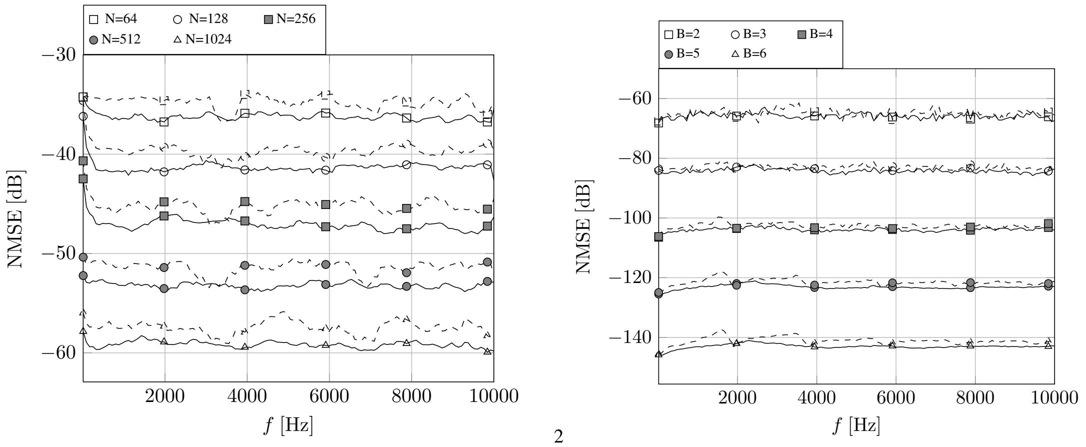
(e) (f)

Figure 6.2: *(a) RST magnitude, (b) RST phase, (c) FRST magnitude, $B = 1$, (d) FRST phase , $B = 1$, (e) FRST magnitude, $B = 5$, (f) FRST phase, $B = 5$*

In order to objectively measure the accuracy of the FRST, we define the Normalized Mean Squared Error as

$$\mathrm{NMSE} = 10\log_{10}\left(\frac{1}{WL}\sum_{l=0}^{L-1}\sum_{w=0}^{W-1}\frac{|[\mathbf{Z}]_{l,w}(\omega) - [\widehat{\mathbf{Z}}]_{l,w}(\omega)|^2}{|[\mathbf{Z}]_{l,w}(\omega)|^2}\right). \tag{6.21}$$

In Figure 6.3 we show the NMSE as a function of f in a range between $10\,\mathrm{Hz}$ and $10\,\mathrm{kHz}$. Without recurring to a complete evaluation for all possible positions of the sources, from preliminary results we selected the two most significant positions. In particular, the first source is placed in $\mathbf{r}'_1 = [5\,\mathrm{m}, 3.15\,\mathrm{m}]^T$ and the second in $\mathbf{r}'_2 = [0.2\,\mathrm{m}, 10\,\mathrm{m}]^T$. Hence, the former is in the end fire while the latter in the broadside with respect to the array. Figure 6.3(a) shows the NMSE as a function of the frequency for different oversampling factors N with $B = 1$. The solid lines are referred to the source in $\mathbf{r}'_1$, while the dashed lines to the source in $\mathbf{r}'_2$. As expected higher N values correspond to a smaller NMSE for both the sources, but a small differences can be observed between the two positions. In Figure 6.3(b), instead, we keep $N = 128$ and vary B. As expected, choosing a higher interpolation factor B leads to better



(a)

(b)

Figure 6.3: NMSE *as a function of temporal frequency when (a) varying the FFT length N and (b) varying the B nearest neighbors used for the interpolation. Solid lines refer to the source in* $\mathbf{r}'_1$ *while dashed lines refer to the source in* $\mathbf{r}'_2$.

results in terms of NMSE. If we compare Figure 6.3(b) with the curve corresponding to $N = 128$ in Figure 6.3(a) we can see that, even for $B = 2$, we achieve higher accuracy. Moreover, the differences between the two sources considerably decrease. This suggests that choosing an interpolation order $B > 1$ leads to more accurate results with respect to the nearest neighbor approach.

Computational Complexity We provide, here, an analysis of the computational complexity of the FRST and the RST based on the considerations introduced in Section 6.1.2. For a given frame i, the FRST requires $\mathcal{O}(N \log N + WB)$ operations when the ratio N/I is not an integer and $\mathcal{O}(N \log I + WB)$ operations when the ratio N/I is integer. On the other hand, the RST requires $\mathcal{O}(IW)$ operations.

Figure 6.4 shows the number of needed operations as a function of the number of microphones I. In particular, I assumes the values in the set

$$\{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$$

and, since we fix $N = 2I$, the ratio N/I is an integer. In Figure 6.4(a) we keep $W = 51$ fixed for each value of I, while in Figure 6.4(b) we choose $W = I$. In both cases we set $B = 5$. By inspecting both Figure 6.4(a) and Figure 6.4(b) it is clear how our implementation of the RST requires less operations, especially as the number of microphones I increases e.g., RST requires more than 3x operations with respect to the FRST for $I = 64$. For example, the evaluation of RST (6.7) with $I = W = 64$ implemented in MATLAB [164] required $3.7\,\mathrm{ms}$ on a consumer laptop[1], while the FRST (6.19) was evaluated in $1.0\,\mathrm{ms}$ with $B = 1$.

6.2 Multichannel NMF model

In this section, we formulate the problem specification of the MNMF based source separation. In particular, we provide the underlying microphone signal model and the

[1] Apple MacBook Pro (15-inch, 2018) with Intel "Core i9" processor (8950HK) and 32GB of 2400 MHz DDR4 RAM.

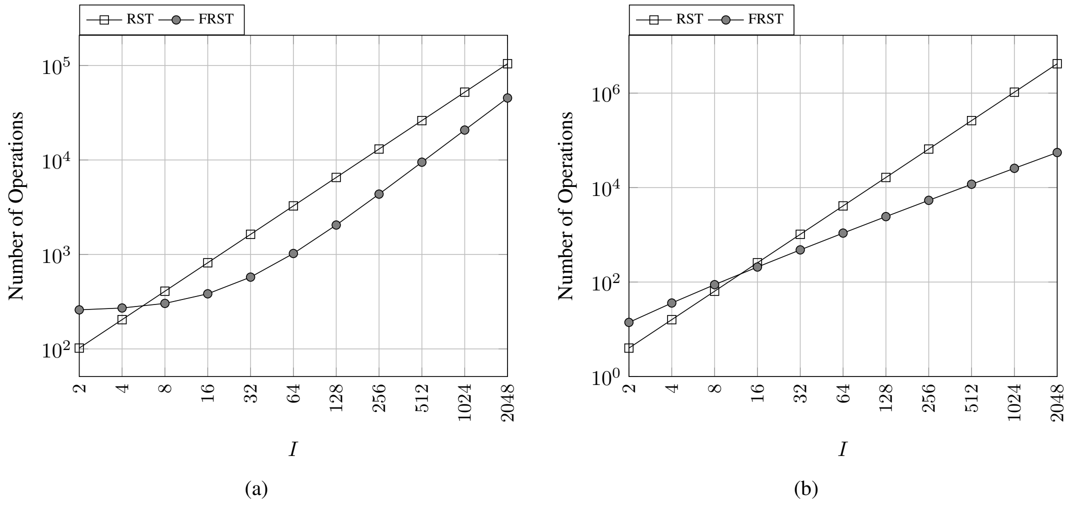


Figure 6.4: *Number of operations as a function of number of microphones I. (a) considers $W = 51$ spatial frequencies, while (b) $W = I$. The x and y axis are on a logarithmic scale.*

probabilistic model adopted by MNMF. Successively, we introduce a MNMF technique closely related to the proposed ray space based MNMF. This approach relies on the transformation of the multichannel signal onto the beamspace domain i.e., a plane-wave-based representation of the array data.

6.2.1 Data model and problem formulation

Let us consider the same setup of Section 6.1.1 made up by a uniform linear microphone array of I microphones (see Figure 6.1(a)). In the presence of N acoustic sources, the time-frequency representation of the ith microphone signal can be written as [189]

$$P_i(\omega, t) = \sum_{n=1}^{N} h_{i,n}(\omega) S_n(\omega, t) + E_i(\omega, t), \tag{6.22}$$

where $i = 1, \dots, I$, is the microphone index, t is the index of the time frame, $\omega = 2\pi f$ the angular frequency with $f > 0$ the temporal frequency, $h_{i,n}(\omega)$ is the transfer function between the ith microphone and the nth source, $S_n(\omega, t)$ is the nth source signal and $E_i(\omega, t)$ models the ith microphone self noise.

MNMF can be formulated based on a so-called local Gaussian model (LGM) [79] that allows modeling and combining spatial and spectral cues in a systematic way. The LGM [79] assumes that each source contribution (also referred as source image), i.e., I-length complex-valued vector $\mathbf{p}_i(\omega, t) = [P_{1,n}(\omega, t) ... P_{I,n}(\omega, t)]^T \in \mathbb{C}^I$, is modeled as a zero-mean circular complex Gaussian random vector as follows

$$\mathbf{p}_n(\omega, t) \sim \mathcal{N}_{\mathbb{C}}\left(0, \mathbf{\Lambda}_n(\omega, t) \lambda_n(\omega, t)\right), \tag{6.23}$$

where the complex-valued covariance matrix is positive definite Hermitian, and is composed of two factors:

1. a spatial covariance $\mathbf{\Lambda}_n(\omega, t) \in \mathbb{C}^{I \times I}$ representing the spatial characteristics of the nth source image at the time-frequency point (ω, t);

2. a spectral variance $\lambda_n(\omega, t) \in \mathbb{R}$ representing the spectral characteristics of the nth source image at the time-frequency bin (ω, t).

As a matter of fact, these source variances $\lambda_n(\omega, t)$ can be modeled using a classical NMF structure as

$$\lambda_n(\omega, t) = \sum_{k=1}^{K} b_{n,k}(\omega) a_{n,k}(t) \tag{6.24}$$

where $b_{n,k}(\omega)$ and $a_{n,k}(t)$ with $k = 1, \ldots, K$ represent the basis functions and the corresponding time-varying gains, respectively. The time varying components $a_{n,k}(t)$ are also referred as the activation functions. It is worth noticing that the model in (6.23) is more general than (6.22). In fact, the signal model in (6.22) corresponds to (6.23) with rank-1 Λ_n matrices. In the case of static sources and assuming the random vectors $\mathbf{p}_n(\omega, t)$ to be independent in time, frequency and between sources, the mixture STFT coefficients in the multichannel mixing in eq. (6.23) may be shown distributed as

$$P_i(\omega, t) \sim \mathcal{N}_{\mathbb{C}} \left(0, \sum_{n=1}^{N} \lambda_n(\omega) \sum_{k=1}^{K} b_{n,k}(\omega) a_{n,k}(t) \right), \tag{6.25}$$

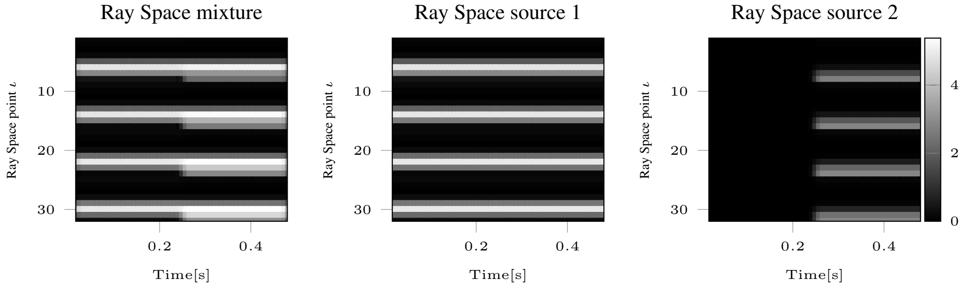
where $\mathbf{\Lambda}_n(\omega)$ could be modeled as a rank-1 SCM or as a full-rank matrix. Several full-rank methods have been proposed recently to provide computationally-efficient solutions [127, 172, 242] that are less sensitive to initialization of the parameters [242]. Alternatively, other multichannel NMF approaches based on the multiplicative update (MU) algorithm in [189] provide lower computational costs at the price of discarding the mutual information between the channels. This is equivalent to set to zero the off-diagonal elements of $\mathbf{\Lambda}_n(\omega)$ in eq. (6.25). Consequently, these approaches do not allow exploiting the interchannel phase differences (IPDs), but only the interchannel level differences (ILDs). However, the IPDs may be very important for multichannel source separation. In fact, using IPDs becomes even more critical for the far-field case (i.e., when the distances between the microphones are much smaller than the distances between the sources and microphones), where the information carried by the ILDs becomes almost non-discriminating. A possible solution to improve the separation performance of these approaches is the integration of spatial information within the observation model, for instance by performing a transformation of the multichannel signals in a domain where the IPDs are inherently exploited.

6.2.2 Beamspace-Domain Multichannel NMF (BS-MNMF)

The beam space domain is introduced in [150] to exploit the inter-channel phase difference during the separation. It describes the microphone signals in terms of the directional components of the sound field. In practice, it starts from a plane-wave decomposition [286] of the signals over a set of M directions, known as the beam space transform [150]

$$\tilde{\mathbf{p}}(\omega, t) = \mathbf{W}_{\mathrm{BT}}^{H} \mathbf{p}(\omega, t), \tag{6.26}$$

where $\mathbf{p}(\omega, t) = [P_1(\omega, t), \ldots, P_I(\omega, t)]^T$ is the vector of the array signals, $\tilde{\mathbf{p}}(\omega, t) = \left[\tilde{P}_1(\omega, t), \ldots, \tilde{P}_M(\omega, t)\right]^T$ is the vector of the beam space signals and $\mathbf{W}_{\mathrm{BT}} \in \mathbb{C}^{I \times M}$



(a) (b) (c)

Figure 6.5: *(a) An example of the ray space data* (6.29)*. Two sources emit a sinusoidal signal at* $\omega = 3\,\mathrm{kHz}$ *with DoA* $\theta_1 = \theta_2 = 45°$*. The second source (c) starts at* $n = 0.25\,\mathrm{s}$*. Although the sources reach the array with the same DoA, the different locations of the sources in space are effectively reflected by the patterns in the ray space data (b) and (c).*

is the beam space transform matrix [296], composed by the steering vectors whose elements are [152]

$$u_i(\theta_m, \omega) = e^{-j\omega \frac{d \sin(\theta_m)}{c}(i-1)}, \quad i = 1, \ldots, I, \tag{6.27}$$

with d the distance between two consecutive microphones, c the speed of sound and j the imaginary unit. Then, inspired by [189], the square magnitude for each beam space signal $\hat{P}_m(\omega, t)$ can be modeled as

$$\hat{P}_m(\omega, t) = \sum_{n=1}^{N} g_{m,n} \sum_{k \in \mathrm{K}_n} b_k(\omega) a_k(t), \tag{6.28}$$

where $g_{m,n} \in \mathbb{R}_+$ represents the mixing weights of the nth source and the mth beam space bin, while K_n is the subset of the basis pertaining the nth source. Note that $g_{m,n}$ in (6.28) is frequency independent since all the frequency components pertaining to the same signals are assumed to have the same DoA. In general, $g_{m,n}$ reaches its maximum when θ_m (6.27) equals the DoA of the nth source [150].

Unlike [189], the beam space domain is based on the plane wave representation of the sound field (see Section 3.6.1). This allows the authors to exploit the IPD (6.27) in the MNMF optimization rather than the magnitude difference (ILD). Nevertheless, the source location information is limited to the DoA only, due to the far field model of (3.13). This limitation can be overcome thanks to the adoption of the ray space in the multichannel NMF signal model [37, 159].

6.3 Ray Space based Multichannel NMF (RS-NMF)

In this section we introduce the ray space representation in the context of MNMF source separation (RS-MNMF). In particular, we show how the MNMF can be applied to the ray space data in order to perform BSS.

First, the multichannel signal of a uniform linear array is mapped onto the ray space domain as described in Section 6.1.1. Successively, the MNMF data model, introduced in Section 6.2, is applied to the ray space data in order to perform the separation of the source signals. We aim at exploiting the inherent representation of the source location in the ray space data in order to enhance the separation capabilities of the MNMF algorithm.

For the ease of the reader, we report the mapping of the array multichannel signal onto the ray space domain performed as (6.3), obtaining as a result the ray space data vector $\mathbf{z} \in \mathbb{C}^{LW \times I}$

$$\mathbf{z} = \mathbf{\Psi}^H \mathbf{p}. \tag{6.29}$$

where $\mathbf{z} \in \mathbb{C}^{LW \times I}$ is the ray space data vector and $\mathbf{\Psi}$ the linear operator defined in (6.7). Due to the indexing of the ray space data vector $\mathbf{z}$ (6.3), the acoustic sources are mapped on comb-like patterns as depicted in Figure 6.5.

The inverse RST (6.10) can be used in order to obtained the time-frequency domain microphone signals from the ray space data (6.29) as (6.11), here reported for the ease of the reader

$$\mathbf{p}(\omega, t) \approx \tilde{\mathbf{\Psi}} \mathbf{z}(\omega, t). \tag{6.30}$$

Inspired by the BS-MNMF solution and the MNMF signal model, we assume that the ray space data can be modeled as

$$Z_\iota(\omega, t) = \sum_{n=1}^{N} r_{\iota,n} S_n(\omega, n) + E_\iota(\omega, t), \tag{6.31}$$

where $r_{\iota,n}$ describes the contribution of the nth source to the ιth ray space element.

The original formulation of the cost function in [150, 189] used the Itakura Saito (IS) divergence [105]. In this thesis, we propose a more general cost function based on the β-divergence [105] which also takes the IS divergence as a special case ($\beta = 0$),

$$C_{\mathrm{RS}}(\Theta) = \sum_{\iota,\omega,t} d_\beta \left(|Z_\iota(\omega, t)|^2 \middle| \hat{Z}_\iota(\omega, t) \right), \tag{6.32}$$

where Θ represents the algorithm parameters, namely the mixing weights, the basis functions and the activation functions. $\hat{Z}_\iota(\omega, t)$ is the estimated square magnitude of the ray space data, modeled as

$$\hat{Z}_\iota(\omega, t) = \sum_{n} g_{\iota,n} \sum_{k \in \mathrm{K}_n} b_k(\omega) a_k(t), \tag{6.33}$$

where $g_{\iota,n} = |r_{\iota,n}|^2$. It is worth noting that the proposed model is characterized by a frequency-independent mixing model, since it depends only on the position of the sources and the same line pattern is expected for every frequency corresponding to the same active source. Additionally, other than the IS divergence, well-known divergences can be obtained by properly setting the parameter β in (6.32). In particular, $\beta = 1$ and $\beta = 2$ correspond to the Kullback-Leibler (KL) divergence and the squared Euclidean (EUC) distance, respectively, as described in [257]. We can therefore exploit the similarity with the instantaneous algorithm of [189] to derive the updated algorithm of the MU method:

$$g_{\iota,n} \leftarrow g_{\iota,n} \frac{\mathrm{sum}\left[\hat{\mathbf{Z}}_\iota^{\cdot \beta - 2} \cdot \mathbf{Z}_\iota \cdot (\mathbf{B}_n \mathbf{A}_n) \right]}{\mathrm{sum}\left[\hat{\mathbf{Z}}_\iota^{\cdot \beta - 1} \cdot (\mathbf{B}_n \mathbf{A}_n) \right]}, \tag{6.34}$$

$$\mathbf{B}_n \leftarrow \mathbf{B}_n \cdot \frac{\sum_{\iota=1}^{LD} g_{\iota,n} \left(\hat{\mathbf{Z}}_\iota^{\cdot\beta-2} \cdot \mathbf{Z}_\iota\right) \mathbf{A}_n^T}{\sum_{\iota=1}^{LD} g_{\iota,n} \hat{\mathbf{Z}}_\iota^{\cdot\beta-1} \mathbf{A}_n^T}, \tag{6.35}$$

$$\mathbf{A}_n \leftarrow \mathbf{A}_n \cdot \frac{\sum_{m=1}^{M} \left(g_{\iota,n}\mathbf{B}_n\right)^T \left(\hat{\mathbf{Z}}_\iota^{\cdot\beta-2} \cdot \mathbf{Z}_\iota\right)}{\sum_{\iota=1}^{LD} \left(g_{\iota,n}\mathbf{B}_n\right)^T \hat{\mathbf{Z}}_\iota^{\cdot\beta-1}}. \tag{6.36}$$

where $\mathrm{sum}[\mathbf{M}]$ is the sum of all the members in $\mathbf{M}$ and $\cdot$ represents element-wise matrix operations. The matrices $\mathbf{B}_n = [b_k(\omega)]_{\omega,k\in\mathrm{K}_n}$ and $\mathbf{A}_n = [a_k(t)]_{k\in\mathrm{K}_n,t}$ are the weight and basis matrices, as in [189], while $\hat{\mathbf{Z}}_\iota = [\hat{z}_\iota(\omega,t)]_{\omega,t}$ contains the estimated mixture data in the ray space domain.

We can obtain an estimate of the ray space source image in terms of the minimum mean squared error (MMSE) as in [150, 189]

$$\tilde{S}_n^{(\iota)\mathrm{im}}(\omega,t) = \frac{g_{\iota,n} p_n(\omega,t)}{\hat{Z}_\iota(\omega,t)} Z_\iota(\omega,t), \tag{6.37}$$

where $\tilde{S}_n^{(\iota)\mathrm{im}}(\omega,t)$ is the estimated contribution of the nth source at the ιth ray space bin and $p_n(\omega,t) = \sum_{k\in\mathrm{K}_n} b_k(\omega)a_k(t)$ represents the estimated power spectral density of the source factorized through the basis and activation functions. Finally, an estimate of the sources at each microphone can be obtained applying the inverse RST (6.10)

$$\hat{\mathbf{S}}^{\mathrm{im}}(\omega,t) = \tilde{\mathbf{\Psi}}(\omega,t)\tilde{\mathbf{S}}^{\mathrm{im}}(\omega,t), \tag{6.38}$$

where $\tilde{\mathbf{S}}^{\mathrm{im}}(\omega,n) = [\tilde{S}_n^{(1)\mathrm{im}}, \dots, \tilde{S}_n^{(LD)\mathrm{im}}]^T$ and

$$\hat{\mathbf{S}}^{\mathrm{im}}(\omega,t) = [\hat{S}_n^{(1)\mathrm{im}}, \dots, \hat{S}_n^{(I)\mathrm{im}}]^T \tag{6.39}$$

is the vector of the nth estimated source signal at the microphones.

6.3.1 Validation and Results

The performance of the RS-MNMF is compared with respect to BS-MNMF and recent state-of-the-art multichannel NMF techniques. In particular, we considered FastMNMF [242], DOA-MNMF [57], WN-MNMF [172] and ILRMA [135].

In order to evaluate the proposed technique, we performed a set of experiments in a reverberant environment. We adopted an ULA of $I = 32$ microphones composed by two *eSticks* [202]. In order to increase the variety of input data, the RIR between each position of the source and the microphones were estimated using sweep excitation [91, 176]. Array signals have been computed through the convolution between the acquired RIR and the first $3\,\mathrm{s}$ of the source signals (male and female speakers and no-drum music signals) taken from dev1 dataset of [186] such that $J = 3$ or $J = 2$ sources are active simultaneously.

The measurements were performed in an office room of dimension $5.5\,\mathrm{m} \times 3.4\,\mathrm{m} \times 3.3\,\mathrm{m}$ with an estimated average $\mathrm{T60} \approx 0.4\,\mathrm{s}$. The RIRs between the microphones and 9 source locations were acquired using a Genelec 8020C [188] loudspeaker. Source locations are organized on a grid with distances between $0.3\,\mathrm{m}$ and $0.9\,\mathrm{m}$ from the lying line of the array. The setup is illustrated in Figure 6.6.

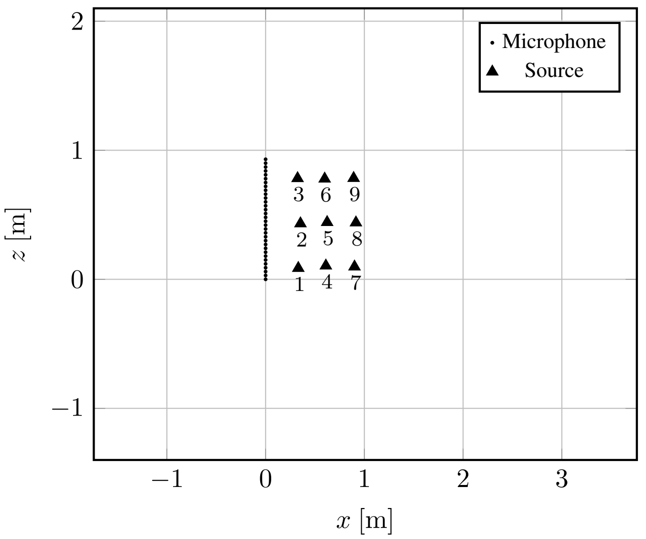


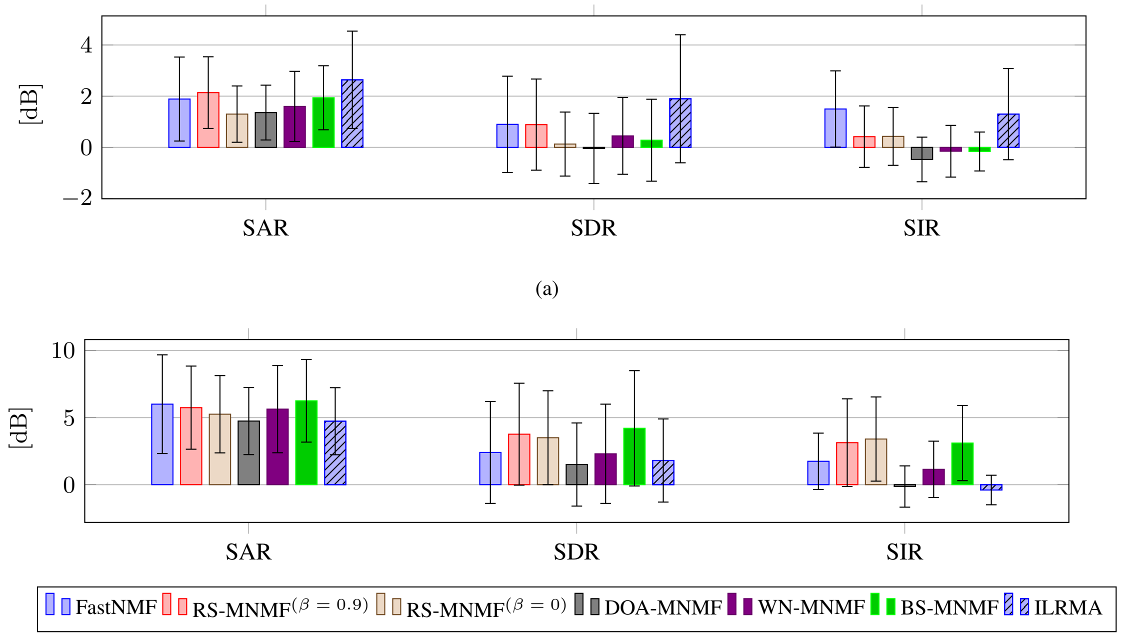
Figure 6.6: *2D graphical representation of the office room setup (top view).*

The procedure is implemented in MATLAB [164][2]. Signals are processed at a sampling rate of $8\,\mathrm{kHz}$, the STFT adopts a hamming window of size 256 with 75% overlap and 512 FFT points. The number of bases for each source $\#\mathrm{K}_j$ and the number of iterations were empirically set to 12 and 100, respectively, and we adopted the same values for all the employed algorithms. For what concerns the beamspace, we adopted $M = I = 32$ angles (3.13) uniformly sampled in the range $\left\{-\frac{\pi}{2}, \frac{\pi}{2}\right\}$. The number of Ray Space points (6.7) is set to be equal to the number of microphones $LD = 32$, with $L = 8$ subarrays and $D = 4$ directions uniformly sampled such that $\mu_{\mathrm{w}} \in \{-0.09, 0.09\}$. We empirically found that $\beta = 0.9$ in (6.32) provides an overall improved performance with respect to $\beta = 0$ (IS divergence). Results with both β values are reported. For what concerns the reference algorithms, we set the parameters following the authors' suggestions available in the related manuscripts. For all the techniques, the values of the parameters were tuned observing the results obtained with a validation dataset concerning $J = 2$ sources with $3\,\mathrm{s}$ speech signals taken from [264] in a subset of locations of Figure 6.6. In order to asses the performance of the proposed algorithm we compute the SAR, SDR and SIR metrics [275] for each microphone signal and the average value over all the microphone signals of these metrics is considered.

Results

In Figure 6.7 the average and standard deviation of the metrics computed for every combination of $J = 3$ with the different algorithms are reported. The results in Figure 6.7(a) shows the average between the metrics obtained with $J = 3$ male speech source signals and $J = 3$ female speech source signals. It is observed that FastMNMF provides the best SIR performance, while ILRMA outperforms the other techniques in terms of SAR and SDR. Compared to FastMNMF, RS-MNMF achieves on average higher SAR and comparable SDR. As regards the SIR, RS-MNMF records a better performance with respect to both the DOA-MNMF and the WN-MNMF other than the

[2] `https://github.com/polimi-ispl/rs-mnmf`



(a)

(b)

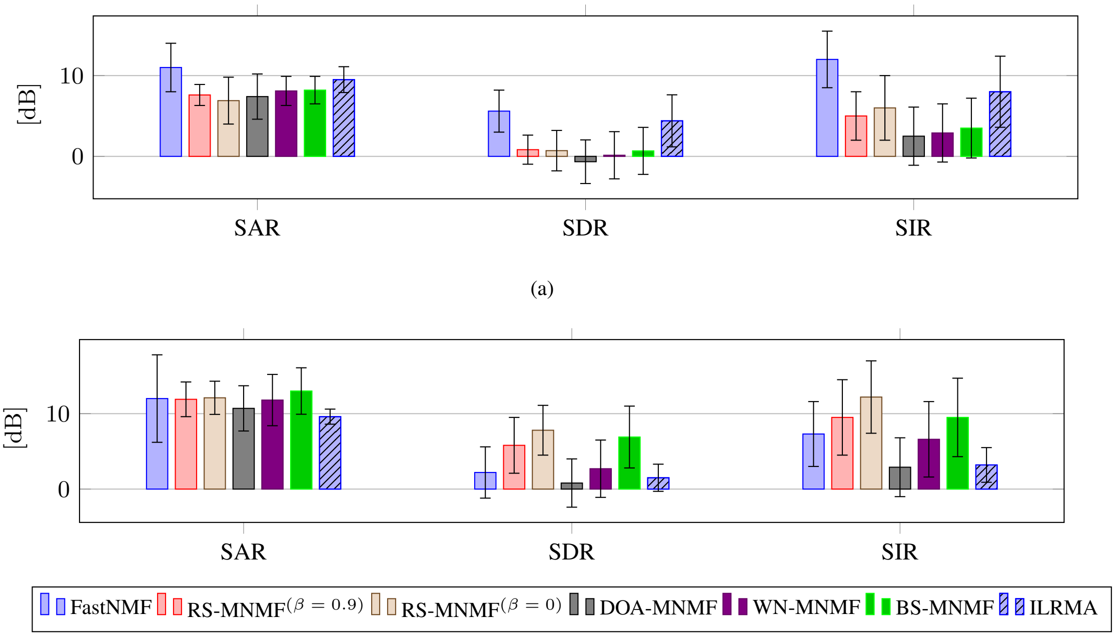
Figure 6.7: *The* SAR, SDR, *and* SIR *averages and standard deviations obtained by each BSS algorithm under analysis for $J = 3$ sources with (a) speech and (b) music signals.*

BS-MNMF. The results in Figure 6.7(b) are obtained with music source signals. In this case the RS-MNMF and the BS-MNMF show the best performance. In particular the proposed RS-MNMF achieves the highest SIR on average. In general, the possibility of varying the cost function allows tuning the performance to favor separation (SIR) (e.g. with $\beta = 0$) over distortion and/or artifacts. In fact, it is known [115,140,254,274] that higher separation capabilities usually correspond to an increased artifact level.

A further analysis concerns a scenario with two sources active simultaneously. We consider the same setup described previously, adopting all the combinations of $J = 2$ sources. The source signals are the same employed in the three sources scenario and taken from [186]. Hence, we considered male and female speech signals and music signals.

From an overall inspection of Figure 6.8, we can observed the same trend in the performance of the techniques as in the three sources scenario. In particular, in the case of speech source signals (Figure 6.8(a)), here, the FastMNMF provides on average the highest values of the three metrics ($\mathrm{SAR} = 11\,\mathrm{dB}$, $\mathrm{SDR} = 5.6\,\mathrm{dB}$, $\mathrm{SIR} = 12\,\mathrm{dB}$) followed by ILRMA with ($\mathrm{SAR} = 9.5\,\mathrm{dB}$, $\mathrm{SDR} = 4.4\,\mathrm{dB}$, $\mathrm{SIR} = 8\,\mathrm{dB}$). Nonetheless, RS-MNMF outperformed DOA-MNMF, WN-MNMF and BS-MNMF in terms of SDR and SIR. As a matter of fact RS-MNMF$^{(\beta=0.9)}$ records $\mathrm{SDR} = 0.83\,\mathrm{dB}$ and $\mathrm{SIR} = 5\,\mathrm{dB}$ on average. In addition, RS-MNMF$^{(\beta=0.9)}$ provides the lowest standard deviation for all the three metrics.

As regards the separation of $J = 2$ music sources, i.e., a flute and a guitar signal, the results associated to each technique are reported in Figure 6.8(b). Similarly to the three sources scenario, also in this case RS-MNMF is able to provide the best performance in terms of SIR. Moreover, here, the proposed technique records SDR higher than



(b)

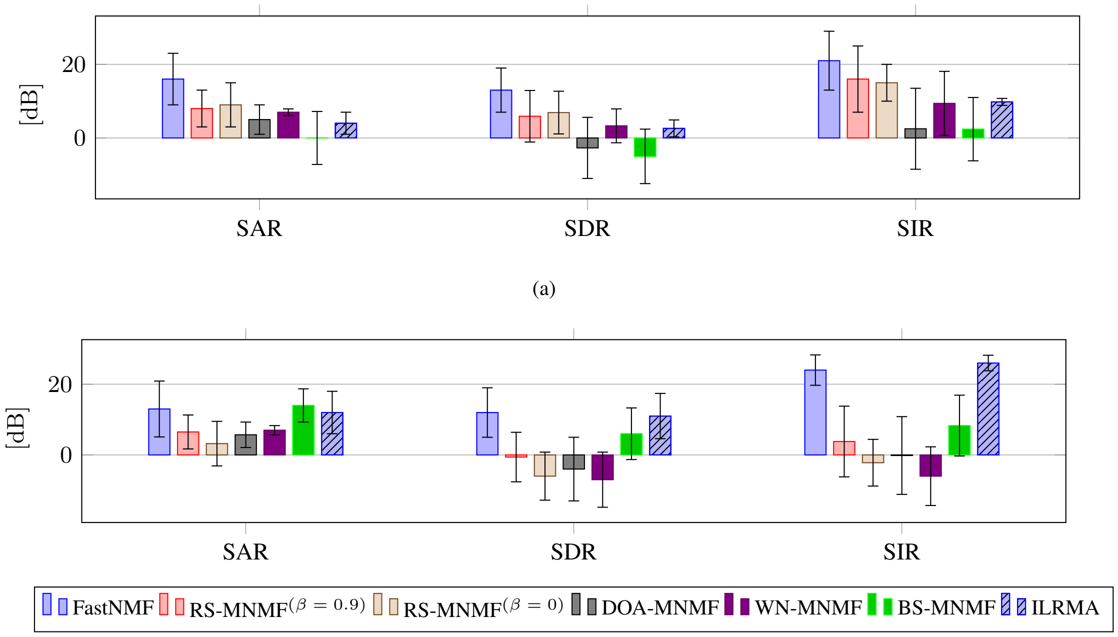
Figure 6.8: *The* SAR, SDR, *and* SIR *averages and standard deviations obtained by each BSS algorithm under analysis for $J = 2$ sources with (a) speech and (b) music signals.*

BS-MNMF, namely $\mathrm{SDR} = 7.8\,\mathrm{dB}$ for RS-MNMF$^{(\beta\,=\,0)}$ and $\mathrm{SDR} = 6.9\,\mathrm{dB}$ for BS-MNMF.

Finally, in order to evaluate the robustness of the proposed technique with respect to different initializations, we performed an analysis of the separation performances randomly varying the initial values of the parameters. In particular, we considered one fix setup concerning $J = 2$ sources with two male speech signals. The locations of the sources are labeled in Figure 6.6 as positions 1 and 3. We run the algorithms 10 times varying the initial parameters randomly. Successively, the estimates of the sources are evaluated in terms of SAR, SDR and SIR.

In Figure 6.9 the average and the standard deviations of the metrics given by the multiple executions are reported. In general from the inspection of Figure 6.9, we can note that the proposed RS-MNMF presents standard deviations of the metrics in line with respect to the other techniques under analysis such as FastMNMF, DOA-MNMF and BS-MNMF. Nonetheless, the lowest values in standard deviation are given by WN-MNMF SAR (Figure 6.9(a)) and ILRMA SIR (Figure 6.9(b)). Noteworthy, the standard deviations provided by the RS-MNMF are rather consistent for the two sources, independently of the metric. In fact, the difference in the performance is below $1\,\mathrm{dB}$ for both RS-MNMF$^{(\beta\,=\,0.9)}$ and RS-MNMF$^{(\beta\,=\,0)}$. Only the DOA-MNMF achieves the same result for all the three metrics.

The results showed that the ray space is a suitable representation for applying MNMF algorithm and it is effective for the application in real world scenarios. The adoption of the ray space let us enhance the performance with respect to the other unconstrained MNMF algorithms and we obtained competitive results with the lastest constrained MNMF techniques.



(a)

(b)

Figure 6.9: *Average and standard deviation of the metrics obtained with ten executions of the algorithms. (a) Results for the first source. (b) Results relative to the second source.*

Part III

Modeling Virtual Sound Sources for Extended Reality

CHAPTER 7

Measurement of the Violin Directivity Pattern

In this thesis, we tackle the implementation of a virtual source (VS) through a study case. We focus our attention on the violin as a virtual source for the EAR. The violin is an interesting musical instrument that attracted the attention of many researchers due to its peculiar mechanical and acoustical characteristics. Considering the implementation of violin as VS, we could rely on different strategies. Distinctions can be made accordingly to the invasiveness of the approaches. On the one side, we can retrieve the VS parameters from measurements taken on real instruments. On the other side, we can rely simulations where the acoustic radiation of the instrument is predicted and consequently the VS model parameters are estimated.

In this chapter, we estimate the directivity pattern of a set of historical violins. The goal of this work is twofold.

First of all, we collect measurements that can be exploited for the VSs. As a matter of fact, we can approach the implementation of violin VSs measuring the directivity directly from the instruments themselves. This allows us to create VS replicas of actual instruments *virtualizing* the violin through the VS parameters retrieved from the acquired data. Here, we rely on acoustic measurements performed while the instruments are played. Obviously, this scenario is close to the actual listening conditions, but it presents an high invasive solution, due to the interaction of the player with the instrument.

The second aim of this chapter is to provide a characterization of the directional radiation properties of violins. Although the violin looks like a simple wooden object, the geometry and material properties of its components and their interaction with the player are rather complex. It follows that a complete and exhaustive characterization of violins is still under analysis, especially when it comes to the comparison between instruments. In the context of the Musical Acoustics Lab, we have the possibility to collaborate with the "*Museo del Violino*"[1] settled in Cremona (Italy), where some of the most renowned violins built by the great masters are preserved. This gave us the unique opportunity to perform acoustic analysis on a relevant number of valuable historical violins. In particular, for the first time for such refined instruments, we measured their directivity patterns and we characterized their directional properties on a wide frequency range, providing a set of tools for quantitive description and comparison of different instruments. While in Section 7.1, we provide a review of the acoustics of violins, the data model and the measurement methodology are described in Section 7.2. Due to their complex directional sound radiation, a comprehensive analysis and comparison between instruments is difficult to provide. Therefore, in Section 7.3, we introduce different tools that focus on the identification of the principal radiation regions, i.e., direction of high sound energy radiation, and metrics of similarity between the directivity patterns of different instruments. Other than the historical instruments, two *twin* violins were measured in order to provide a similarity benchmark for violins with the same geometry and wood.

Results offered in Section 7.4 show that the *twins* provide similar directivity pattern, while interesting observations on the historical instruments can be drawn. As a matter of fact, common directional behavior can be observed, which highlighted shared characteristics among the harmonics of strings; significant examples can be found for A and E strings. As far as the comparison between individual instruments is concerned, similarities among violins of the same luthier and time of building are revealed by the proposed tools. In parallel, some instruments stand out for being different from other violins of the set. The differences among violin directivity patterns, shown in this chapter, demonstrate that an accurate modeling the VS directivity cannot be neglected since each single instrument present original directional characteristics.

7.1 Review of the Acoustic of Violins

The quality of a violin is deeply related to the ability of its maker. Instruments made by the old great Italian masters, like *Antonio Stradivari* or *Giuseppe Guarneri del Gesù* are masterpieces used as inspiration by modern violin makers allover the world. Nevertheless, the preferences of listeners and players, and consequently, the final price of an instruments may be extremely different from violin to violin demonstrating that the instrument quality lies in the subtle details. This result is not surprising, if we consider the complexity of the violin as a physical system.

We can roughly describe the violin as a wooden box with two arched guitar-shaped plates that generates sound through the vibrations caused by its bowed strings. In practice, it acts as a converter transforming the energy of the vibrating strings into sound pressure that is radiated by the body. Therefore, the strings do not actually radiate sound but rather provide the driving force that finally produces the acoustic waves. It follows that the sound quality is also related to the string motion generated by the bowing of the player. It is known that the string excitation applied to the body through the bridge of the violin is generated by highly nonlinear interactions. In particular, a proper bowing generates oscillations known as Helmholtz wave through the nonlinear interaction given by the friction between the string and hair of the bow. The Helmholtz wave presents a saw tooth waveform that is able to excite the body with its transverse

[1] `https://www.museodelviolino.org`

force applied on the bridge. As a consequence, through the bowing action, a skilled player, is able to control the sound providing expressiveness to the performance.

In the mechanism of sound generation, the coupling between the vibrating strings and the modes of body shell is fundamental. Therefore, many studies [40,41,113] focus on the characterization of the dynamical behavior of the body and the influence on the final radiated sound. As regards the acoustic radiation, other works [169,196,278,284], instead, focus on the directional properties of sound radiation providing insights on the spatial radiation characteristics of violins.

Different methodologies have been introduced in the literature in order to measure the directivity of violins and they can be mainly divided according to the excitation type and measurement setup.

In [40, 41, 67, 157] the violin is required to be mounted on specifically designed stand in order to let the instrument freely vibrate. The instrument is excited by means of an impulsive force applied to the bridge and provided by an impact hammer or a sinusoidal excitation [169], while the strings are dampened in order to retrieve the body response only. The radiated sound field is then captured by means of one [67, 169] or more microphones organized in arrays [40, 41, 157] that are placed or moved in order to cover all the possible directions around the instrument.

An interesting inverse-like methodology based on the reciprocity principle is adopted in [283, 284], where the violin is excited by means of an incoming acoustic wave emitted by a loudspeaker and the vibrational response of the instrument is measured at the bridge.

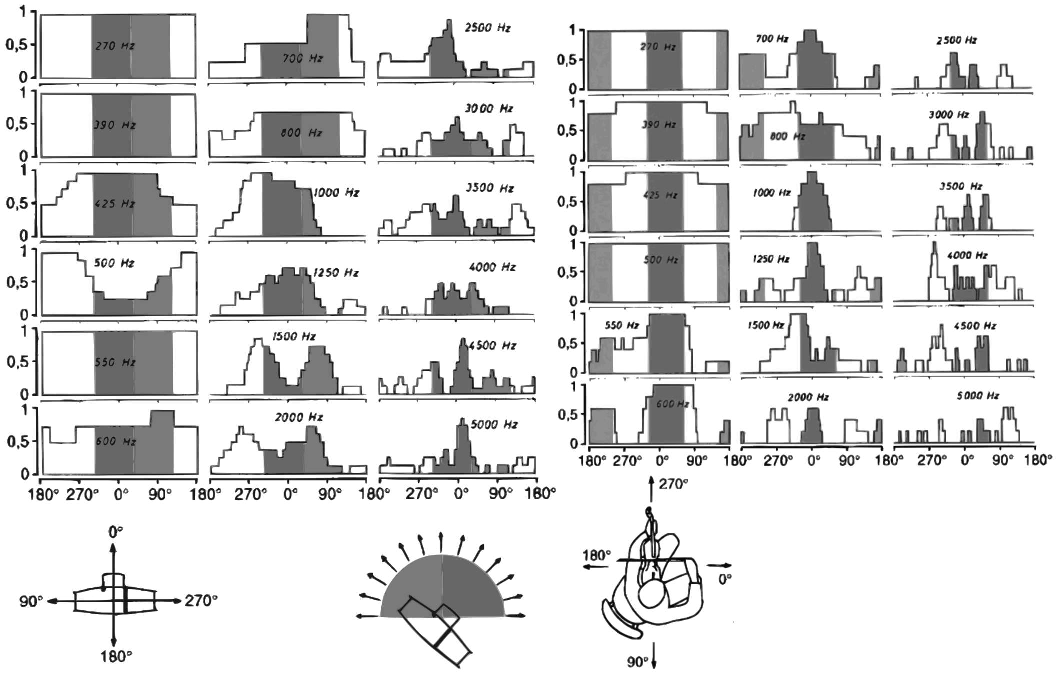
The main limitation of the aforementioned techniques is related to the type of input excitation. In particular, they do not consider the actual force at the bridge applied by the bowed strings. The non-linear excitation of the bowed string is known to be a fundamental aspect of the dynamical behavior of the violin and this is reflected on the quality of the emitted sound [112, 289].

In [278] the sound radiation of the violin is analyzed by means of near-field acoustic holography [165, 288]. The violin is supported by an ad-hoc structure and it is excited by the action of a mechanical bowing machine, while the near-field acoustic pressure is captured by a roving microphone array.

This wide range of techniques for the estimation of the violin directivity is based on artificial excitation such as impulsive forces, sound waves or mechanical bowing, that present the advantage of being repeatable and accurately measurable. Nevertheless, they do not consider the effect of the human body on the sound radiation and the influence of the player on the emitted sound [289].

In order to take into account the presence of the violinist, in [93, 193, 196, 220] the directivity is measured while the instrument is played adopting well-defined positions both for the player and the violin.

In general, all the mentioned techniques are based on measurements performed in an anechoic room. While providing an acoustically controlled environment ideal for artificially-excited methods, this setup could compromise the performance of the player when required. In particular, the *dry* acoustics of the anechoic room is unusual for a violinist and typically it results in an uncomfortable perception that might affect the quality the performance in addition to the prescribed unnatural playing positions required during the directivity sampling.



(a)

(b)

Figure 7.1: *Histogram of the principal radiation directions (0–3dB) for the violins in [169, 170]. (a) Directions taken on the bridge plane. (b) Directions taken on the horizontal plane. Dark gray areas: angular region for the first violin oriented directly toward the audience. Light gray areas: corresponding angular region for the second violins with European seating. The figures are taken from [170].*

Therefore, in [55] the authors propose a methodology for the estimation of the directivity of violins that enables the measurement with a flexible setup. The violinist is allowed to move during the measurement while the violin location and orientation are tracked using a depth map camera and a gyroscope, respectively. The radiated sound is then captured using a pleancoustic camera, namely a rectangular microphone array supported by pleancoustic signal processing [160], while the input signal is acquired by means of a proximity microphone on the instrument. This enables the measurement of the violin directivity even in low reverberant conditions.

In this chapter, we adopt an enhanced and generalized version of the algorithm of [55] that has been recently proposed in [54]. In particular, in [54] the authors employ two pleanacoustic cameras both for measuring the radiated sound and for localizing the source, while the robustness against reverberation is improved by means of further signal processing. With respect to the other methods, this allows us to analyze and compare the directivity of the violins with measurements taken in a context closer to the actual playing conditions.

The well-known work by Meyer [169] examines the directivity patterns of bowed stringed instruments, i.e., violin, viola, cello and double bass in terms of principal directions covering a wide range of frequencies. As regards the violin, directivity patterns are evaluated over two orthogonal planes, the first is the bridge plane perpendicular to the top plate (see Figure 7.1(a)), while the second horizontal plane is parallel to the violin body (see Figure 7.1(b)). The directional behavior of violins is summarized for

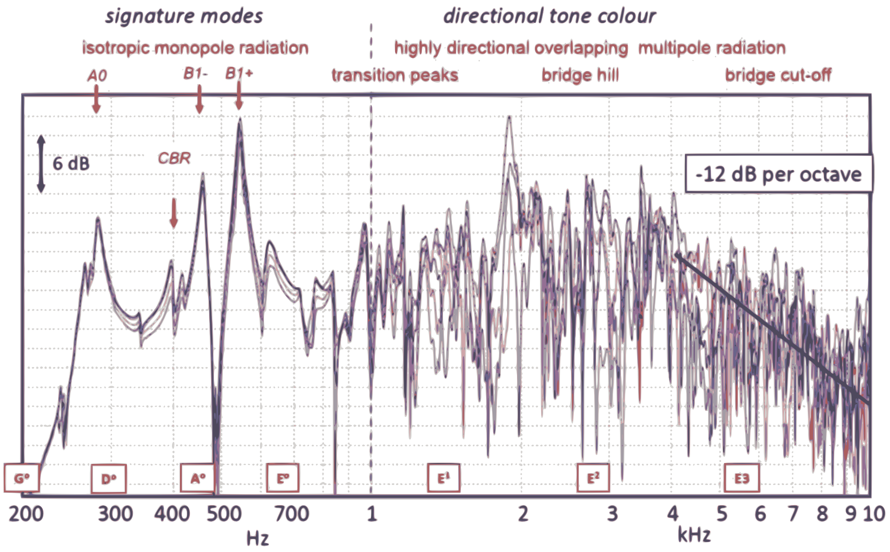


Figure 7.2: *Superimposed spectra of the radiated sound pressure measured in the bridge plane at different directions. The figure is taken from [114].*

each plane by means of histograms that provide the probability for a given direction to be a principal radiation direction. In general, the histograms reveals the omnidirectional behavior of the instrument up to around $800\,\mathrm{Hz}$ for both the planes. At the middle frequencies $[1, 2]$ kHz, a dipole radiation is present, with two main maxima on the bridge plane pointing toward the upper front and back sides of the violin (see Figure 7.1(a)). As regards the horizonal plane, in the same frequency range, one main direction towards the audience can be noted (see Figure 7.1(b)). Inspecting Figure 7.1 at higher frequencies, more than one main direction is present on the horizontal plane, while on the bridge plane the directivity is mainly perpendicular to the top plate. The same analysis approach was also adopted for characterizing the directivity patterns of other classes of musical instruments in [170].

In [278], the sound radiation of three different violins is compared and evaluated up to $3\,\mathrm{kHz}$. The results highlight the relevant contribution of the top plate on the overall radiation, especially at high frequencies, while the back plate barely radiates sound. Moreover, the authors of [278] corroborate the baffling effect of the violin body [284] above $880\,\mathrm{Hz}$, when the acoustic wavelength becomes comparable to the radius of the violin (approximately $6\,\mathrm{cm}$). This value of the radius corresponds to the distance between the top and the back at the waist of the instrument. In addition, the predominance of the isotropic radiation is confirmed for the lowest frequencies, while the acoustic center is located approximately at the soundpost on the top plate. In general, the studies suggest that the frequency response of the radiated soundfield can be roughly divided into three main overlapping regions that present very different characteristics [40, 114, 289] (see Figure 7.2).

7.1.1 The Signature Mode Region

In the lowest frequency range, up to around $1\,\mathrm{kHz}$, the radiation is governed by a small number of well-defined body resonance modes [39, 40, 113]. Such modes mainly concerns the *cavity* or *Helmholtz-like* modes usually referred as $A0$ and $A1$ and the *corpus* modes identified as CBR, $B1-$ and $B1+$ (see Figure 7.2). They are considered as an acoustic fingerprint for each individual violin, since their frequency and radiation intensity are typical of the instrument [39, 40]. In particular, in [40] the radiativity of the *Helmholtz-like* mode $A0$ emerged as a relevant feature for the discrimination of the violin quality with higher radiativity associated to excellent instruments. The relationship between the signature modes and the structure of the violin has been studied in [113] by means of FEM simulations providing a model for determining the influence of the different components to the final instrument sound. In this frequency region, the violin mainly presents an overall isotropic monopole-like radiation caused by the strongly radiating *Helmholtz-like* mode $A0$ and the body shell bending modes $B1-$ and $B1+$, which are characterized by a significant air volume flow through the f-holes of the top plate. We refer the reader to the supplementary material [2] of [114] for videos illustrating the shell modes motion.

7.1.2 The Transitional Region

The frequencies going from around 0.8-$1\,\mathrm{kHz}$ up to 2-3kHz are characterized by a complex acoustic response resulting from a superposition of multiple modes. This is due to a general higher density of the resonances and modal damping with respect to the signature mode region [40]. In practice, the violin radiates sound as a multipole with directional properties and the characterization of the directivity in this frequency range is rather complex since it is known to be variable in frequency [284]. As a matter of fact, in Figure 7.2, a greater variability of the acoustic radiation sensed from different directions can be noted in this frequency range. Such behaviour, known as *directional tone color* is referred by Weinreich in [284] to provide "*the illusion that each note played by a solo violin comes from a different direction, endowing fast passages with a special flashing brilliance.*" and it is assumed to be related to the *projection* quality of the instrument. In [284] and later in [196] the authors measure the *directional tone color* by computing the radiation ratio between two different directions as a function of the temporal frequency. Despite making clear evidence of the phenomenon, this metric does not provide a full analysis of the fluctuating directivity patterns of the instrument in terms of pattern shapes and principal directions.

7.1.3 The High-frequency Region

In this frequency region the resonance density and overlap makes barely impossible to perform modal analysis and a statistical approach for describing the acoustic response is adopted in [290]. Here, the main features are related to the bridge influence on the overall acoustic response of the violin body. In particular, around $3\,\mathrm{kHz}$ the response is boosted (see Figure 7.2) due to the so called *bridge-hill* effect [289], mainly caused by the strong resonance of the bridge placed in this frequency range. In [81] the authors argue that the bridge hill must be attributed also to the interaction of the bridge and the f-holes wings. Moreover, at frequencies higher than $3\,\mathrm{kHz}$ a rather rapid roll-off of the radiated energy can be observed due to the bridge damping effect [290] (see Figure 7.2).

[2] https://acousticstoday.org/supplementary-text-violinacoustics-colin-e-gough/

In conclusion, the sound radiation of the violin has been broadly studied in the literature, and many insights on the relationship between the mechanical behavior and the acoustic response at the low frequencies have been given. At higher frequencies, the overall trend of the instruments has been studied, but a specific and complete characterization could not be easily performed. Therefore, usually a qualitative comparison of different instruments, or an investigation on an interesting but limited set of directions has been performed. In this thesis, we aim to propose a quantitative analysis of the directivity patterns that can be used for characterizing the directional properties of the radiation and comparing different instruments. The proposed methodology is validated on violins built from the same wood pieces and with very similar geometries, and applied on a set of historical violins.

7.2 Characterization of the Violin Directivity Pattern

When it comes to the characterization and the comparison of the directivity, we are mainly interested in the evaluation of the violin directional behavior both as the deviation from an isotropic omnidirectional radiation and in terms of its principal directions, namely the directions of maximum acoustic energy emission. It is worth to underline that the directivity is a frequency-dependent quantity and as a result a compact but complete and intuitive description is difficult to provide. In this chapter, we propose a set of tools for the characterization and comparison of the violin directivity pattern, i.e., the magnitude of the directivity function. Such tools can be generally applied regardless of the measurement technique and their goal is to capture the directional characteristics of the instruments in an intuitive fashion. On the one hand, we take advantage of the spherical harmonics representation [54,285] (see Section 3.6.2) in order to describe the directivity pattern as a compact set of features that allows a general characterization of the directional properties. On the other hand, we rely on signal processing for analyzing the principal directions and comparing the shape of the patterns. The proposed metrics are used to characterize the directivity patterns of a set of 10 historical valuable instruments made by the well-known great Italian masters: *Antonio Stradivari*, *Giuseppe Guarneri del Gesu*, *Nicolò Amati* and *Andrea Amati*. In addition, we characterize and compare the directivity patterns of a pair of twins violins. The twins violins have been built by the same violin maker in order to share the very same geometry and wood. Hence, the twins violins are expected to show very similar directional characteristic.

The developed directivity pattern models could be employed with the aim of creating synthetic VS directivity whose main features mimic those of a given historical instrument, allowing a user to virtually listen to the sound field of a desired instrument.

7.2.1 Data Model and Experimental Methodology

Let us consider the directivity pattern measurement setup shown in Figure 7.3 with the reference system centered on the violin. The acoustic pressure radiated by the instrument to the ith omnidirection microphone located in the far-field [286] of the source at $\boldsymbol{r}_i$ is defined in the frequency domain as the sum of the direct and reverberant sound field components (ref. (5.38))

$$X\left(t,\omega,\boldsymbol{r}_i\right) = X_{\mathrm{dir}}(t,\omega,\boldsymbol{r}_i) + X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i). \tag{7.1}$$

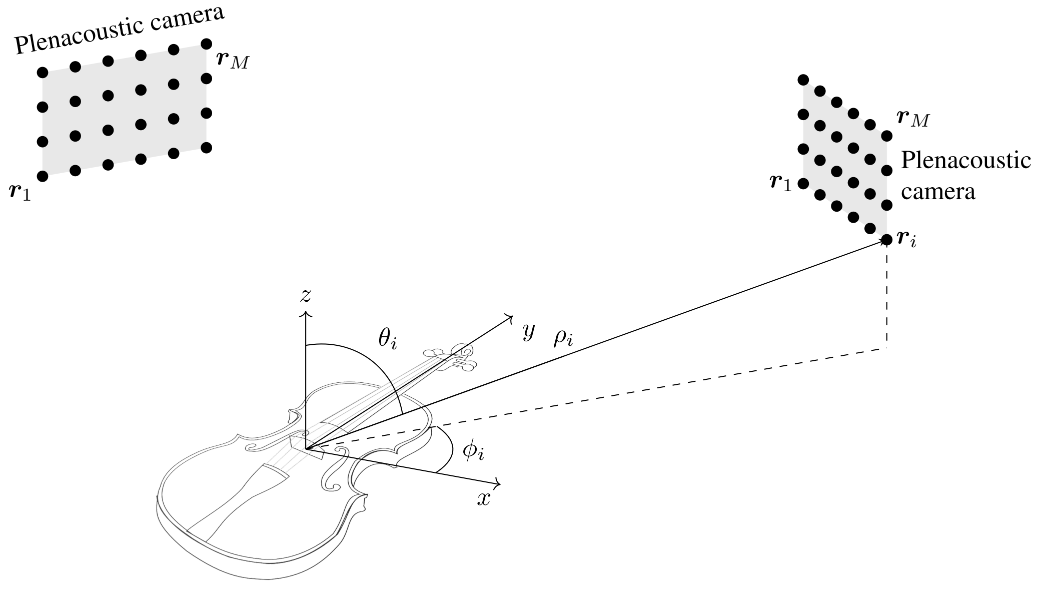


Figure 7.3: *The setup for the directivity pattern measurement with the adopted violin-centric reference system.*

Note that we implicitly assume a noiseless signal. From hereinafter the dependence on time of the signal $X(\cdot)$ and other quantities will be omitted for simplicity. Similarly to (5.3) we model the direct sound as the combination of the Green's function (3.29), the directivity and the source signal as

$$\begin{aligned} X_{\mathrm{dir}}(\omega, \boldsymbol{r}_i) &= \tilde{D}(\theta_i, \phi_i, \omega) H(\omega, \boldsymbol{r}_i) S(\omega) \\ &= \tilde{D}(\theta_i, \phi_i, \omega) \frac{e^{-j\frac{\omega}{c}\rho_i}}{\rho_i} S(\omega), \end{aligned} \tag{7.2}$$

where $\tilde{D}(\cdot)$ is the complex directivity function and in order to conveniently highlight the directional dependency of the signal we express the microphone location adopting spherical coordinates (see Section 3.1) centered in the acoustic center of the violin as shown in Figure 7.3 , hence

$$\boldsymbol{r}_i = [\rho_i \sin\theta_i \cos\phi_i, \rho_i \sin\theta_i \sin\phi_i, \rho_i \cos\theta_i]^T$$

with ϕ_i the azimuth, θ_i the inclination and ρ_i the radial distance from the source. We are interested in the estimation of the directivity pattern, namely the magnitude of the directivity $\tilde{D}(\cdot)$, and from the knowledge of the direct sound field component, we can retrieve the directivity pattern inverting (7.2)

$$D(\phi_i, \theta_i, \omega) = |\tilde{D}(\phi_i, \theta_i, \omega)| = \frac{\rho_i |X_{\mathrm{dir}}(\omega, \boldsymbol{r}_i)|}{|S(\omega)|}. \tag{7.3}$$

The directivity pattern can be conveniently expressed using the radial component of the spherical harmonics expansion (3.52), described in Section 3.6.2, as [54, 219, 285]

$$D(\phi_i, \theta_i, \omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} C_{lm}(\omega) Y_{lm}(\theta_i, \phi_i), \tag{7.4}$$

where $C_{lm}(\omega)$ are the coefficients of the spherical harmonics $Y_{lm}(\theta_i, \phi_i)$ (3.25).

In practice, the direct sound field (7.2) can only be estimated from (7.1) and the presence of the reverberant component will affect the final estimate of the directivity pattern (7.3).

Let us introduce the discrete version of the directivity pattern (7.4) obtained from the B sampled directions (θ_b, ϕ_b), $b = 1, \dots, B$ given by measurement procedure [54] as

$$\hat{\mathbf{d}}(\omega) = \mathbf{Y}\mathbf{c}^{(L)}(\omega), \tag{7.5}$$

where $\hat{\mathbf{d}}(\omega) \in \mathbb{R}^{B\times 1}$ is the vector of the estimated samples of the directivity pattern, $\mathbf{c}^{(\mathrm{L})}(\omega)$ is the vector of the spherical harmonic coefficients of order L associated to the directivity pattern

$$\mathbf{c}^{(L)}(\omega) = [C_{00}(\omega), C_{1-1}(\omega) \dots, C_{LL}(\omega)]^T \tag{7.6}$$

and

$$\mathbf{Y} = \begin{bmatrix} Y_{00}\left(\theta_1, \phi_1\right) & Y_{1-1}\left(\theta_1, \phi_1\right) & \cdots & Y_{LL}\left(\theta_1, \phi_1\right) \\ \vdots & \vdots & \ddots & \vdots \\ Y_{00}\left(\theta_B, \phi_B\right) & Y_{1-1}\left(\theta_B, \phi_B\right) & \cdots & Y_{LL}\left(\theta_B, \phi_B\right) \end{bmatrix}, \tag{7.7}$$

is the matrix containing the spherical harmonics as in (3.25). As explained in [54], an estimate of the spherical harmonics coefficients $\mathbf{c}^{(\mathrm{L})}(\omega)$ can be obtained by the inversion of (7.5)

$$\hat{\mathbf{c}}^{(L)}(\omega) = \mathbf{Y}^{\dagger}\hat{\mathbf{d}}(\omega), \tag{7.8}$$

where † represents the least-squares inverse operator.

The adoption of the spherical harmonics expansion gives us two main advantages. On the one hand, it provides the interpolation of the directivity pattern for arbitrary directions and on the other hand, the vector of the coefficients $\hat{\mathbf{c}}^{(L)}(\omega)$ (7.8) represents a compact and complete description of the data which can be exploited for the comparison of the patterns.

In order to retrieve the direct sound field component (7.2) and accurately estimate the directivity pattern (7.5), here we adopt the methodology of [54], where the authors exploit the plenacoustic analysis [7, 160, 161] of the sound field for robustly estimating the directivity pattern of sources.

The measurement setup is depicted in Figure 7.3, where two plenacoustic cameras are employed for analyzing the sound field. Each camera is composed by M microphones, hence the total number of sensors is $I = M \times 2$ and we assume that the location of each ith microphone is known and the signals are synchronized. The adoption of two plenacoustic cameras (Figure 7.3) let us analyze the acoustic radiation from two points of view enabling the accurate localization of the source and speeding up the procedure by measuring more directions simultaneously. The source signal $S(\omega)$ required in (7.3) is acquired by a microphone placed on the violin in the proximity of the bridge, while the instrument orientation is tracked by a 9 degrees-of-freedom (DOFs) Inertial Measurement Unit (IMU), consisting of a 3-axis compass, a 3-axis gyroscope and a 3-axis accelerometer sensors synchronized with the source signal.

In order to capture a sufficiently dense set of B directions in (7.5), the violinist is asked to continuously change the position and orientation while constantly playing a given note. We refer the reader to [54] for the detailed description of the measurement technique.

7.3 Tools for the Analysis of the Directivity Patterns

We introduce the tools that we use to compare the directivity patterns of the violins, and in general of any acoustic source.

The goal of our approach is twofold. On the one side, we aim at providing quantitative metrics that enable a summarized comparison of the directivity pattern, and on the other side we aim at defining quantities that emphasize the subtle differences that make one pattern different from the others.

7.3.1 Spherical-harmonics-based Tools

In order to characterize the directivity pattern in a compact fashion, the metrics are based on the spherical harmonic representation.

Normalized Cross Correlation Index (NCC) The Normalized Cross Correlation (NCC) of the directivity pattern of two violins is defined as

$$\mathrm{NCC}_{n,j}(\omega) = \frac{\hat{\mathbf{c}}_n^{(L)}(\omega)\hat{\mathbf{c}}_j^{(L)}(\omega)}{||\hat{\mathbf{c}}_n^{(L)}(\omega)||\,||\hat{\mathbf{c}}_j^{(L)}(\omega)||}, \tag{7.9}$$

where $\hat{\mathbf{c}}_n^{(L)}$ and $\hat{\mathbf{c}}_j^{(L)}$ are the Lth-order estimated spherical harmonics coefficients (7.8) of the nth and jth violin, respectively. Note that NCC is frequency dependent, therefore in order to provide further summarized comparison of the directivity patterns, we define the averaged NCC as

$$\overline{\mathrm{NCC}}_{n,j} = \frac{1}{W}\sum_{w=1}^{W}\mathrm{NCC}_{n,j}(\omega_w), \tag{7.10}$$

where $w = 1, \dots, W$ is the index of the analyzed frequency bin.

7.3.2 Analysis of the Principal Radiation Regions

In addition to a compact characterization of the directivity pattern, we are also interested in identifying the principal regions of emission from the directivity pattern, and analyzing how the directivity is spread around the principal directions.

Similarly to [169] we introduce tools aimed at extracting the principal radiation regions from the directivity pattern. The principal radiation regions of a directivity pattern are defined as the set of angles $\mathcal{P}$ for which the value of the directivity pattern is larger than a threshold τ

$$\mathcal{P}(\omega) = \left\{(\bar{\theta}, \bar{\phi}) : \hat{\mathbf{d}}_{\mathrm{dB}}(\omega, \theta, \phi) \geq \tau\right\}, \tag{7.11}$$

where $\hat{\mathbf{d}}_{\mathrm{dB}}(\cdot) = 10\log_{10}\frac{|\hat{\mathbf{d}}(\cdot)|}{\max|\hat{\mathbf{d}}(\cdot)|}$ is the normalized directivity pattern in decibel and we adopt $\tau = -3\,\mathrm{dB}$. It is worth to underline that the angles (ϕ, θ) of $\hat{\mathbf{d}}_{\mathrm{dB}}$ in (7.11) can be different from the sampled ϕ_b, θ_b of the measurements in (7.5) thanks to the interpolation given by the spherical harmonics representation (7.8).

Consequently, we can define the segmented directivity pattern according to its principal radiation regions as

$$\overline{\mathbf{d}}(\theta, \phi, \omega) = \begin{cases} 1 & (\theta, \phi) \in \mathcal{P}(\omega) \\ 0 & \text{otherwise} \end{cases} . \tag{7.12}$$

The binary pattern $\overline{\mathbf{d}}(\cdot)$ in (7.12) is used for comparing directivity patterns of different instruments.

The segmentation (7.12) obtained through (7.11), results in areas of the directivity pattern related to the maximum energy emission. Such areas can be arbitrarily shaped. They do not extract, therefore, a single preferred direction of emission, rather they focus on an area of maximum emission. Nevertheless, this approach is preferable with respect to considering as a principal direction the direction related to the maximum of $\hat{\mathbf{d}}(\omega)$, as the proposed approach is more robust to measurement errors.

In order to obtain a point-like descriptor of a principal radiation region in a directivity pattern, we define the center of mass

$$\mathbf{r}(\omega) = \frac{1}{\boldsymbol{M}} \sum_{b \in \mathcal{P}(\omega)} \mu_b \boldsymbol{r}_b, \tag{7.13}$$

where $\boldsymbol{r}_b = \left[\sin\bar{\theta}\cos\bar{\phi}, \sin\bar{\theta}\sin\bar{\phi}, \cos\bar{\theta}\right]^T$ are the points belonging to the principal radiation regions on a unitary sphere which are weighted by the normalized directivity pattern

$$\mu_b = \frac{\hat{\mathbf{d}}(\theta, \phi)}{\max \hat{\mathbf{d}}}, \tag{7.14}$$

and

$$\boldsymbol{M} = \sum_{b \in \mathcal{P}(\omega)} \mu_b. \tag{7.15}$$

Principal radiation region probability map ($\mathcal{M}$) Following the same methodology of the histograms of the principal radiation directions in [169], we can define, for a set of N violins, the principal direction probability map as

$$\mathcal{M}(\theta, \phi, \omega) = \frac{1}{N} \sum_{n=1}^{N} \overline{\mathbf{d}}_n(\theta, \phi, \omega). \tag{7.16}$$

where $\overline{\mathbf{d}}_n(\cdot)$ is the segmented binary directivity pattern of the nth violin. Therefore, for a fixed frequency ω, $\mathcal{M}(\theta, \phi, \omega)$ is a sample estimate of the probability for the direction (θ, ϕ) to belong to a principal radiation region.

Jaccard similarity index (JSI) The JSI between the binarized directivity patterns (7.12) of two violins is defined as

$$\mathrm{JSI}_{n,j}(\omega) = \frac{|\overline{\mathbf{d}}_n(\omega) \cap \overline{\mathbf{d}}_j(\omega)|}{|\overline{\mathbf{d}}_n(\omega) \cup \overline{\mathbf{d}}_j(\omega)|}, \tag{7.17}$$

where $\overline{\mathbf{d}}_n(\omega)$, $\overline{\mathbf{d}}_j(\omega)$ are the binarized principal direction patterns of the nth and jth violin, respectively.

In other words, the JSI measures the similarity of two binary patterns, therefore, $\mathrm{JSI}_{n,j}(\omega) = 0$ when they are disjoint, i.e., no portion of principal radiation region is shared between the two violins, and $\mathrm{JSI}_{n,j}(\omega) = 1$ when they perfectly match. Similarly to $\overline{\mathrm{NCC}}$ (7.10), with $\overline{\mathrm{JSI}}$ we refer to the average over frequency of JSI (7.17).

Center of mass distance (CMD) In order to evaluate the difference in terms of direction between two directivity patterns, we define the distance between the centers of mass as

$$\mathrm{CMD}_{n,j}(\omega) = \arctan\left(\frac{|\mathbf{r}_n(\omega) \times \mathbf{r}_j(\omega)|}{\mathbf{r}_n(\omega) \cdot \mathbf{r}_j(\omega)}\right), \tag{7.18}$$

where the operators × and · denote the vector cross and dot products, respectively, and $\mathbf{r}_k(\omega)$ is the center of mass (7.13) of kth violin directivity patterns. It is worth noticing that when more than one region of principal directions is found, we average the $\mathrm{CMD}_{n,j}(\omega)$ of the closest centers of mass **r** of the two directivity patterns. Furthermore, similarly to the other quantities, let us denote with $\overline{\mathrm{CDM}}_{n,j}$ the frequency averaged version of (7.18).

7.4 Violin Directivity Pattern Analysis

The proposed tools for the analysis and comparison of directivity patterns allow us to analyze a set of violins extracting the main directional characteristics of their radiation. We provide the details of the measurement setup and the set of analyzed violins. Moreover, A general characterization of the directivity pattern of the violins and a comparison of the instruments exploiting the proposed tools are given.

7.4.1 Setup and Dataset Description

The measurements were performed with the setup of [54] in an environment characterized by a reverberation time of $T_{60} \approx 20\,\mathrm{ms}$. Each pleanacoustic camera (Figure 7.3) consists of $M = 4 \times 8 = 32$ omnidirectional microphones (Beyerdynamic MM1), hence the total number of employed sensors is $I = M \times 2 = 64$. The estimate of the source signal is provided by a proximity microphone (T.Bone Ovid System CC 100) mounted on the instrument close to the bridge, while the violin orientation is tracked by the 9-DOFs IMU (Phidgets Spatial Precision 3/3/3). For the details on the measurement setting and parameters setup, we refer the reader to Sec. VI-VII of [54].

The violins were played by a professional violinist. In particular, we asked to play the four open strings (G-D-A-E, respectively) at an intensity corresponding in the standard music notation to *mezzo-forte*. Therefore, the radiated acoustic energy is concentrated over a discrete set of frequencies associated to the fundamental and the harmonics of each string. We denote with $\omega^{(s)}$, $s \in (\mathrm{G}, \mathrm{D}, \mathrm{A}, \mathrm{E})$ the set of frequencies related to each string that go from the fundamental to an arbitrary maximum frequency. This allows us to measure the directivity for a total of $W = 60$ frequencies going from around $196\,\mathrm{Hz}$ (the G string fundamental) up to $5\,\mathrm{kHz}$.

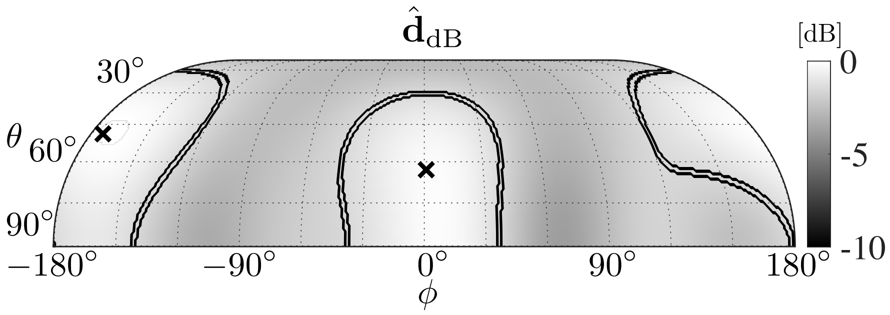


Figure 7.4: *The normalized directivity pattern* $\hat{\mathbf{d}}_{\mathrm{dB}}(\omega)$ *of the violin Vesuvius by A. Stradivari. The* $\hat{\mathbf{d}}_{\mathrm{dB}}$ *is computed at* $4400\,\mathrm{Hz}$ *with* $L = 4$*. The principal radiation regions* $\mathcal{P}$ *are delimited by the black lines, while the centers of mass* $\mathbf{r}$ *are marked with a black cross.*

n	Violin maker	Violin Name	Year
0	*A. Amati*	*Carlo IX*	1566
1	*N. Amati*	*Hammerle*	1658
2	*N. Amati*	*Ex Collin*	1669
3	*A. Stradivari*	*Clisbee*	1669
4	*Guarneri del Gesù*	*Quarestani*	1689
5	*A. Stradivari*	*Joachim Ma*	1714
6	*A. Stradivari*	*Il Cremonese*	1715
7	*A. Stradivari*	*Vesuvius*	1727
8	*A. Stradivari*	*Scotland University*	1734
9	*Guarneri del Gesù*	*Stauffer*	1734
A	*Elena Bardella*	*Twin A*	2015
B	*Elena Bardella*	*Twin B*	2015

Table 7.1: *The set of violins under study.*

We limited our analysis to the upper hemisphere of the directivity pattern, i.e., taking only $\theta \leq \pi/2$ in (7.5), as it is known that the lower hemisphere contributes to the overall radiation in a minor fashion [278].

We limited our analysis to the upper hemisphere of directivity pattern, i.e., taking only $\theta \leq \pi/2$ in (7.5), as it is known that the lower part contributes to the overall radiation in a minor fashion [278]. In Figure 7.4, an example of the directivity pattern (7.5) expressed in decibel is shown. The axes in Figure 7.4 are referred to the reference system of Figure 7.3 and they are adopted throughout the analysis. In addition to the directivity pattern, in Figure 7.4 the region of the principal radiation (7.11) is enclosed by a bold line, and the corresponding centers of mass (7.13) are highlighted by black crosses. We organize the violins in two groups: historical and twin violins, as reported in Table 7.1. The first set of analyzed violin is composed of $N = 2$ *twin* violins. The aim of the analysis on *twin* violins is to establish a benchmark of similarity, as we expect that they will yield very similar patterns if compared to the other instruments.

The two instruments were made by *Elena Bardella*, an Italian violin maker. These two instruments were built purposely for research goals. The geometry is identical, under the limits of a manual construction (precision up to $\frac{1}{20}$ mm). The mechanical properties of the wood are equal for the two violins, as they were carved from the same wood blocks. The only aspect for which the two instruments differ is the varnishing, as alcohol and oil varnishes have been applied. Therefore, we expect the twins to provide very similar directivity patterns. In order to distinguish the twin violins from the historical instruments, we indicate them with the subscripts A and B (see Table 7.1).

The second set includes $N = 10$ prestigious historical instruments made available by the "*Museo del Violino*". These violins were built by some of the most renowned old Italian masters. The list of the instruments along with the violin maker and building year is reported in Table 7.1. To the best of our knowledge, this is the first analysis on the directivity pattern on a large set of valuable instruments by different masters. More specifically, we measured 5 violins by *Antonio Stradivari*, 2 by *Giuseppe Guarneri del Gesù*, 2 by *Nicolò Amati* and 1 by *Andrea Amati*.

In Table 7.1 we assign to each instrument an index $v = 0, \ldots, 9$, to be used in Sec. 7.3.

7.4.2 General Characterization

As a first analysis, we aim at characterizing the general directional behavior of the violins adopting the tools presented in Section 7.3.

An important parameter, for the accurate estimation of the directivity pattern is the spherical harmonic order L in (7.8). On the one hand, a high order L allows more precision, tracking the details of the shape of the pattern. On the other hand, by including irrelevant spherical harmonics orders we increase the measurement noise in the data. In order to empirically determine a suitable value of L, we examined the energy associated to the coefficients as a function of the order. As expected, the energy associated to higher spherical harmonics order increases with the frequency, i.e. low-frequency patterns are more omnidirectional than high-frequency ones [39, 41, 113], and therefore, have most of the energy concentrated in the zeroth order coefficient. We discovered that, on average, $99\,\%$ of the signal energy is associated to $L \leq 4$. Based on this preliminary analysis, we introduce a frequency dependent order L_ω defined as

$$L_\omega = \begin{cases} 1 & 0 < \omega \leq 500\,\mathrm{Hz} \\ 2 & 500\,\mathrm{Hz} < \omega \leq 800\,\mathrm{Hz} \\ 3 & 800\,\mathrm{Hz} < \omega \leq 3.3\,\mathrm{kHz} \\ 4 & 3.3\,\mathrm{kHz} < \omega \leq 5\,\mathrm{kHz} \end{cases}, \tag{7.19}$$

The frequency-dependent spherical harmonic order turns out to be useful for analyzing the directivity pattern shapes with the tools for the analysis of the principal radiation regions (see Section 7.3.2), since an overestimation of the spherical harmonics expansion order could affect the shapes of the principal radiation regions. As regards, the NCC (7.9) we fix $L = 4$ for all the analyzed frequencies such that the $\overline{\mathrm{NCC}}$ is referred to the same spherical harmonics order. In Figure 7.5, Figure 7.6, Figure 7.7 and Figure 7.8 $\mathcal{M}(\theta, \phi, \omega^{(s)})$ (7.16) related to each string is depicted. Note that a different number of patterns is associated to each string. In particular, for $G,\ D\ A,\ E$ strings we measure 25, 17, 11, 7 patterns, respectively. As regards the A string 11 $\mathcal{M}(\theta, \phi, \omega^{(\mathrm{A})})$ are depicted in Figure 7.7, while 7 harmonics related to the E string are available in the analyzed frequency range (see Figure 7.8).

At frequencies lower than $800\,\mathrm{Hz}$ we can observe that the violins tend to radiate in an omnidirectional fashion. Nevertheless, a region of principal directions toward $\phi = 0°$ can be observed for $(294, 587)$Hz in Figure 7.6, while a dipole-like radiation

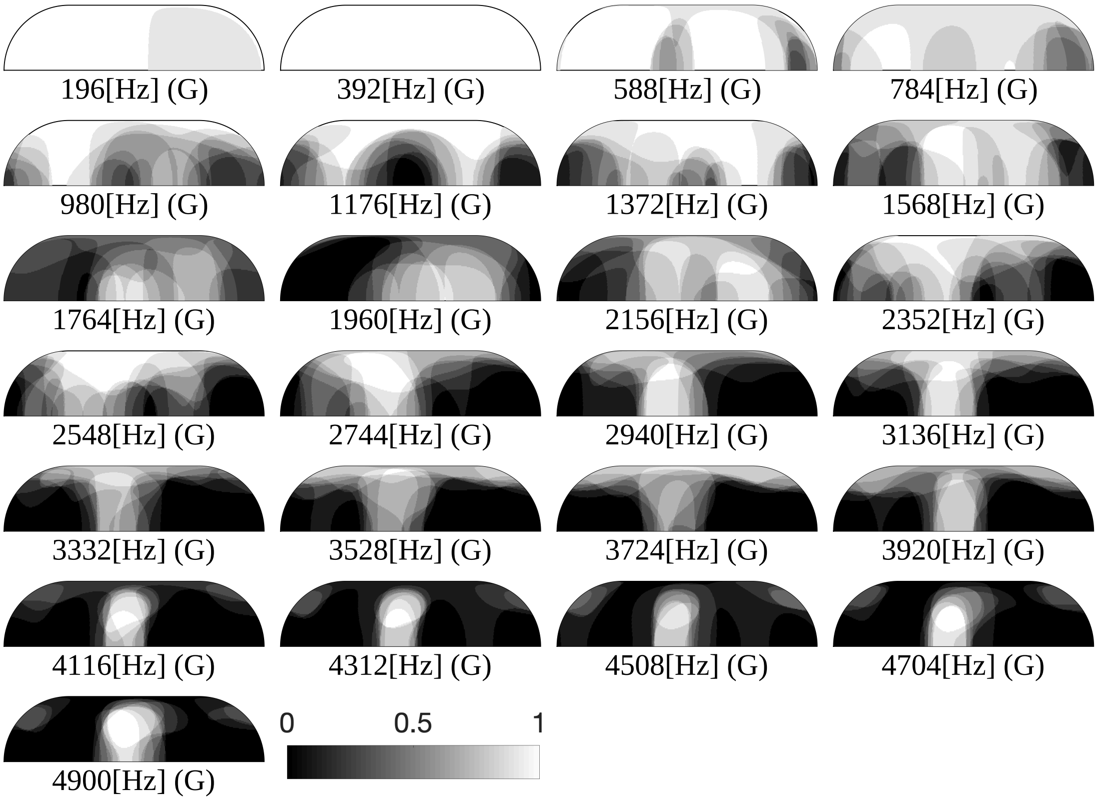


Figure 7.5: *The principal radiation region probability maps $\mathcal{M}(\theta, \phi, \omega^{(\mathrm{G})})$ associated to the harmonics of the* G *string. Only the historical violins are considered.*

occurs at $588\,\mathrm{Hz}$ directed towards $\phi = (80, -80)°$. These results are in line with the ones reported by [169, 170].

As regards the G string, in Figure 7.5, we can note that in the range $(588 - 1372)\mathrm{Hz}$ the principal radiation regions are similar to a dipole-like radiation, with lobes on average directed towards, $\phi = (80, -80)°$ and $\theta = 0°$. At high frequencies ($\omega^{(\mathrm{G})} \geq 4116\,\mathrm{Hz}$), $\mathcal{M}(\theta, \phi, \omega^{(\mathrm{G})})$ clearly shows a narrower region of principal directions at $\phi = -15°$ and $\theta = 40°$ that is shared by all the violins, while one third of the analyzed instruments radiates strongly also around $\phi = -180°$ and $\theta = 40°$. As expected, in the transitional frequency range [289], the regions of the principal directions are large and only a few directions of preferential emissions are shared by multiple instruments. This confirms the variability of the directivity pattern in that frequency range, as also documented in [196, 284].

A similar behaviour can be observed for the harmonics of the D string, shown in Figure 7.6. In this case the transitional region extends at frequencies above $3\,\mathrm{kHz}$. At frequencies higher than $4\,\mathrm{kHz}$ the principal radiation regions are located on the upper part of the hemisphere ($\theta_b = 0°$) i.e., directions perpendicular to the top plate.

As far as the A string is concerned, in Fig. 7.7 we can identify a preferred region of emission at $\phi_b = 0°$ that is shared by almost all the harmonics in the low and transitional frequency ranges. Starting from $3520\,\mathrm{Hz}$, a clear dipole-like radiation is shown by the majority of the violins. This behavior is similar to what can be observed from

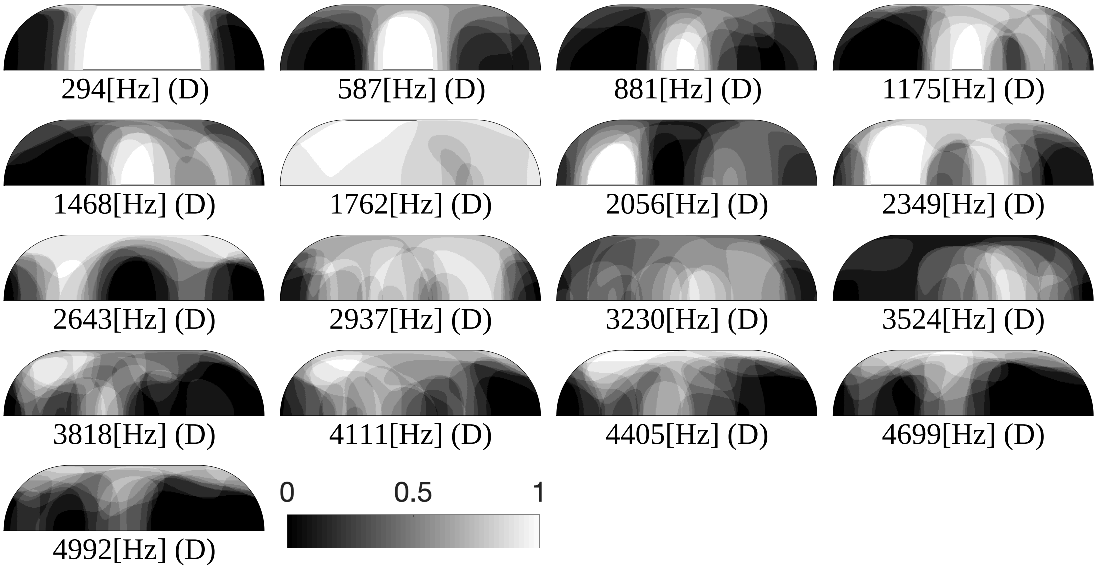


Figure 7.6: *The principal radiation region probability maps $\mathcal{M}(\theta, \phi, \omega^{(\mathrm{D})})$ associated to the harmonics of the D string. Only the historical violins are considered.*

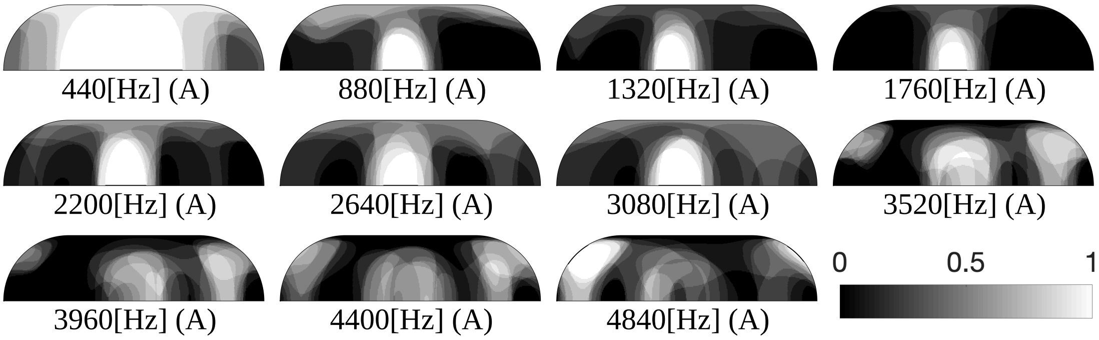


Figure 7.7: *The principal radiation region probability maps $\mathcal{M}(\theta, \phi, \omega^{(\mathrm{A})})$ associated to the harmonics of the A string. Only the historical violins are considered.*

the principal radiation region probability maps of the highest G string harmonics (see Fig. 7.5).

Finally, $\mathcal{M}(\theta, \phi, \omega^{(\mathrm{E})})$ in Figure 7.8, shows that the principal directions are less variable with respect to the other strings. In Figure 7.8, the regions of the principal directions are wide and they tend to be steady in the range $\theta = (90, 0)^\circ$ and $\phi = (-20, -90)^\circ$.

Overall, we can observe that, except the lowest frequencies, where the directivity patterns tend to be less directional, all the violins under analysis express a clear directional characteristic. This is particularly evident for frequencies belonging to the transitional region, where only very narrow regions of principal radiation are shared by all the instruments. Nonetheless, a relevant overlap of the principal radiation regions for a wide range of the analyzed frequencies is present. This suggests that all the instruments exhibit a common radiation behavior. Therefore, we can identify some shared principal radiation regions that can be observed for the directions towards the audi-

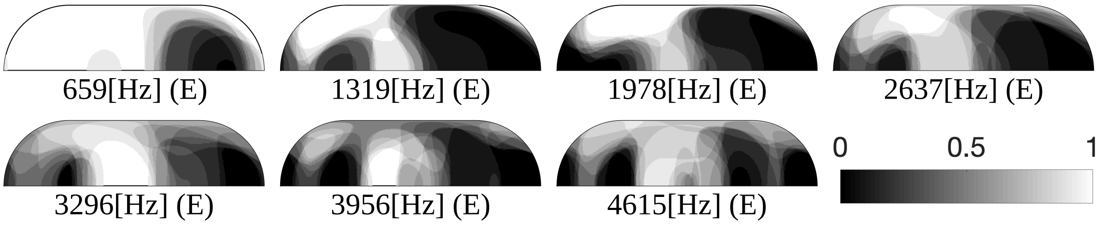


Figure 7.8: *The principal radiation region probability maps $\mathcal{M}(\theta, \phi, \omega^{(\mathrm{E})})$ associated to the harmonics of the* E *string. Only the historical violins are considered.*

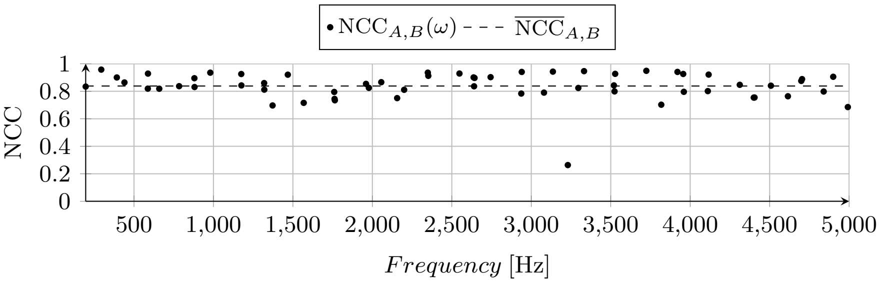


Figure 7.9: *The Normalized Cross Correlation* $\mathrm{NCC}_{A,B}(\omega)$ *of the twins violin as a function of the frequency along with the average value* $\overline{\mathrm{NCC}}_{A,B}$.

ence, approximately around $\theta \in (45°, 90°)$, $\phi \in (-20°, 20°)$ and above the instrument around $\theta \in (30°, 0°)$, $\phi \in (-90°, 0°)$.

7.4.3 Comparison between Instruments

Analysis on twin violins First, in order to validate the proposed metrics, we propose the comparison between the twin violins. In Figure 7.9 we report $\mathrm{NCC}_{A,B}(\omega)$ along with the average value $\overline{\mathrm{NCC}}$. We can note that for a wide range of the analyzed frequencies ($> 70\,\%$) $\mathrm{NCC}_{A,B}(\omega) > 0.8$ while the average value over the whole frequency range is $\overline{\mathrm{NCC}} = 0.84$. In the frequency range between $1\,\mathrm{kHz}$ and $2\,\mathrm{kHz}$, a decrease of $\mathrm{NCC}_{A,B}(\omega)$ can be observed with a minimum of 0.7 at $1372\,\mathrm{Hz}$. This decrease in $\mathrm{NCC}_{A,B}(\omega)$ is not unexpected, since in the transitional frequency range the directivity patterns are known to be more variable. The absolute minimum of $\mathrm{NCC}_{A,B}(\omega)$ is registered at $3230\,\mathrm{Hz}$, where the violin A presents an overall omnidirectional directivity pattern, while the other instrument is relatively more directional. In order to evaluate the differences in terms of principal radiation regions, we compare the binary patterns $\overline{\mathbf{d}}(\cdot)$ (7.12) in terms of JSI (7.17). In Figure 7.10 we show $\mathrm{JSI}_{A,B}(\omega)$ along with its average value $\overline{\mathrm{JSI}}_{A,B} = 0.67$. In general, the trend of $\mathrm{JSI}_{A,B}(\omega)$ confirms the similarity of the directivity patterns observed through the analysis of the Normalized Cross Correlation shown in Figure 7.9. More specifically, more than the $60\,\%$ of the analyzed frequencies have a $\mathrm{JSI}_{A,B}(\omega) > \overline{\mathrm{JSI}}_{A,B}$. Similarly to $\mathrm{NCC}_{A,B}(\omega)$, in the frequency range between $1\,\mathrm{kHz}$ and $2\,\mathrm{kHz}$ there is a decrease in $\mathrm{JSI}_{A,B}(\omega)$.

It is possible to notice that, with respect to $\mathrm{NCC_{A,B}}$, $\mathrm{JSI}_{A,B}(\omega)$ is on average lower. This is due to the fact that the Jaccard Similarity Index tends to emphasize minor differences in the principal directions between the violins under analysis. In fact the JSI is devoted to the analysis of the principal radiation regions only, hence it evaluates a

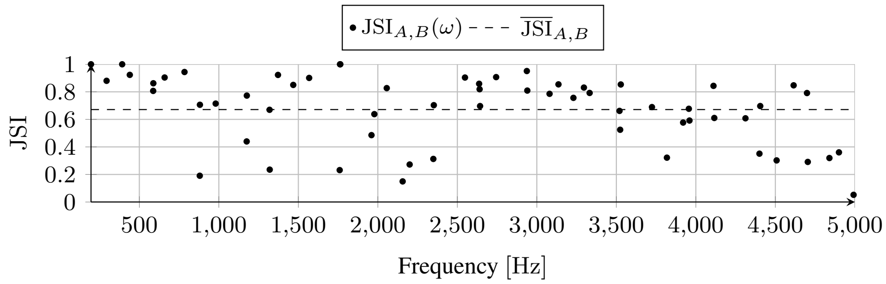


Figure 7.10: *The Jaccard Similarity Index* $\mathrm{JSI}_{A,B}(\omega)$ *of the twins violin as a function of the frequency along with the average value* $\overline{\mathrm{JSI}}_{A,B}$.

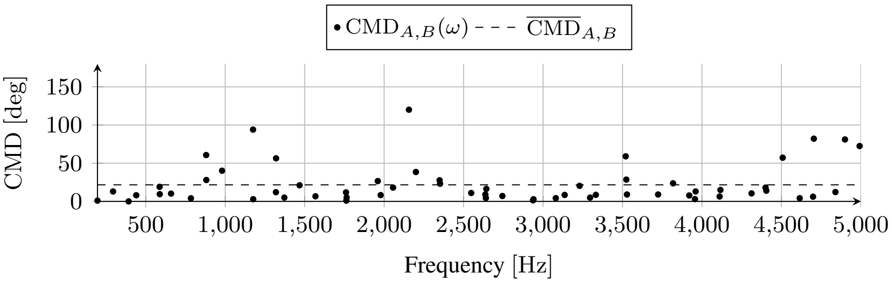


Figure 7.11: *Distance between the centers of mass,* $\mathrm{CMD}_{A,B}(\omega)$*, of the twins violin as a function of the frequency along with the average value* $\overline{\mathrm{CMD}}_{A,B}$.

section of the patterns.

Finally, in Figure 7.11, the distance between the centers of mass $\mathrm{CMD}_{A,B}(\omega)$ is depicted for the analyzed frequency range along with the average value. The average value of the distance between the centers of mass is $\overline{\mathrm{CMD}}_{A,B} = 21°$ and we can note that for a wide range (70 %) of the analyzed frequencies, $\mathrm{CMD}_{A,B}(\omega) < \overline{\mathrm{CMD}}_{A,B}$. It is worth noting that the overall trend of $\mathrm{CMD}_{A,B}(\omega)$ is related with $\mathrm{JSI}_{A,B}(\omega)$. This behavior is not surprising, since similar regions of principal directions are likely to present similar centers of mass $\mathbf{r}$ (7.13). The results provided by the proposed tools for the *twin* violins validate the proposed metrics, as instruments with almost identical geometries and built from the same wood blocks are likely to exhibit the same directional behavior. These results provide a benchmark for the comparison of the historical instruments, as we expect that the similarity obtained for the *twin* violins will be unmatched by other violin pairs.

Analysis of the directivity patterns of the historical violins As a first comparison, we compute $\mathrm{NCC}_{n,j}$ with $n = 0, \ldots, 8$, $j = 1, \ldots, 9$ and $n \neq j$ between all the historical violins. By looking at the definition of $\mathcal{M}(\cdot)$ in Section 7.4.2, we expect a dependency on the specific string under analysis (see Figure 7.7 and Figure 7.8 as opposed to Fig. 7.6 and Figure 7.5). For this reason, in Figure 7.12 we report $\overline{\mathrm{NCC}}_{n,j}(\omega^{(s)})$, computed at the frequencies related to each string. Overall, $\overline{\mathrm{NCC}}^{(G)}$ in Figure 7.12(a) and $\overline{\mathrm{NCC}}^{(D)}$ Figure 7.12(b) present a less variable correlation with respect to $\overline{\mathrm{NCC}}^{(A)}$ in Figure 7.12(c) and $\overline{\mathrm{NCC}}^{(E)}$ Figure 7.12(d). A clear difference between the individual violins can be observed in Figure 7.12(d), especially as regards the violin $n = 2$ *Ex Collin* by *N. Amati*

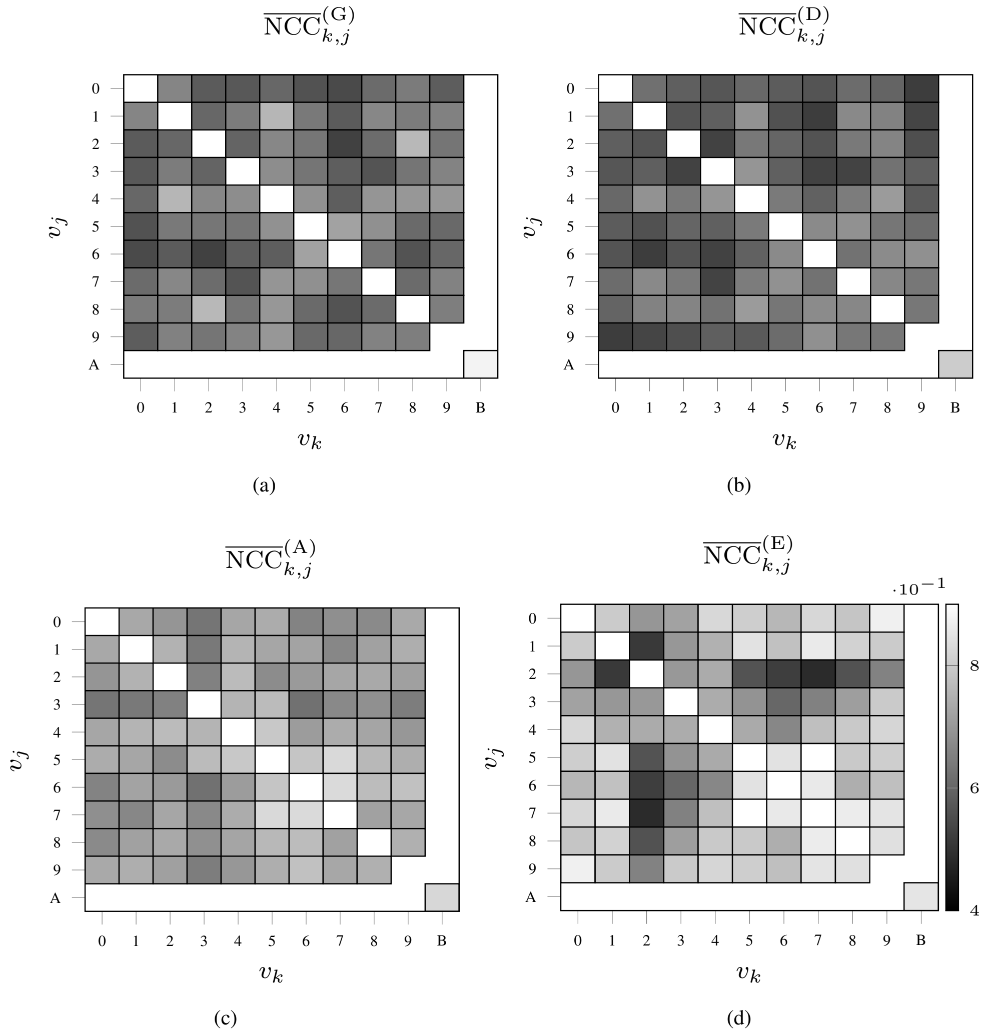


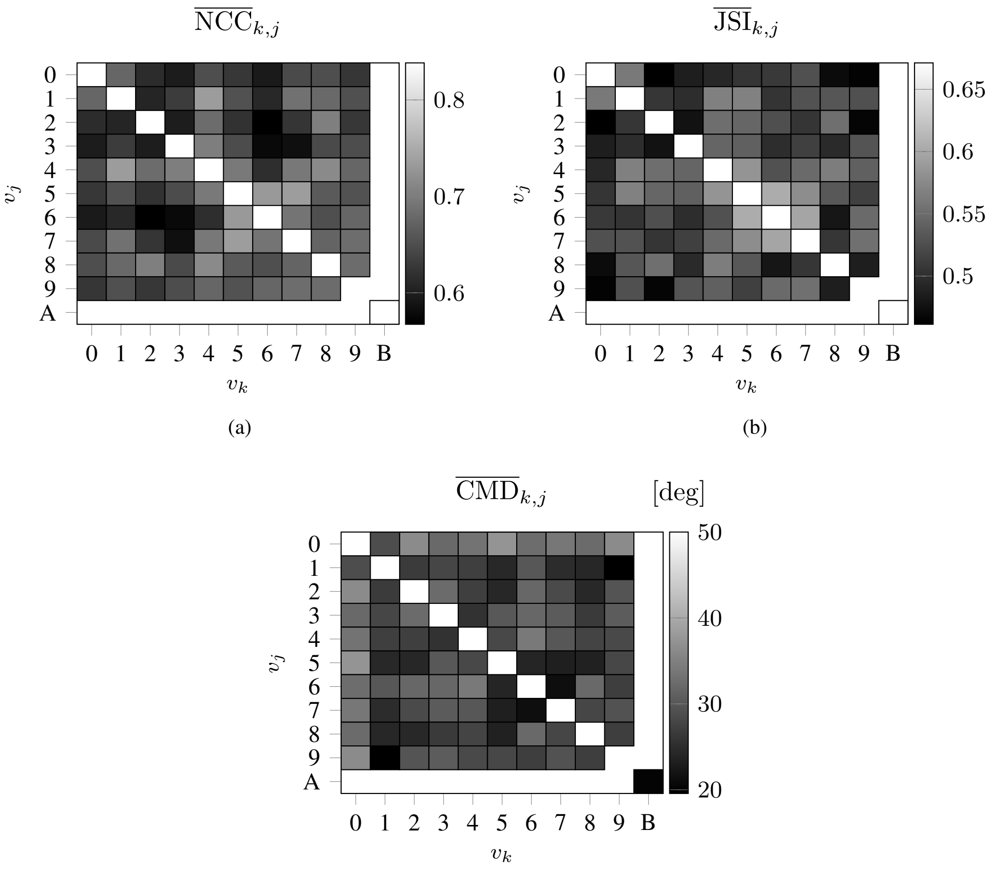
Figure 7.12: *Normalized Cross Correlation of the directivity pattern of the historical violins subdivided by strings. Metrics values related to the twin violins ($v = \mathrm{A}, \mathrm{B}$) are reported for comparison.*

(see Table 7.1) which presents the lowest correlation values.

In order to provide a compact comparison of the violins, the Normalized Cross Correlation is averaged over the whole W frequencies to yield $\overline{\mathrm{NCC}}$ and this is shown in Figure 7.13(a).

Some interesting conclusions can be drawn.

First of all, the average correlation among the historical instruments is 0.65. The lowest value is $\overline{\mathrm{NCC}}_{2,6} = 0.55$, and it corresponds to the pair *Ex Collin* and *Stradivarius*' *Il Cremonese*, while the highest value $\overline{\mathrm{NCC}}_{5,7} = 0.73$ is between *Vesuvius* and *Joachim Ma*, both by *Stradivarius*. Interestingly, the violin $v = 6$ *Il Cremonese* by *A. Stradivarius* reports a good correlation (higher than the average), with other instrument by Stradivarius of the same period ($v = 5, 7$, *Joachim Ma* and *Vesuvius*, respectively) and later instruments rather than older violins. These instruments by Stradivarius also exhibit the highest value of $\overline{\mathrm{NCC}}^{(E)}$ as in Figure 7.12(d). Note that the maximum value $\overline{\mathrm{NCC}}_{5,7} = 0.73$ in Figure 7.13(a) is around 0.1 lower than the average correlation



(a)

(b)

(c)

Figure 7.13: *(a)* $\overline{\mathrm{NCC}}$*, (b)* $\overline{\mathrm{JSI}}$*, and (c)* $\overline{\mathrm{CMD}}$ *of the set of historical violins. Metrics values related to the twin violins (*$v = \mathrm{A}, \mathrm{B}$*) are reported for comparison.*

obtained for the *twin* violins $\overline{\mathrm{NCC}}_{A,B} = 0.84$ (see Figure 7.9).

For what concerns the instruments by *Guarneri del Gesù*, $\overline{\mathrm{NCC}}_{9,j}$ tends to be monotonically increasing with j, i.e., they are more similar to modern instruments rather than older. A relatively higher value is shown also for $v = 4$ *Quarestani*, the other instrument of the set made by this historical maker.

Among the oldest violins, it is possible to notice that the *Hammerle* and the *Quarestani* ($v = 1, 4$) present a *good* correlation, as shown from $\overline{\mathrm{NCC}}^{(G)}_{1,4}$ in Figure 7.12(a) and $\overline{\mathrm{NCC}}^{(D)}_{1,4}$ in Figure 7.12(b). Considering the whole set of instruments, the oldest violin *Carlo IX* by *A. Amati* ($v = 0$) appears less correlated with the rest of the set (see Figure 7.13(a)). In fact, $\overline{\mathrm{NCC}}$ for this instrument is consistently below the average $\overline{\mathrm{NCC}}$ of the set, except for *Hammerle* $\overline{\mathrm{NCC}}_{0,1} = 0.67$ i.e., the violin closest in time to *Carlo IX*. Comparing the results of $\overline{\mathrm{NCC}}$ for the strings in Figure 7.12 with $\mathcal{M}$ of the general analysis in Section 7.4.2, we can confirm the similarity between the directivity patterns of different instruments for the harmonics of the A and E strings, while G and D strings tend to differ among the violins.

In Figure 7.14 $\mathrm{JSI}_{n,j}(\omega)$ is reported for all the pairs of violin in the set. As expected, at the lowest frequencies we can observe the highest $\mathrm{JSI}_{n,j}(\omega)$, since almost every

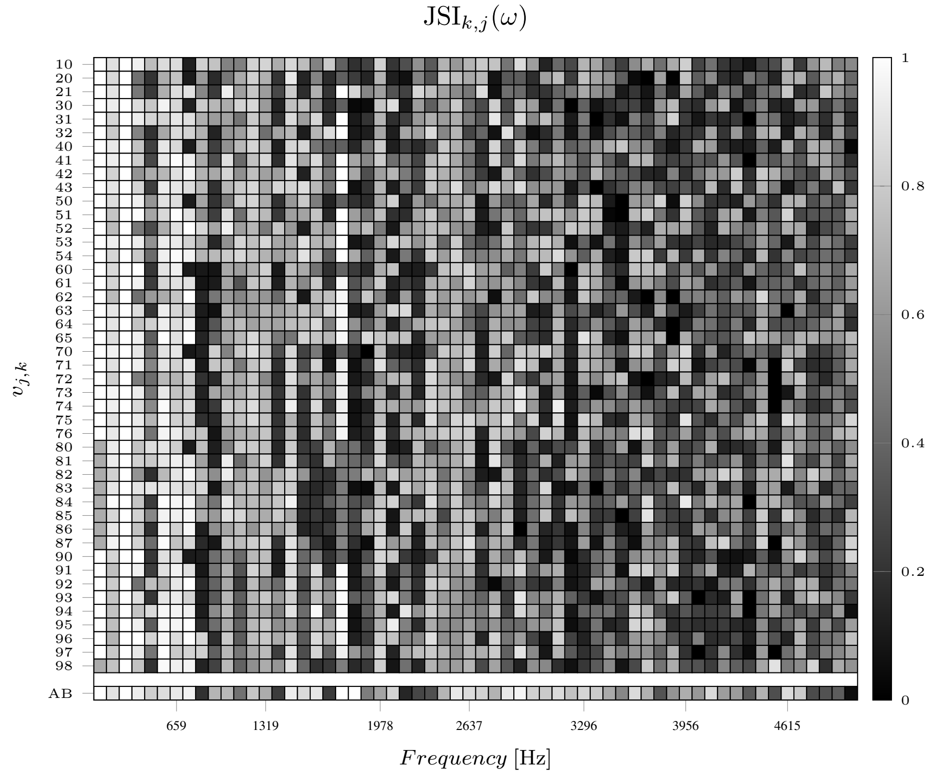


Figure 7.14: *The* $\mathrm{JSI}(\omega)$ *metrics of the historical violins.* $\mathrm{JSI}_{\mathrm{A,B}}(\omega)$ *related to the twin violins is reported for comparison.*

violin exhibits an omnidirectional radiation. It is worth noting that at 196 Hz, only the *Scotland University* ($n = 8$) is characterized by JSI $\neq 1$. In general, JSI tends to decrease as the frequency increases. Nevertheless, we can note that at some specific frequencies the principal directions are shared among the violins, resulting in vertical stripes in Figure 7.14. As an example, at 1762 Hz, all the violins report JSI > 0.9 (see also Figure 7.6) with the exception of $v = 0$ (*Carlo IX*) and $v = 8$ *Scotland University*. The violin $v = 0$ *Carlo IX* shows a low JSI with respect to the other violins also at 784 Hz.

The common directional behavior of the first harmonics of the E string observed in Fig. 7.8, can be identified in Figure 7.14, where $\mathrm{JSI}(\omega^{(E)})$ is greater than the average JSI of the whole dataset up to 2637 Hz.

Other than the lowest frequencies where, as expected, the JSI is relatively high, in the frequency range between around 1 kHz and 1.4 kHz we can observe in Figure 7.14 an increase in the JSI meaning that instruments exhibit similar principal radiation regions. A similar trend is observed also at frequencies around 2.3 kHz and 2.7 kHz (see Figure. 7.14). Ultimately, JSI quantifies the differences between the directional characteristics of individual instruments and the trend in the results confirms the well-known variability of the directivity patterns in the transitional modes frequency region [289], also observed in the principal radiation maps of Section 7.4.2.

In Figure 7.13(b) we report $\overline{\mathrm{JSI}}$ averaged over the whole frequency range under analysis. From the inspection of Figure 7.13(b), we can note that the overall trend of

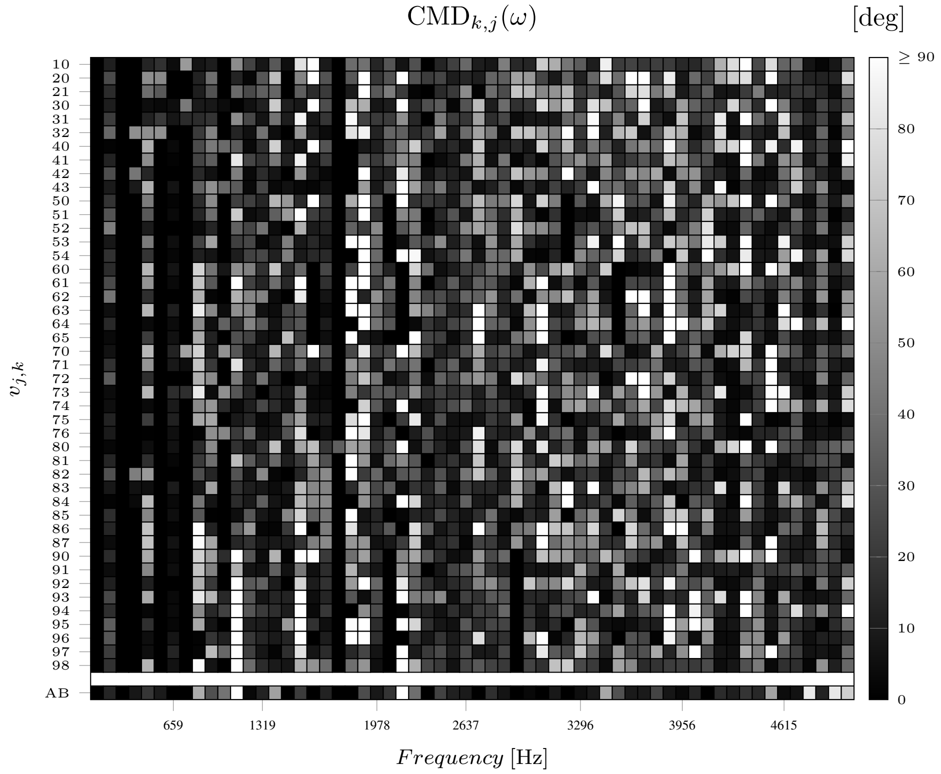


Figure 7.15: *The* $\mathrm{CMD}(\omega)$ *metrics of the historical violins.* $\mathrm{CMD}_{\mathrm{A,B}}(\omega)$ *related to the twin violins is reported for comparison.*

$\overline{\mathrm{JSI}}$ is similar to $\overline{\mathrm{NCC}}$ of Figure 7.10. Again, $\overline{\mathrm{JSI}}_{A,B}$ of the *twins* (see Figure 7.10) is higher than the maximum reported in Figure 7.13(b) that corresponds to $\overline{\mathrm{JSI}}_{5,6} = 0.6$. Similarly to what happens with the NCC metrics, the *Carlo IX* ($n = 0$) presents the lowest overall $\overline{\mathrm{JSI}}$ (see Figure 7.13(b), with peculiar differences as noted in Figure 7.14.

Noteworthy is the fact that while the second highest value of $\overline{\mathrm{NCC}}$ is associated to violins *Hammerle* ($v = 1$) and *Quarestani* ($v = 4$), their $\overline{\mathrm{JSI}}$ is lower. This result can be interpreted as deviations of their principal radiation regions that lower $\overline{\mathrm{JSI}}$, while the overall patterns provides high correlation values. On the other hand, the violins by *Stradivarius* that provide the highest $\overline{\mathrm{NCC}}$ values in Figure 7.13(a) show a good agreement of the their principal radiation regions, as confirmed by JSI in Figure 7.13(b).

It is important to notice that while NCC is based on the spherical harmonic coefficients of the directivity patterns, JSI is defined for comparing the principal radiation regions of the patterns. Hence, from the obtained results we can infer that the principal radiation regions carry most of the information of the directivity patterns.

For what concerns the CMD metrics defined in (7.18), we report in Figure 7.15 CMD of the different combinations of violins as a function of the frequency. Inspecting Figure 7.15, we can note that as expected and similarly to the case of the *twin* violins, $\mathrm{CMD}(\omega)$ reflects the trend of $\mathrm{JSI}(\omega)$ (see Figure 7.14). In fact, the frequency ranges that provide higher JSI in Fig.7.14 are generally associated to low CMD in Figure 7.15, since similar principal radiation regions relates to close centers of mass. Conversely, $\mathrm{CMD}(\omega)$ is higher when two principal radiation regions are "far" from each other, namely when the sound radiation of the violins is pointing towards different directions. This is particularly evident in the transitional and high (greater than $3\,\mathrm{kHz}$) frequency ranges. Accordingly to the results related to $\overline{\mathrm{NCC}}$ and $\overline{\mathrm{JSI}}$, high $\overline{\mathrm{CMD}}$ are observed for violin *Carlo IX* $v = 0$ and for combinations of $v = 6$ *Il Cremonese* and older instruments. In Figure. 7.13(c), a "small cluster" of *Stradivarius* instruments ($v = 5, 6, 7$) can be noted. However, Fig. 7.13(c) shows that according to the $\overline{\mathrm{CMD}}$ metric the most similar violins are the $v = 1$ *Hammerle* and the $v = 9$ *Stauffer*. This result differs from $\overline{\mathrm{JSI}}$ in Figure 7.13(b) and $\overline{\mathrm{NCC}}$ in Figure7.13(a). As a matter of fact, $\mathrm{CMD}_{n,j}(\omega)$ computes the distances between the centers of mass (7.13) of the principal directions, while $\mathrm{JSI}_{n,j}(\omega)$ directly compares the shapes of the regions of the principal directions (7.11) in the segmented directivity patterns (7.12). Hence, low values of $\mathrm{CMD}_{n,j}(\omega)$ can also be associated to different shapes in $\bar{\mathbf{d}}$ (7.12) but with similar centers of mass.

In conclusion, an overall metrics for the similarity of the directivity patterns is given by NCC which is also reported by string for an eased comparison with the principal radiation region probability maps. The results for the $\mathrm{JSI}(\omega)$ metric gave insights on the principal radiation regions differences in the frequency. As a matter of fact, it allowed us to easily identify frequency ranges for which the instruments present similar principal radiation regions. Finally, the results on CMD confirmed the trend observed in the other metrics and highlighted subtle differences among directivity patterns. This fact underlines the complexity of the directivity pattern analysis, and the need of considering multiple metrics at the same time in order to discriminate the subtle differences between the instruments.

CHAPTER 8

Deep-Leaning-based Nearfield Acoustic Holography

In this chapter we introduce a novel techniques for the analysis of acoustic sources. The proposed method relies on data-driven approaches to perform Nearfield Acoustic Holography (NAH). Nearfield acoustic holography (NAH) represents an interesting method for the analysis of acoustic sources with a low invasiveness. In fact, NAH enables the contactless analysis of acoustic sources, an interesting possibility in different scenarios. As instance, when analyzing particularly fragile instruments and conventional measurements cannot be performed or where deployment of the sensors is difficult to achieve. Contactless analysis is also preferred when lightweight objects are considered, since no additional mass needs to be added. Differently from contactless optical techniques, e.g., Laser Doppler Vibrometer (LDV), NAH can be employed with objects made of reflective materials. In practice, through NAH one can estimate the vibrational field of an acoustic source from acoustic measurements performed in its proximity. This allows the vibroacoustic analysis of a structure without the use of contact devices such as accelerometers. The acoustic pressure is typically captured by a microphone array deployed on a plane, known as holographic plane. The sensors are placed close to the vibrating surface in order to retrieve the evanescent wave components [286]. It is known [286] that the far field radiation of a source can be inferred from the knowledge of its surface vibrational field, hence NAH is an appealing contactless technique for the estimation of the directivity pattern of violins or any other source that cannot be conventionally measured.

With the aim of estimating the velocity field of the source from the pressure on the holographic plane, NAH relies on the inversion of the well-known Kirchhoff-Helmholtz (KH) integral [181, 286]. This operation is known to be a highly ill-conditioned problem, thus different strategies for solving NAH have been proposed in the literature. A direct approach to NAH considers the discretization of the KH integral, which leads to the Boundary Element Method (BEM) [28,60], adopted to solve the forward problem. Therefore, NAH is implemented through the inversion of BEM (IBEM) [240,270], using Tikhonov regularization. This technique is able to provide accurate results, but it is severely limited by the extreme computational cost. An alternative regularization strategy is represented by Compressed Sensing (CS) in [58,59], where the solution to NAH is approximated by a sparse set of plane wave components. However, the use of [58] is limited to star-shaped planar plates.

The Equivalent Source Method (ESM) [139,149] assumes that the measured acoustic pressure field radiated by the source can be equivalently expressed as the sound field generated by a set of point-like virtual sources located within or in proximity of the real source. The techniques based on ESM typically involve two phases. First, the solution to the inverse problem is involved in order to find weights of the equivalent sources, and in the second step a suitable propagator function is applied to the fictitious sources to infer the velocity on the target surface. The main problem of ESM is the computation of the optimal set of equivalent sources. In order to deal with this problem, ESM techniques based on CS [56,94] have been proposed with the aim of finding small and sparse subsets of equivalent sources.

In [56] a dictionary-based ESM (DESM) is proposed in order to consider a sparse domain for solving ESM limitations. The ESM solution space is restricted to a suitable compressed dictionary whose components are retrieved from several sets of equivalent sources. The dictionary is built from synthetic data varying the mechanical parameters of the object, while the object dimensions are fixed and known. The resulting set of equivalent sources weights is reduced using principal component analysis and then it constitutes the learned dictionary. Nevertheless, the location and the number of equivalent sources used to build the dictionary are still an open problem. This is especially true when the geometry of the objects under study are complex, e.g., whose surface exhibit curvatures.

The NAH methodology introduced in this chapter represents a novel data-driven approach to the problem. This solution is inspired by the effectiveness of learned features for NAH [56] and the well-known feature learning capabilities of Deep Neural Networks (DNN) [4,38,52,123]. In particular, we employ a well-known deep learning (DL) architecture called convolutional neural network (CNN) for performing NAH. Hence, we let the network learn during the training the optima features for solving NAH without imposing constraints on the source geometry, boundary conditions and material. First, in Section 8.1 we briefly review the characteristics of CNNs and the training procedure of deep learning architectures. Subsequently, in Section 8.2 we describe in details the proposed NAH technique. We employ the network in order to estimate the velocity field of vibrating plates from the pressure sampled on a rectangular grid over the surface. The CNN is trained using datasets of synthetic data generated using FEM simulations of rectangular and violin plates. The results shows that the proposed CNN is able to effectively estimate the vibrational field of sources with arbitrary geometry and orthotropic mechanical properties of the materials.

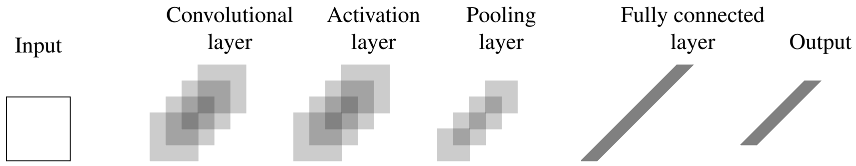


Figure 8.1: *Graphical representation of a "dummy" CNN architecture. A 2D input is processed through typical CNN layers i.e., convolutional, activation, pooling and fully connected layers.*

8.1 Convolutional Neural Network Overview

In this section we provide an overview of Deep Learning (DL) techniques. In particular, we focus on Convolutional Neural Network (CNN), a class of DL architectures widely adopted in different fields including acoustics [38]. We describe the main elements that constitute CNNs. Moreover, we explain the main principles and the training procedure of CNNs which is valid, in general, for any DL architecture.

Deep learning approaches based on CNNs provided outstanding results in several fields, including computer vision [123, 263], image processing [75, 147, 226], speech processing and music information retrieval [117, 183, 279], geophysics [137, 138], and not least acoustics [38, 63, 153]. CNNs are a variation of DL architecture usually employed for temporally or spatially correlated signals.

The computational model of CNNs, similarly to other DL architecture, is inspired by the neural system. In practice, they consist of a huge number of interconnected computational nodes, the neurons. A node performs a simple operation on its input, while every connection among the other neurons has a numeric weight whose value is tuned based on experience. The set of connections forms the neural network and typically nodes are organized in multiple stacked layers. The network parameters, i.e., the connection weights, are used to learn complex functions. Usually, CNNs are composed of the following operations: convolution, energy normalization, non-linear activation, and pooling. Through the minimization of a cost function at the output of the CNN, the parameters of the network are adapted in order to infer patterns in the input data. Consequently, a set of inherent features are extracted automatically from the data. The feature extraction process is driven by the data only, differently from traditional machine learning algorithms where features are "handcrafted", namely a-priori defined following signal models. The training of DL architectures such as CNNs, relies on backpropagation together with gradient descent optimization over a huge set of training data. In general, three main steps are needed for training a CNN:

- definition of the CNN architecture, i.e., the number of layers, the operations to perform at each layer, the size and the number of the convolutional filters etc;
- definition of a cost function distinctive for the required task that is minimized during the training of the network;
- dataset preparation for training, validating and testing the CNN.

In Figure 8.1, a *dummy* example of a small CNN with customary layers is reported. In the context of this thesis, we consider only 2D CNN as in Figure 8.1 which are characterized by two-dimensional input data and convolutional filters. In the following we provide a brief description of the architecture shown in Figure 8.1. Let us define the ith layer input as a feature map of size $H_i \times W_i \times P_i$ and its relative output as a feature map of size $H_{i+1} \times W_{i+1} \times P_{i+1}$. Note that for 1D layers, $H_i = W_i = 1$. The different types of layers in Figure 8.1 are

- **Convolutional Layer**: this layer performs 2D convolution, along the first and second dimensions, with stride S_h and S_w, between the input of the layer and P_{i+1} filters. Each filter has size $K_h \times K_w \times P_i$, while the output of the layer has dimensions $H_{i+1} = \left\lfloor \frac{H_i - K_h + 1}{S_h} \right\rfloor$, $W_{i+1} = \left\lfloor \frac{W_i - K_w + 1}{S_w} \right\rfloor$ and P_{i+1}.
- **Activation layer**: a non linear activation is applied on each element of the feature map. Typically in CNNs, the non linearity is a ReLU [177] that transform each input element x into the output y as $y = \max(0, x)$.
- **Pooling layer**: it performs an extraction of elements from the feature maps with stride S_h, S_w along the first and second axis, respectively. Each element is extracted from a $K_h \times K_w$ neighborhood of 2D sections of the input. It follows that the output feature map has size $H_{i+1} = \left\lceil \frac{H_i - K_h + 1}{S_h} \right\rceil$, $W_{i+1} = \left\lceil \frac{W_i - K_w + 1}{S_w} \right\rceil$ and $P_{i+1} = P_i$. In the literature different policies for the element extraction have been presented such as average pooling, sum pooling and the widely adopted max pooling that extracts the maximum value of the neighborhood.
- **Fully-connected layer**: this layer performs the dot multiplication between the input elements and the matrix of the weights that has P_{i+1} rows and as many columns as the number of input elements ($H_i \cdot W_i \cdot P_i$). The output is a vector of features with length P_{i+1}.

During the training phase, the parameters of the CNN, i.e., the weights of the convolutional and fully-connected layers are learned using backpropagation [31] algorithm, typically with gradient descent [46] optimization exploiting data belonging to a training dataset. The loss function, defined according to the application goals, is computed over the predicted and reference (expected) output, providing prediction error and gradients used for updating the parameters. The training data is divided in mini-batches [32] of samples. The goal of mini-batches is twofold, on one side, the data is split in subsets that can fit into the available memory, and on the other side we average the computation over a set of samples, proving a better approximation of the ideal gradient descent method which would otherwise require the estimation of the gradients over the whole dataset. Hence, the training is composed of iterations which consist of two steps for each mini-batch: a forward phase where the predictions are estimated and a backward phase in which the gradients, computed according to the loss function are averaged and "backpropagated" in order to update the weights of the network.

As mentioned, the dataset preparation involves the split of data into training set and validation set. The purpose of the training set is to perform the optimization of the network parameters, while the validation set is used to compute the loss function over unknown data, i.e., samples unseen during the training phase. Hence, the goal of the validation set is to tune the network in order to improve the performance of the CNN over new data. During the learning process of a CNN, the training is repeated for multiple "epochs", namely we train multiple times across the training dataset. The learning

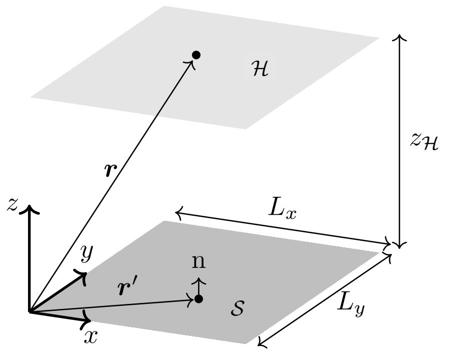


Figure 8.2: *Setup for NAH. The vibrating surface $\mathcal{S}$ is a finite plane, vectors $\mathbf{r}$, $\mathbf{r}'$ and n are defined according to a Cartesian reference system located at the bottom-left corner of $\mathcal{S}$. In Acoustic Holography, the radiated sound field is acquired at points belonging to a 2D surface $\mathcal{H}$, called hologram.*

ends when the loss related to the validation dataset reaches its minimum. Finally, once the training process is completed, the CNN is ready to perform on the test dataset of unseen new samples.

8.2 CNN-based Nearfield Acoustic Holography

In this section we propose a novel data driven method for NAH. In particular, we employ a DNN with the structure inspired by convolutional autoencoder architectures, in order to retrieve the vibrational data of a vibrating source from the pressure field captured by a microphone array in its proximity. The training data is generated through a FEM simulation campaign in which we compute the velocity of the surface and its relative pressure field of rectangular and violin plates. We varied plate dimensions, boundary conditions and mechanical properties in order to obtain a good dataset variability. The performance of the proposed solution was assessed comparing the predicted vibrational fields with the groundtruth coming from FEM simulations. Moreover, we investigate the robustness of the architecture against noisy input data, positioning errors in the sampling grid and missing data during the training.

8.2.1 Data Model and Problem Formulation

Let us consider the setup for NAH concerning a rectangular plate of dimensions L_x and L_y lying on the xy plane, as depicted in Figure 8.2. The vibrating surface $\mathcal{S}$ generates a 3D sound pressure field, which is measured on the hologram plane $\mathcal{H}$.

The exterior radiation of $\mathcal{S}$ in the air medium can be formulated in the frequency domain by means of the well-known Kirchhoff-Helmholtz (KH) integral (3.32) [181,286] as discussed in Section 3.5, hence, the sound pressure at a given point $\boldsymbol{r} = [x, y, z]^T$ on the hologram $\mathcal{H}$ is defined as [139]

$$P(\boldsymbol{r}, \omega) = \int_{\mathcal{S}} P(\boldsymbol{r}', \omega) \frac{\partial}{\partial \mathrm{n}} G(\boldsymbol{r}, \boldsymbol{r}', \omega) d\boldsymbol{r}' - j\omega\rho_0 \int_{\mathcal{S}} V_{\mathrm{n}}(\boldsymbol{r}', \omega) G(\boldsymbol{r}, \boldsymbol{r}', \omega) d\boldsymbol{r}', \quad (8.1)$$

where $P(\boldsymbol{r}', \omega)$ is the sound pressure evaluated at each point $\boldsymbol{r}' = [x', y', z']^T$ belonging to surface $\mathcal{S}$, $V_{\mathrm{n}}(\boldsymbol{r}', \omega)$ is the surface velocity along the outward normal direction n, ω is the angular frequency and $\rho_0 \approx 1.2\,\mathrm{kg\,m^{-3}}$ is the air mass density at $20\,°\mathrm{C}$. The term $G(\cdot)$ represents the Green's function (3.29), which models the acoustic wave propagation in the free-field as discussed in Section 3.4.

Let us consider the holographic plane $\mathcal{H}$ and the surface $\mathcal{S}$ be sampled on a regular grid of $N \times M$ points in locations $\boldsymbol{r}_{mn}$ and $\boldsymbol{r}'_{mn}$, with $m = 0, \dots, M-1$, $n = 0, \dots, N-1$, and where M, N are the number of points along the x and y axes, respectively. Hence, the discretized pressure field of (8.1) associated to the sampled surface and hologram can be expressed in matrix form with the introduction of the discrete estimator $\mathcal{F}$ as

$$\mathbf{P}_{\mathcal{H}}(\omega) \approx \mathcal{F}(\mathbf{P}_{\mathcal{S}}, \mathbf{V}, \omega), \tag{8.2}$$

where $\mathbf{P}_{\mathcal{H}} \in \mathbb{C}^{N \times M}$ and $\mathbf{P}_{\mathcal{S}} \in \mathbb{C}^{N \times M}$ are the pressure matrices at the hologram and the surface, respectively, and the term $\mathbf{V} \in \mathbb{C}^{N \times M}$ refers to the sampled velocity field.

In the context of NAH, we are interested in the estimation of the velocity matrix $\mathbf{V}(\omega)$ starting from the pressure measurements of the hologram $\mathbf{P}_{\mathcal{H}}(\omega)$ only. As a matter of fact, it is not possible to acquire the pressure on the surface $\mathbf{P}_{\mathcal{S}}$ available in (8.2) since the measurement is performed at the holographic plane $\mathcal{H}$. In practice this boils down to the inversion of (8.2), i.e.

$$\hat{\mathbf{V}}(\omega) \approx \mathcal{F}^{-1}(\mathbf{P}_{\mathcal{H}}(\omega)), \tag{8.3}$$

where $\hat{\mathbf{V}}(\omega)$ is the estimate of the normal velocity field. The estimation in (8.3) is known to be an ill-posed problem [287] and several regularization techniques were proposed in the literature such as the Tikhonov regularization [107], the L-curve analysis [122], the Equivalent Source Method [139, 149] and sparse regularization [58].

Here we propose a solution for the estimation of the velocity magnitude in (8.3) using CNNs (introduced in Section 8.1), in which $\mathcal{F}^{-1}$ has the structure inspired by a Convolutional Autoencoder [110] and the input of the CNN is the sound pressure magnitude. The main advantage of this approach is that it avoids explicit matrix inversions, since the inverse operator is learned by the network. The details of the proposed architecture are described in Section 8.2.2.

8.2.2 Neural Network Architecture Description

The adopted CNN model is inspired by the architecture of the renowned UNet [224]. This architecture consists of three main components: the contraction, the bottleneck, and the expansion sections. The contracting component $\mathcal{A}$, also known as encoder, is designed to extract a feature map from input X. This latent representation obtained in the encoding phase is located at the bottleneck. In the expansion section $\mathcal{Z}$, also known as decoder, the network exploits the embedding at the bottleneck to provide the desired output Y, namely $Y = \mathcal{Z}(\mathcal{A}(X))$. Autoencoder-like structures have been successfully employed in denoising tasks [92, 108, 155]. In fact, thanks to their compressive nature, these architectures prioritize the learning of useful input data description in order to approximate the output, showing robustness against noisy data. In the context of NAH this represents an appealing feature, since the inversion of (8.2) is prone to deviations due to noise in the data.

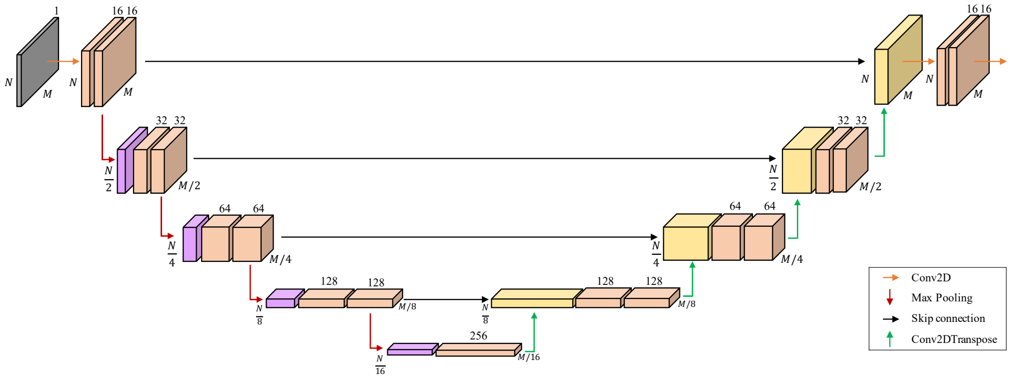


Figure 8.3: *The proposed UNet architecture. Orange blocks represent layers made by convolutions and ReLU activation functions, purple blocks are the max pooling operations, while yellow blocks indicate the upsampling stages. The input and the output of the network are depicted in grey and blue, respectively.*

Therefore, the relation between the estimated velocity magnitude and the magnitude of the corresponding sound pressure field defined in (8.3) can be expressed in terms of the autoencoder components as

$$\hat{\mathbf{\Phi}}(\omega) \approx \mathcal{Z}\left(\mathcal{A}(\mathbf{\Psi}(\omega); \mathbf{w})\,; \mathbf{w}\right), \tag{8.4}$$

where $\mathbf{\Psi}(\omega) = |\mathbf{P}_{\mathcal{H}}(\omega)|$, $\hat{\mathbf{\Phi}}(\omega) = |\hat{\mathbf{V}}(\omega)|$ and the parameters $\mathbf{w}$ are learned by optimizing the network predictions in the Mean Square Error (MSE) sense through

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{Y} - \mathcal{Z}(\mathcal{A}(\mathbf{X}; \mathbf{w}); \mathbf{w})\|_2^2, \tag{8.5}$$

with $\mathbf{Y}$ and $\mathbf{X}$ representing collections of the normal velocity fields and the sound pressure fields on the holographic plane, respectively (see Section 8.2.1).

In Figure 8.3 the detailed structure of the proposed CNN is provided along with the description of the layers dimensions.

Input/Output data: The dataset $\mathbf{X}$ is the input of the network representing the sound pressure magnitude evaluated on the holographic plane $\mathbf{\Psi}(\omega)$. Each matrix has $\mathrm{N} \times \mathrm{M}$ values, with $\mathrm{N} = 16$ and $\mathrm{M} = 64$. For the ease of the reader, we will refer to it as pressure image. The dataset $\mathbf{Y}$ is the desired output shown to the network during the training phase. It collects the normal velocity matrices $\mathbf{\Phi}(\omega) = |\mathbf{V}(\omega)|$ measured over the structure surface. Each matrix, which will be referred to as velocity image, has the same dimensions $\mathrm{N} \times \mathrm{M}$ of the input.

Encoder: The proposed encoder $\mathcal{A}$ consists of a series of four dowsampling blocks. Each block includes two consecutive layers of 2D convolutions with filter size 3×3, each followed by a Rectified Linear Unit function (ReLU) [177]. After each block, a 2×2 max pooling operation is applied to achieve the compression. The downsampling starts with 16 extracted features and we double the number of feature channels at each step. Therefore, we reach 256 features at the innermost layer, thus representing the information with a structure of $1 \times 4 \times 256$.

Decoder: Every step of the decoder $\mathcal{Z}$ operates an "up-convolution" by means of a Conv2DTranspose layer [78] with stride 2×2. The skip connections [123] between each downsampling block and its corresponding upsampling layer enables the reuse of the encoded features in the decoding process. Moreover, two 3×3 convolutions with ReLU activations follow each upsampling step. In this way, the decoder has also a large number of feature channels, which allow the network to propagate context information to higher resolution layers. As a consequence, the decoder is symmetric to the encoder, and yields a u-shaped architecture.

8.2.3 Dataset Generation

In this section we describe the dataset generation used for the evaluation of the CNN.

The dataset used for training, test and validation has been generated through a FEM simulation campaign using the *COMSOL Multiphysics*® software [64]. We evalute the NAH method on two different datasets. First, focused on isotropic rectangular plates with dimensions comparable to the body of small bowed-string instruments, while as a second test we apply the NAH technique on violin plates. We perform the simulations at the eigenfrequencies $\bar{\omega}$ of each plate, with $\bar{\omega} \in [0, \omega_{\mathrm{MAX}}]$ where ω_{MAX} is defined such that $\frac{\omega_{\mathrm{MAX}}}{2\pi} = 2000\,\mathrm{Hz}$. The synthesized data is sampled on an uniform grid of $\mathrm{N} \times \mathrm{M}$ points with $\mathrm{N} = 16$ and $\mathrm{M} = 64$. We computed the spatial sampling steps $\bar{x}$, $\bar{y}$ to honor the Shannon-Nyquist conditions, namely

$$\bar{x} \leq \frac{L_x}{M-1}, \quad \bar{y} \leq \frac{L_y}{N-1} \quad \text{and} \quad \min(\bar{x}, \bar{y}) < \frac{\pi c}{\omega_{\mathrm{MAX}}}, \tag{8.6}$$

with the hologram plane $\mathcal{H}$ placed at $z_{\mathcal{H}} = \min(\bar{x}, \bar{y})$ from the vibrating surface.

Finally, data is organized in the input and output datasets as

$$\mathbf{X}, \, \mathbf{Y} \in \mathbb{R}^{N \times M \times D}, \tag{8.7}$$

where X is the input (pressure) dataset, Y contains the expected output (velocity) images and D is the number of pressure and velocity images generated. Notice that the datasets consist of the absolute value of the collected data (see Section 8.2.2). From the condition on spatial sampling (8.6) adopted for creating the dataset (8.7), it follows that the NN is not aware on the actual grid size i.e., varying distances between grid points. This simplifies the network structure, which does not consider the different distances, at the cost of providing the correctly sampled data.

Rectangular Plates

Through *COMSOL Multiphysics*® we have simulated the vibroacousic behaviour of different rectangular plates considering three Boundary Conditions (BCs): simply supported, clamped and free edges.

We varied the dimensions of the plate as $L_x \in [0.23, \, 0.36]\mathrm{m}$, $L_y \in [0.15, \, 0.22]\mathrm{m}$ and $L_z \in [0.002, \, 0.007]\mathrm{m}$ and we varied the x and y dimensions with a step of $0.01\,\mathrm{m}$ while the sampling in the z dimension (thickness) had a step of $0.001\,\mathrm{m}$. Therefore, we obtained a set of 672 different rectangular plates. As far as the material properties are concerned we simulated isotropic aluminum plates adopting the standard material definition of *COMSOL Multiphysics*® [64]. As the eigenfrequencies vary as a function

Property	Variable	Value	Unit
Density	ρ	400	$\mathrm{kg\,m^{-3}}$
Young's modulus	$[E_1, E_2, E_3]$	$[10.8, 0.4644, 0.8424]$	GPa
Shear modulus	$[G_1, G_2, G_3]$	$[0.6588, 0.6912, 0.0324]$	GPa
Poisson's ratio	$[\nu_1, \nu_2, \nu_3]$	$[0.467, 0.372, 0.435]$	1

Table 8.1: *Sitka spruce material properties from [225].*

of the BCs and the plate dimensions, a different number of pressure and velocity fields occur in the three different cases of BCs applied. Therefore, in order to have a balanced dataset we replicated data associated to less frequent modes, obtaining a total number of $D_1 = 342400$ velocity and pressure fields of $\mathrm{N} \times \mathrm{M} = 16 \times 64 = 1024$ points, collected in the input and output datasets (8.7).

Violin Plates

In order to validate the NAH with object of arbitrary shape, a dataset[1] of violin plates has been employed. The synthetic meshes of violin top plates with constant thickness and arching were generated by the parametric model introduced in [109, 227]. The parameters defining the shapes are varied according to Gaussian distributions centered around the parameters of a reference violin, as described by the authors in [109, 227]. Again the response of violin plates to a time harmonic load through *COMSOL Multiphysics®* has been simulated. The dataset is constituted of 1111 meshes of violin plates and each object presents a different outline, while the arching, the largest dimension $L_x = 0.356\,\mathrm{m}$ and the uniform thickness of $L_z = 27\,\mathrm{mm}$ are shared among all the plates. As regards the material parameters, an orthotropic model of sitka spruces [225] was adopted, whose parameters are reported in Table 8.1. Similarly to the rectangular plate case, we sampled the pressure and velocity fields with $N \times M = 16 \times 64 = 1024$ points. It is worth noting that, in order to cover all the violin plate, the rectangular sampling grid defined according to (8.6) provides locations that might fall outside the plate outline. Therefore, for such points the corresponding velocity value cannot be computed. In order to deal with points of the velocity images outside the violin plate, we adopt a binary mask to force the reconstruction only in meaningful locations. The binary mask assumes zero value when a point of the velocity image is outside the violin plate and one otherwise. Through a point-wise matrix multiplication the binary mask is applied to the output of the network in order to remove the contribution of points located outside the violin plate.

Differently from the rectangular plate case, here we limited the simulation to free BCs. In fact, free BCs are customarily considered by liuthiers when building and tuning violin plates. Using this setup and the considered frequency range we obtained around 40 eigenfrequencies for each violin plate, resulting in a final dataset (8.7) composed of $D_2 = 48207$ elements.

[1] The violin plate dataset was kindly made available by Dr. Davide Salvi and Sebastian Gonzalez, Ph.D. from Politecnico di Milano.

8.2.4 Evaluation Metrics

The performance of the proposed CNN is assessed by comparing the estimated vibrational field with the simulated ground truth. In order to evaluate the prediction accuracy (8.4), we measured the Normalized Mean Square Error (NMSE) between the reconstructed plate velocity field $\hat{\mathbf{\Phi}}(\omega)$ and the synthesized ground truth $\mathbf{\Phi}(\omega)$, computed as

$$\mathrm{NMSE}\left(\hat{\mathbf{\Phi}}(\omega), \mathbf{\Phi}(\omega)\right) = 10\log_{10}\left(\frac{\|\hat{\mathbf{\Phi}}(\omega) - \mathbf{\Phi}(\omega)\|_2^2}{\|\mathbf{\Phi}(\omega)\|_2^2}\right). \tag{8.8}$$

Another metric that we used is the Normalized Cross Correlation (NCC) between the predicted velocity image and the ground truth, namely

$$\mathrm{NCC}\left(\hat{\mathbf{\Phi}}(\omega), \mathbf{\Phi}(\omega)\right) = \frac{\hat{\mathbf{\Phi}}^T(\omega) \cdot \mathbf{\Phi}(\omega)}{\|\hat{\mathbf{\Phi}}(\omega)\|\|\mathbf{\Phi}(\omega)\|}. \tag{8.9}$$

Note that the NCC corresponds to 1 when the output predictions perfectly match the ground truth velocity pattern.

8.2.5 Results

The performance of the proposed NAH technique is not limited to the validation over the two datasets, but additionally, we performed a study on the robustness against noisy input data and sampling positioning (location of microphones) errors. Moreover, for the violin plates we observed the performance of the NAH against missing data in the training set.

Rectangular Plates

The CNN architecture is implemented[2] in Python using Keras [61]. The dataset was splitted in $60\,\%-30\,\%-10\,\%$ for the training, test and validation sets, respectively. The model was trained for 7 epochs using Adam optimizer [133] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate $\alpha = 0.001$ and applying early stopping in order to prevent overfitting. In general, the model was able to retrieve a good prediction of the velocity field and to recognize the different BCs applied. For instance, in Figure 8.4, Figure 8.5, and Figure 8.6 reconstruction examples are reported for the free, clamped and simply supported BCs, respectively. Here, we label the modes accordingly to the progressive number given by the frequencies sorted in ascending order. Since the pressure ($\mathbf{\Psi}(\bar{\omega})$) and velocity ($\mathbf{\Phi}(\bar{\omega})$) images are quite similar with simply supported and clamped BCs, we can notice that the network performs some kind of deblurring and rescaling operations. Nevertheless, even if the pressure and velocity images in the free BCs are very different and they are characterized by more complex patterns, the network predicts correctly the desired output. We interpret this as if the CNN is able to retrieve hidden information on propagation phenomena and it learns to infer the relation defined by the back-propagation operator (8.3).

NMSE and NCC evaluated on the test set in ascending and descending order are shown in Figure 8.7(a) and Figure 8.7(b), respectively. The average NMSE value, $\mathrm{NMSE}_{\mathrm{AVG}} = -30.87\,\mathrm{dB}$ is reported in Figure 8.7(a), note that the $19.59\,\%$ of the test

[2] `https://github.com/polimi-ispl/nah-cnn`

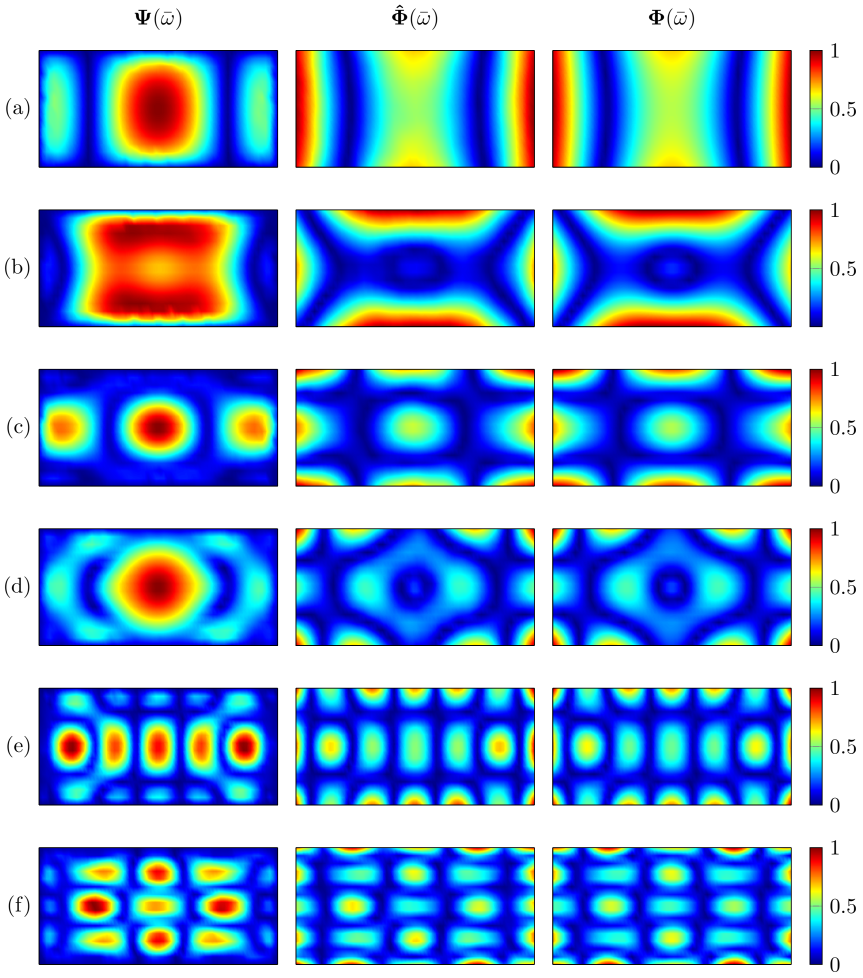


Figure 8.4: *Examples of pressure field Ψ, velocity field estimates $\hat{\Phi}$ and references Φ for rectangular plates with free BCs.*

Label	$[L_x, L_y, L_z]$ [m]	Mode	Frequency [Hz]
(a)	$[0.35, 0.16, 0.002]$	1	85.98
(b)	$[0.32, 0.16, 0.003]$	4	428.06
(c)	$[0.36, 0.15, 0.004]$	9	1311.71
(d)	$[0.34, 0.19, 0.004]$	14	1735.51
(e)	$[0.35, 0.20, 0.002]$	27	1688.11
(f)	$[0.36, 0.21, 0.002]$	31	1999.08

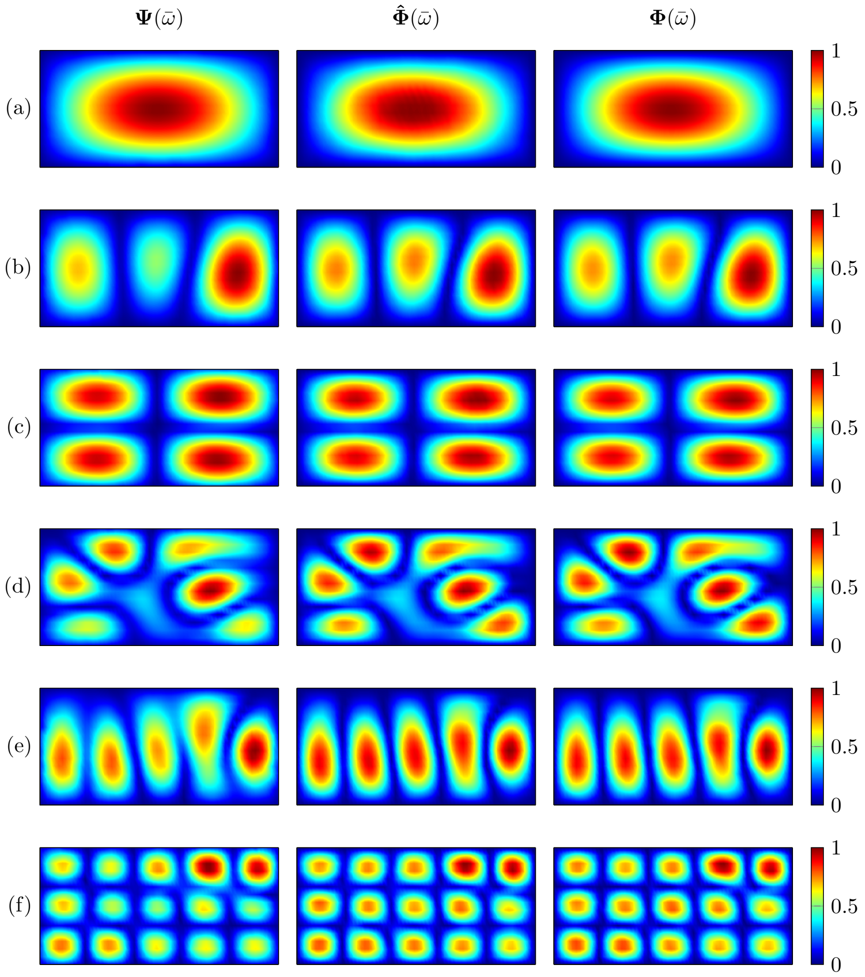


Figure 8.5: *Examples of pressure field $\mathbf{\Psi}$, velocity field estimates $\hat{\mathbf{\Phi}}$ and references $\mathbf{\Phi}$ for rectangular plates with simply supported BCs.*

Label	$[L_x, L_y, L_z]$ [m]	Mode	Frequency [Hz]
(a)	$[0.23, 0.18, 0.002]$	1	243.07
(b)	$[0.31, 0.15, 0.002]$	3	671.05
(c)	$[0.30, 0.19, 0.005]$	5	1873.51
(d)	$[0.33, 0.32, 0.003]$	10	1741.95
(e)	$[0.20, 0.20, 0.002]$	12	1905.58
(f)	$[0.35, 0.21, 0.002]$	19	1956.3

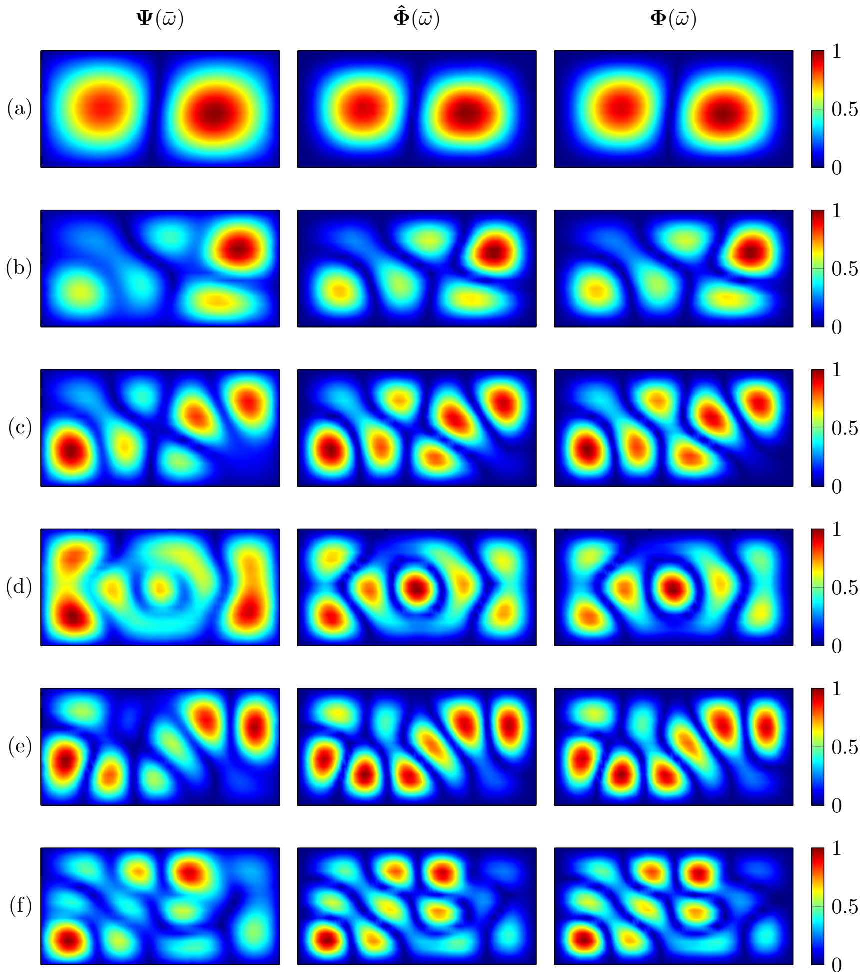
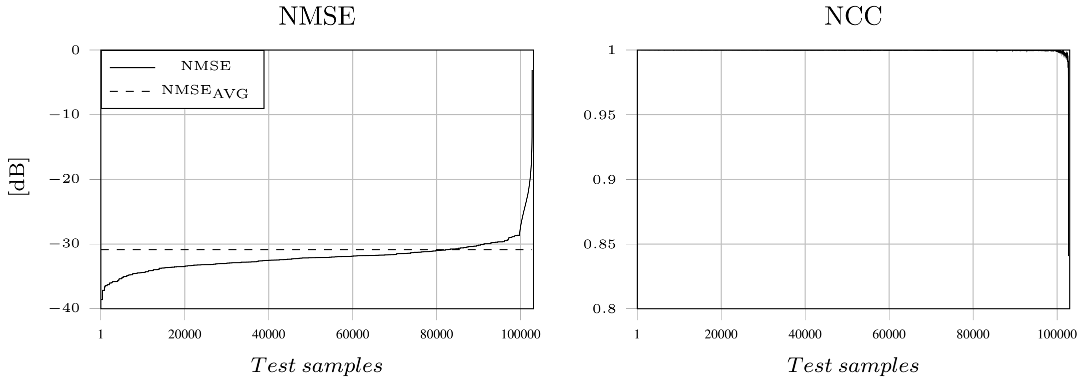


Figure 8.6: *Examples of pressure field* Ψ*, velocity field estimates* $\hat{\Phi}$ *and references* Φ *for rectangular plates with clamped BCs.*

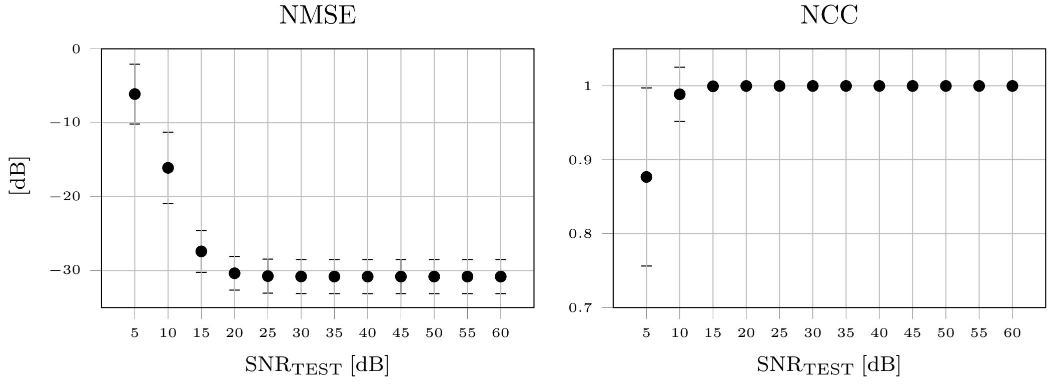
Label	$[L_x, L_y, L_z]$ [m]	Mode	Frequency [Hz]
(a)	$[0.26, 0.20, 0.002]$	2	85.98
(b)	$[0.25, 0.18, 0.002]$	6	428.06
(c)	$[0.32, 0.19, 0.002]$	8	1311.71
(d)	$[0.31, 0.22, 0.002]$	11	1735.51
(e)	$[0.36, 0.20, 0.002]$	13	1688.11
(f)	$[0.36, 0.21, 0.002]$	15	1999.08



(a)

(b)

Figure 8.7: *The ordered metrics evaluated over the test set: (a)* NMSE *along with its average value, (b)* NCC *between the output prediction and the ground truth*



(a)

(b)

Figure 8.8: *Metrics evaluation with respect to the mean and the standard deviation confidence error as a function of* $\mathrm{SNR}_{\mathrm{TEST}}$. *The* NMSE *(a) and the* NCC *(b) between the output prediction and the ground truth, respectively.*

set reports a NMSE value above $\mathrm{NMSE}_{\mathrm{AVG}}$ and only the $1.96\,\%$ has a NMSE value greater than $-25\,\mathrm{dB}$.

Inspecting the results related to the worst predictions, we can observe that they are mainly related to a scaling bias between the reconstruction and the ground truth and do not influence the estimated patterns. Moreover, since NCC highlights the pattern similarity between prediction and ground truth, by looking the graph in Figure 8.7(b) we can confirm that the large majority of the test set images matches the relative velocity field reconstructions. Although the minimum NCC value is 0.84, only the $0.015\,\%$ of the predicted velocity fields has a NCC value less than 0.98.

Noisy Input Data In order to simulate measurements in actual scenarios, we trained the network with an input dataset $\mathbf{X}$ corrupted by additive white noise with a fixed value of $\mathrm{SNR}_{\mathrm{TRAIN}}$ equal to $40\,\mathrm{dB}$. The learned model has then been tested on 12 additional versions of noisy dataset by varying the $\mathrm{SNR}_{\mathrm{TEST}}$ from $5\,\mathrm{dB}$ to $60\,\mathrm{dB}$ with a step of

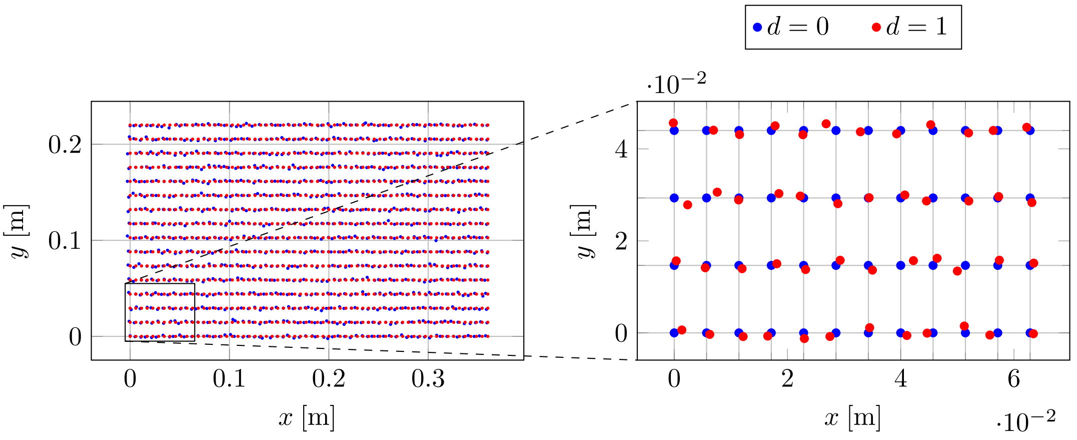


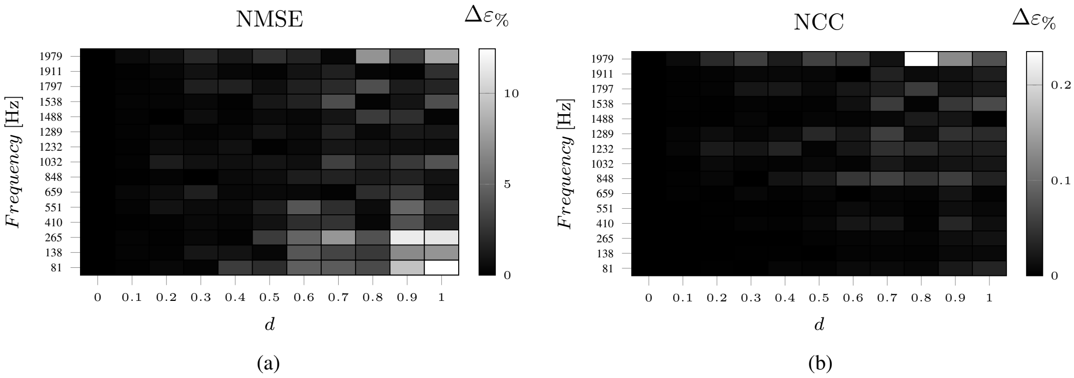
Figure 8.9: *Grids of sampling points (microphones) for a rectangular plate with dimensions* $[0.36, 0.22, 0.002]\mathrm{m}$. *Complete version (left) and a detail (right): in blue the regular grid* ($d = 0$) *and in red the grid affected by the maximum error* ($d = 1$).

$5\,\mathrm{dB}$. In Figure 8.8 the mean value of the metrics evaluated at different $\mathrm{SNR}_{\mathrm{TEST}}$ are shown along with the corresponding standard deviation confidence error. As expected, the model trained at $\mathrm{SNR}_{\mathrm{TRAIN}} = 40\,\mathrm{dB}$ was able to estimate correctly the normal velocity for $\mathrm{SNR}_{\mathrm{TEST}} > 40\,\mathrm{dB}$, with an average NMSE value around $-30\,\mathrm{dB}$. It is noteworthy that the model retrieves similar results up to a value of $\mathrm{SNR}_{\mathrm{TEST}} = 20\,\mathrm{dB}$. This suggests that the architecture is able to learn a set of features in its encoding stages such that it describes the correct pattern while it discards the additive noise contribution, even with data more corrupted than the train set. Comparing the NMSE and the NCC, we can notice that the performance of the network decreases when $\mathrm{SNR}_{\mathrm{TEST}} \leq 15\,\mathrm{dB}$.

Comparing the results with the performances of state of the art techniques such as [56], we can observe that the proposed technique appears as more flexible. A direct comparison with [56] cannot be readily performed, since the authors in [56] only vary the mechanical parameters of the plate. On contrary, the proposed technique is able to work with rectangular plates with different dimensions and boundary conditions obtaining an high NCC for a broad range of plates (see Figure 8.7(b)). Nonetheless, a large dataset is required for training the CNN, while although being data driven [56] adopts limited data. Noteworthy, the proposed CNN shows a remarkable robustness with respect to noisy input, with an average $\mathrm{NCC} > 0.8$ for an $\mathrm{SNR} = 5\,\mathrm{dB}$. Similar studies in [56] instead showed a significant decrease in the $\mathrm{NCC} \approx 0.5$ with similar SNR values.

Positioning Error In actual experimental measurements, microphones are placed on a regular grid to record the acoustic data, but errors on their positioning often occur. Hence, we performed the analysis with an error associated to microphone positioning. We generated new acoustic pressure data with the software *COMSOL Multiphysics®* and we exported them using different noisy grids of $\mathrm{N} \times \mathrm{M} = 1024$ points. In order to sample the hologram plane on inaccurate grids, a discrete set of points can be defined as

$$H = \{[x, y, z_{\mathcal{H}}]^T \in \mathcal{H} \mid x \sim \mathcal{N}(m\bar{x},\, d\sigma_x^2) \wedge y \sim \mathcal{N}(m\bar{y},\, d\sigma_y^2)\}, \tag{8.10}$$
$$m = 0, ..., M-1, \quad n = 0, ..., N-1,$$



(a)

(b)

Figure 8.10: *Metrics evaluation of the relative error in percentage with respect to the regular grid as function of sampling position errors at different frequencies. The percent error in (a) refers to the* NCC*, in (b) to* NMSE*.*

where the x and y locations of each microphone are realizations of two Gaussian distributions with the regular grid locations as mean and two variables controlled by the parameter d ranging in $[0,\ 1]$ as variance. The values of σ_x^2 and σ_y^2 are fixed to

$$\sigma_x^2 = \left(\frac{\bar{x}}{6}\right)^2, \quad \sigma_y^2 = \left(\frac{\bar{y}}{6}\right)^2. \tag{8.11}$$

This particular choice has been done in order to prevent overlapping or inversion of the grid points. It is worth noticing that the discrete set of points corresponds to the regular grid when $d = 0$, while the maximum random positioning errors occur when $d = 1$. We selected a specific plate with dimensions $[0.36, 0.22, 0.002]\mathrm{m}$ and we retrieved the pressure fields associated to the three BCs and different noisy grids by varying the parameter d from 0 to 1 with a step of 0.1. Figure 8.9 shows an example of regular and noisy grids used to retrieve the hologram pressure data. We tested the model with the input pressure images corrupted by different positioning errors on the microphone grid and we evaluated the performance. In Figure 8.10, the relative error of NCC and NMSE ($\Delta\varepsilon_{\%}$) is depicted as a function of frequency and the noise parameter d. The value of the metrics evaluated in the case of a regular grid is taken as reference, hence $\Delta\varepsilon_{\%}$ provides the relative difference of the metrics when data with positioning error is instead considered. Frequencies in Figure 8.10 represent some examples of natural frequencies for the three BCs. As expected, the estimate error increases with the positioning error. In particular, by inspecting the NCC plot we can infer that the pattern predictions are robust at low frequencies with a minimum dependence on d. On the other hand, by increasing the frequency, the dependence on the positioning errors increases up to a maximum NCC relative error of $0.23\,\%$ with respect to the regular grid. In the NMSE plot we can identify three main regions with different error behaviors. In the first one, with $d \in [0,\ 0.2]$, the model reconstructions are not affected by the positioning error. In the second region, with $d \in [0.3,\ 0.6]$, a uniform error increment is present in all the frequency ranges. In the last region, with $d > 0.6$, the NMSE relative error mainly grows at low frequencies with a maximum deviation of $12\,\%$ with respect to the regular grid. It must be noticed that this behaviour is due only to the scaling bias of the reconstruction which does not affect the estimated pattern as shown by the NCC. Nevertheless, although the network presents a good fault tolerance, it is important to place microphones carefully in order to avoid undesirable scaling errors.

Violin Plates

As a second analysis on the proposed NAH method, we evaluate the performance of the CNN using the dataset of violin plates. Here, we split the dataset in $60\,\% - 30\,\% - 10\,\%$ for training, validation and test, respectively. The training was performed for 100 epochs with early stopping and learning rate reduction on plateau.

In Figure 8.11 we report examples of velocity predictions along with the input pressure and reference velocities of the first 6 modes of the plates. These modes are particularly interesting since violin makers usually analyze such modes in order to drive the building of instruments. Overall the network is able to predict the velocity of violin plates also at higher frequencies and some examples are reported in Figure 8.12.

In Figure 8.13(a) the NMSE is reported for the whole test dataset, while Figure 8.13(b) shows the relative NCC. As regards the average performance of NMSE, we report a $\mathrm{NMSE}_{\mathrm{AVG}} = -16.46\,\mathrm{dB}$ and the $30.68\,\%$ of the samples achieved a NMSE greater than $\mathrm{NMSE}_{\mathrm{AVG}}$. We can notice that the $\mathrm{NMSE}_{\mathrm{AVG}}$ reported for the violin plates is higher with respect to the one obtained with the dataset of rectangular plates. We interpret this increase as a result of the complexity of the velocity field shapes and reduced number of elements in the dataset. Nevertheless, by inspecting the definition of the NMSE in (8.8) we can infer that a $\mathrm{NMSE}_{\mathrm{AVG}} = -16.46\,\mathrm{dB}$ corresponds to a relative error of only $2.26\,\%$ between the prediction and expected velocity field. Moreover, we noted that the mismatch between the network output and the groundtruth is generally associated to scaling between the data. The observation is confirmed by the inspection of NCC in Figure 8.13(b), where it is possible to note a good agreement between the predicted and expected patterns. In particular, only the $10.59\,\%$ of the test samples report $\mathrm{NCC} < 0.98$ and it never goes below 0.88.

Noisy Input Data Similarly to the rectangular plates analysis, we test the robustness of the NAH technique against noisy input data with the violin plate dataset. The network was trained with a training dataset $\mathbf{X}$ corrupted with additive white noise with variance set in order to obtain $\mathrm{SNR}_{\mathrm{TRAIN}} = 40\,\mathrm{dB}$. During the test phase, we employed 12 noisy versions of the test dataset obtained varying the $\mathrm{SNR}_{\mathrm{TEST}} \in [5, 60]\mathrm{dB}$ with a $5\,\mathrm{dB}$ step. The average values obtained with the different test datasets are reported in Figure 8.14 along with their standard deviation. We can notice that the performance decreases with $\mathrm{SNR}_{\mathrm{TEST}} \leq 15\,\mathrm{dB}$, similarly to what is achieved with rectangular plates.

Positioning Error Again we evaluate the performance of proposed NAH technique in the case of positioning errors of the pressure field sampling. Here, we test the CNN using three violin plates modifying the sampling locations of the pressure field according to (8.10) varying the parameter $d \in [0, 1]$. For each violin plate, we chose 15 modes and 11 realizations of noisy sampling grids were simulated. In Figure 8.15 an example of sampling grid with positioning error is reported. We consider the values of the metrics obtained with the correct grid positioning as the benchmark of the analysis since no train or adaptation has been performed with “corrupted” grids. Therefore, we evaluated

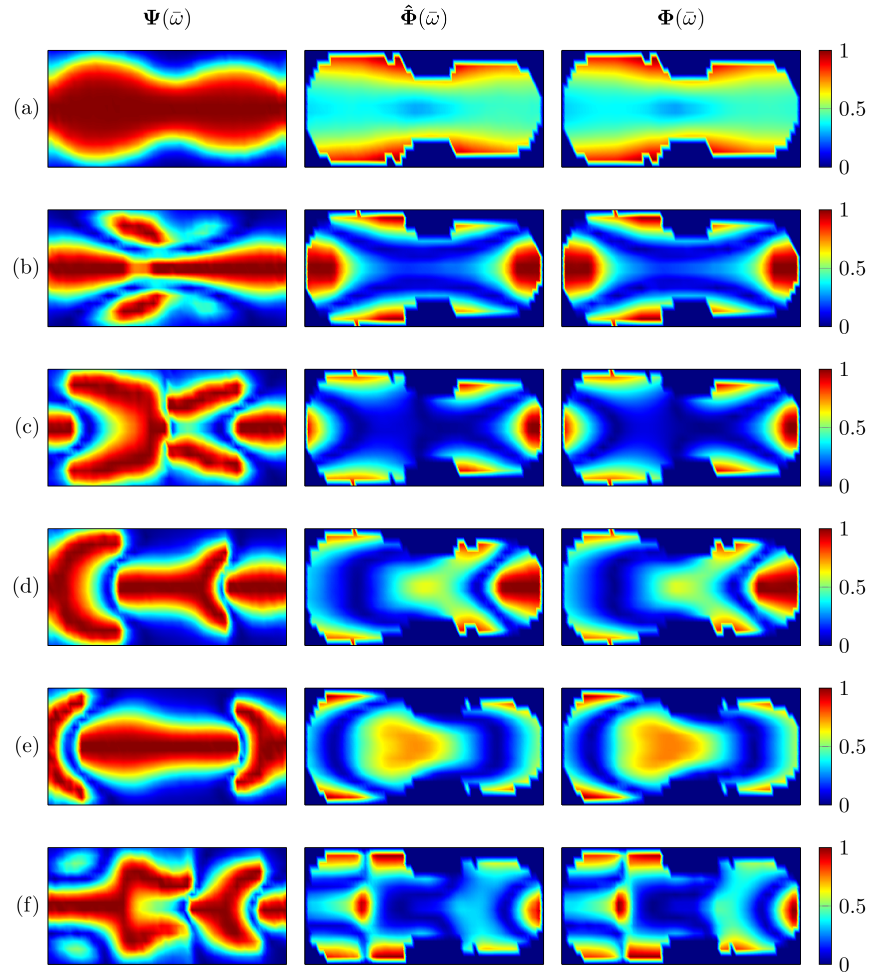


Figure 8.11: *Examples of pressure field Ψ, velocity field estimates $\hat{\Phi}$ and references Φ for violin plates.*

Label	$[L_x, L_y]$ [m]	Mode	Frequency [Hz]
(a)	$[0.356, 0.215]$	1	63.3
(b)	$[0.356, 0.221]$	2	98.95
(c)	$[0.356, 0.205]$	3	191.16
(d)	$[0.356, 0.207]$	4	205.22
(e)	$[0.356, 0.216]$	5	265.29
(f)	$[0.356, 0.206]$	6	316.59

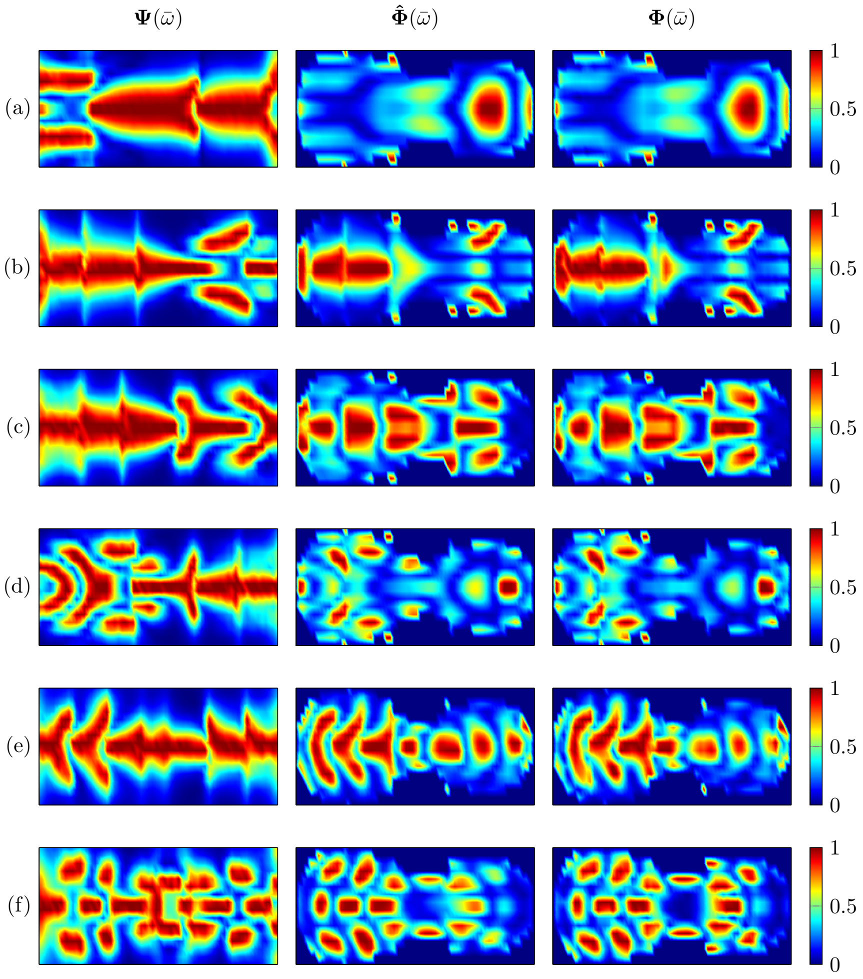


Figure 8.12: *Examples of pressure field Ψ, velocity field estimates $\hat{\Phi}$ and references Φ for violin plates.*

Label	$[L_x, L_y]$ [m]	Mode	Frequency [Hz]
(a)	$[0.356, 0.193]$	15	711.94
(b)	$[0.356, 0.222]$	25	1077.9
(c)	$[0.356, 0.205]$	30	1285.1
(d)	$[0.356, 0.204]$	35	1596.6
(e)	$[0.356, 0.211]$	40	1784
(f)	$[0.356, 0.207]$	44	1964.7

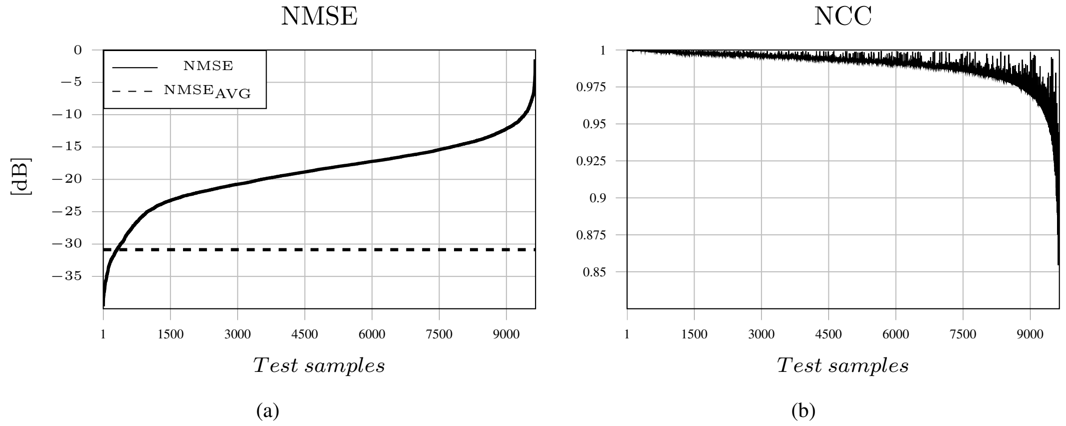


Figure 8.13: *The ordered metrics evaluated over the test set of violin plates: (a)* NMSE *along with its average value, (b)* NCC *between the output prediction and the ground truth*

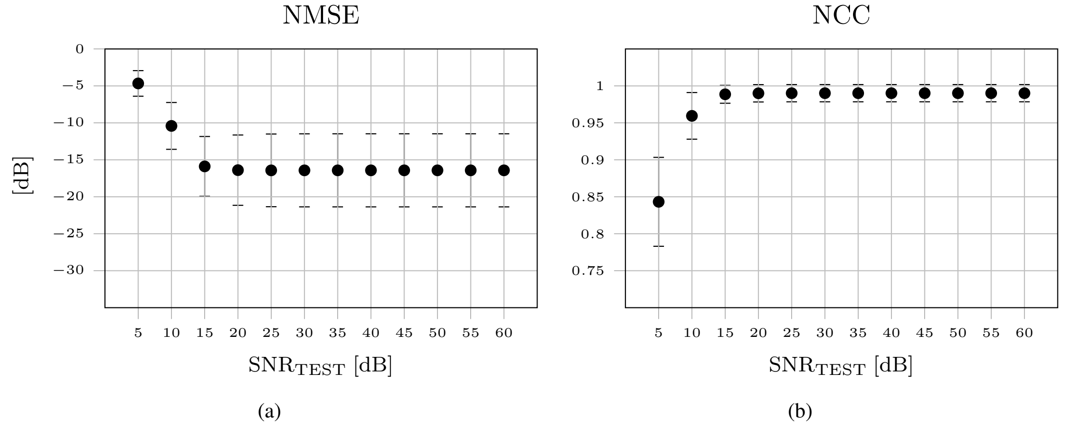


Figure 8.14: *Metrics evaluation with respect to the mean and the standard deviation confidence error as a function of* $\mathrm{SNR}_{\mathrm{TEST}}$. *The* NMSE *(a) and the* NCC *(b) between the output prediction and the ground truth, respectively.*

the relative error of NMSE and NCC with respect to the value of the metrics obtained with the correct grid positioning and it is reported in Figure 8.16.

Inspecting the results in Figure 8.16, we can infer that the estimation is almost unaffected by errors in the sampling locations up to $d = 0.6$. An increase in the location error caused a deviation of both NMSE and NCC up to $18.7\,\%$ and $3.4\,\%$, respectively.

Missing Data In addition to the previous analysis, we investigate the performance of the NAH against unknown pressure fields. In practice, we removed different sets of elements from the training data related to specific frequencies or entire violin plates. Such elements are therefore used in the test phase only, in order to evaluate the generalization of the network to “unseen” data. First we removed from the training set all elements at the frequencies of the first five modes. We first focus on these modes, since they are analyzed by violin makers during the construction of the instruments. In fact, it is known that such plate modes are related to the signature modes of the instru-

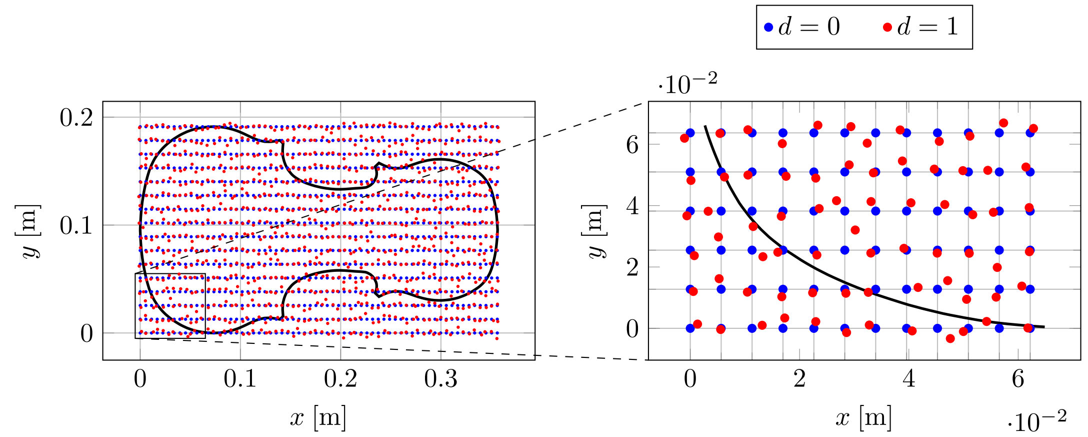


Figure 8.15: *Grids of sampling points (microphones) for a violin plate with dimensions* $[0.356, 0.191]$m. *Regular grid points are in blue (*$d = 0$*) while in red the grid affected by maximum error (*$d = 1$*).*

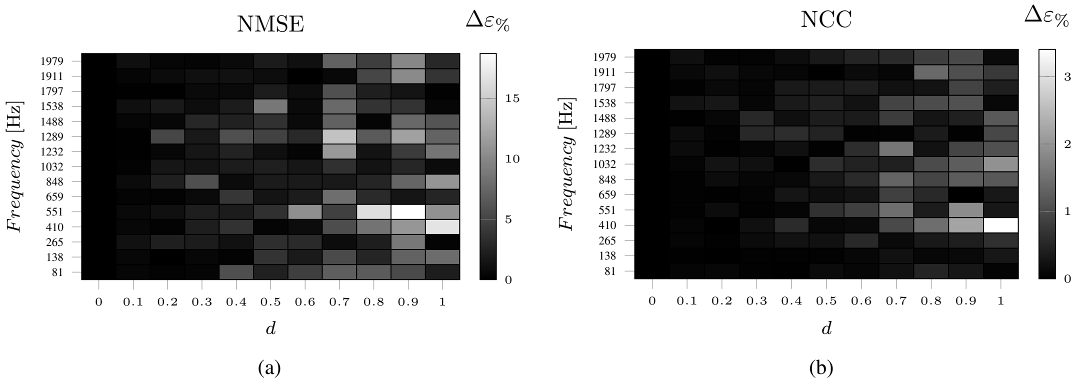


Figure 8.16: *Metrics evaluation of the relative error in percentage with respect to the regular grid as function of sampling location errors at different frequencies. The percent error in (a) refers to the* NCC*, in (b) to* NMSE*.*

ments [113, 289]. In Figure 8.17 an example of reconstructions provided for the first, second and fifth mode are shown. We can notice that the worst prediction is associated to the first mode, the second and fifth mode velocity reconstructions effectively estimate the nodal line in the velocity fields, even if clear differences in the patterns are present. For this subset we obtained a $\text{NMSE}_{\text{AVG}} = -7.58\,\text{dB}$. Further tests are related to the removal of the data in the frequency range $[850, 1477]$Hz and $[1443, 2000]$Hz. In Figure 8.18 and Figure 8.19 we report examples for the two frequency ranges, respectively. At higher frequencies, the estimates improved with $\text{NMSE}_{\text{AVG}} = -12.26\,\text{dB}$ and $\text{NMSE}_{\text{AVG}} = -9.92\,\text{dB}$ for data in $[850, 1477]$Hz and $[1443, 2000]$Hz, respectively. The reconstruction is mainly affected by scaling errors, as the NCC never drops below 0.9 for all the considered cases. Such results are particularly interesting, since we can infer that the proposed NAH technique can be performed on data at higher frequencies with respect to the training one. As last investigation, we removed from the training set all the elements relative to 10 violin plates. This allows us to analyze the performance of NAH with respect to unknown violin plate geometries. In Figure 8.20 some examples of the reconstruction are reported. We can note that the velocity fields

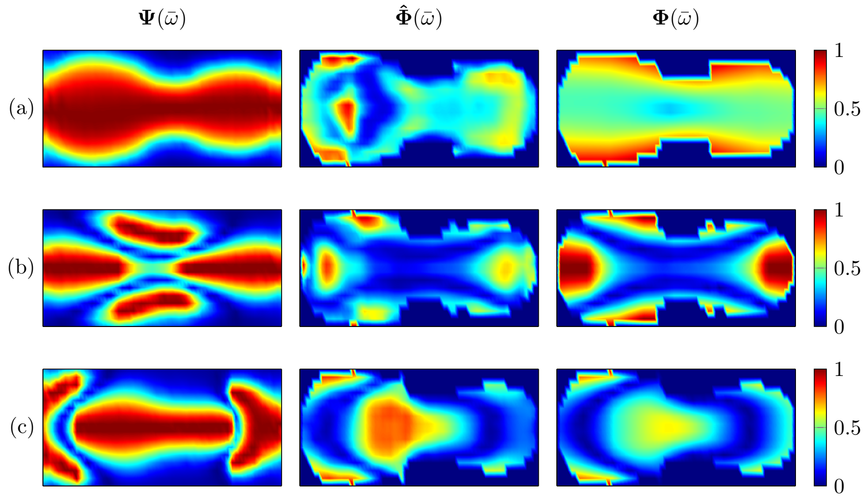


Figure 8.17: *Examples of pressure field Ψ, velocity field estimates $\hat{\Phi}$ and references Φ for modes of violin plates "unseen" during the training phase.*

Label	$[L_x, L_y]$ [m]	Mode	Frequency [Hz]
(a)	$[0.356, 0.203]$	1	66.7
(b)	$[0.356, 0.227]$	2	97.03
(c)	$[0.356, 0.202]$	5	270.76

are correctly estimated, and this is confirmed by the metrics that report values close to reconstructions obtained including such plates in the training datasets. In particular, the NMSE is on average $\mathrm{NMSE}_{\mathrm{AVG}} = -15.66\,\mathrm{dB}$, while NCC never drops below 0.95. Therefore, we envision the feasibility of the proposed NAH technique also to unknown plates. This possibility is particularly interesting in practical scenarios when we can train the network with simulated data and successively employ the NAH method with measurements of actual violin plates with geometries different from those used during the training.

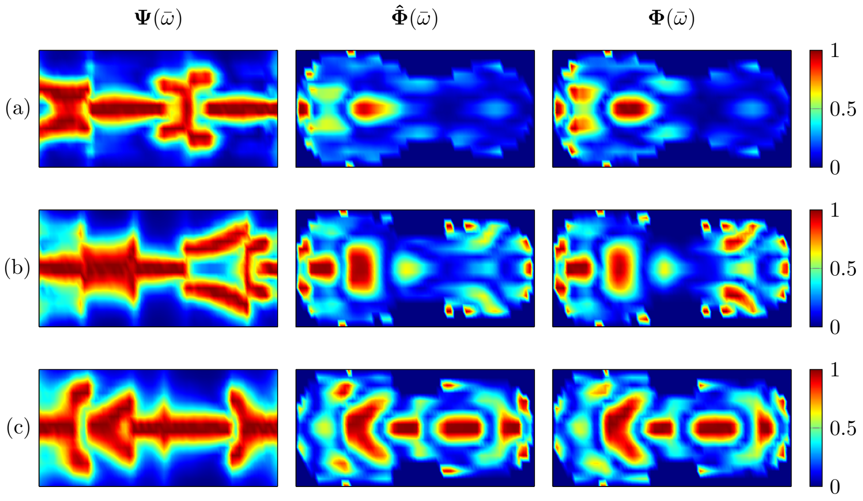


Figure 8.18: *Examples of pressure field $\mathbf{\Psi}$, velocity field estimates $\hat{\mathbf{\Phi}}$ and references $\mathbf{\Phi}$ for modes of violin plates "unseen" during the training phase.*

Label	$[L_x, L_y]$ [m]	Mode	Frequency [Hz]
(a)	$[0.356, 0.234]$	22	913.77
(b)	$[0.356, 0.202]$	24	1161
(c)	$[0.356, 0.219]$	26	1111.4

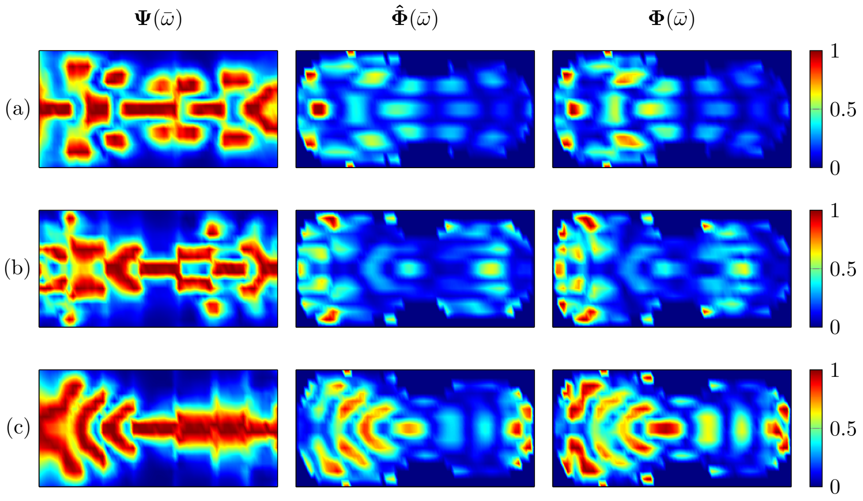


Figure 8.19: *Examples of pressure field $\mathbf{\Psi}$, velocity field estimates $\hat{\mathbf{\Phi}}$ and references $\mathbf{\Phi}$ for modes of violin plates "unseen" during the training phase.*

Label	$[L_x, L_y]$ [m]	Mode	Frequency [Hz]
(a)	$[0.356, 0.233]$	41	1560.3
(b)	$[0.356, 0.197]$	42	1849.1
(c)	$[0.356, 0.202]$	43	1915.8

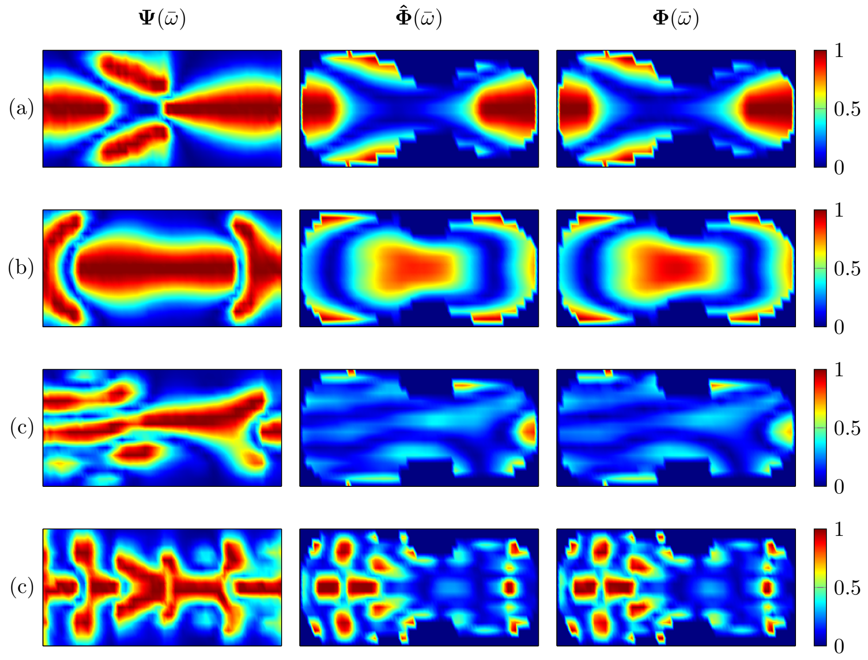


Figure 8.20: *Examples of pressure field* Ψ*, velocity field estimates* $\hat{\Phi}$ *and references* Φ *for violin plates "unseen" during the training phase.*

Label	$[L_x, L_y]$ [m]	Mode	Frequency [Hz]
(a)	$[0.356, 0.203]$	2	116.04
(b)	$[0.356, 0.203]$	5	266.74
(c)	$[0.356, 0.210]$	10	489.56
(d)	$[0.356, 0.200]$	35	1543.7

CHAPTER 9

A Methodological Approach towards the Numerical Prediction of the Directivity Pattern

Numerical simulation represents the least invasive approach to the analysis of the directivity pattern of violins. As a matter of fact, one can imagine to simulate the vibroacoustic dynamic behavior of the instrument by means of Finite Element Method (FEM) simulations and successively, to estimate the VS parameters from the synthetic data.

Clearly, in order to provide effective simulations of the vibroacustic behavior of violins, and consequently a prediction of the directivity pattern, precise models in terms of the geometry of the instrument and material properties, are required. Unfortunately, such accurate models are difficult to obtain since violins are handcrafted instruments composed of different parts (the body, the neck, the strings etc.). Even if we limit ourselves to the violin body, which is the main responsible of the acoustic radiation, we discover that the individual parts making the body present variable thickness and a complex geometry. Actually, each part of the violin body will contribute, even if with variable influences to the final radiated sound and consequently to the directivity pattern.

In order to perform simulation of the violin body components, the geometry of the instrument can be acquired by means of 3D scans. Typically on built instruments, the different body part cannot be taken apart, hence only their outer surface can be acquired. It follows that, the development of techniques that are able to reconstruct the 3D profile of the different violin components is needed.

Among the different parts of the violin, here we focus on the violin plates. In fact its known that violin plates are the main contributors to the acoustic response of the instrument as underlined in Section 7.1.

In Section 9.1, we introduce a practical technique for the reconstruction of the plate 3D model from outer surface scans through the modeling of its inner surface according to the nonuniform thickness profile. The 3D outer shape of the plate is acquired by means of a 3D laser scanner and then smoothed in order to remove artefacts and details that are unnecessary for the acoustics simulation. We assume that the thickness is known at some reference points from which we retrieve the thickness of the whole plate. These reference values could for instance be measured by a thickness gauge on the instrument or given by the literature. In the following, we apply a methodology for the interpolation of the thickness on a regular grid that covers the whole plate area. The combined knowledge of the outer and inner surfaces makes it possible to reconstruct the three-dimensional geometry of the plate. We validate this 3D reconstruction technique by comparing the vibrometric behaviour of the 3D model with data measured on the reference plate, and with simulations on a model with uniform thickness.

Additionally to the geometry of the instrument, an effective characterization of the material mechanical parameters is required in order to provide effective simulations. Moreover, in the field of liuthery and musical acoustics in general, the accurate estimation of the material mechanical properties has a great impact, because relevant mechanical parameters are used to drive the design and building process of musical instruments [282]. In this thesis we approach the estimation of mechanical parameters of wood from the analysis of the wave speed in the material. In fact, from the sound wave velocity we can estimate parameters such as the Young's modulus.

The estimation of the wave velocity in a medium can be tackled through *matched field processing*. This problem has been studied in different domains such as seismology [16], underwater navigation [126, 245, 246] and microphone array processing [17, 18, 214]. Among violin makers, a well-known method for the estimation of wave velocity in wood is the *tap tone* [128]. This technique is widely adopted by luthiers because of its repeatability and non-invasive characteristic. It consists in the estimation of the resonance frequency of the wood block, from which the longitudinal velocity is derived. Unfortunately, it requires a great manual skill in order to correctly identify the resonance frequency by *tapping* the tone wood block.

Alternatively, the longitudinal wave velocity can be easily estimated by measuring the time of flight (TOF) of an impulsive wave between the extremes of the block under analysis [136, 158]. Due to the high propagation speeds, the adoption of expensive analog or digital instrumentation with sampling rate in the ultrasound bandwidth is required. Furthermore, since state-of-the-art techniques measure the TOF of the direct wave only, the estimated velocity turns out to be sensitive to measurement errors.

In Section 9.2, as a first approach towards the mechanical parameter identification for numerical simulations, we consider the estimation of the sound wave velocity in tone wood blocks. The developed technique exploits the impulse response of the wood block, acquired by means of accelerometers. This results in a simple procedure that is highly repeatable and non invasive. In addition, it does not require expensive instrumentation or specific skills. We analyze in a rake receiver [210] fashion the impulse responses, extending the analysis of the TOF to a larger portion of the impulse response, beyond the direct wave. This allows us to work at a sampling frequency in the audible bandwidth, adopting low cost general purpose digital hardware.

Although the two techniques introduced in this chapter are not directly linked, they represent two basic steps towards the implementation of violin directivity pattern simulations. In fact, both the geometry and the material parameters are fundamental for the

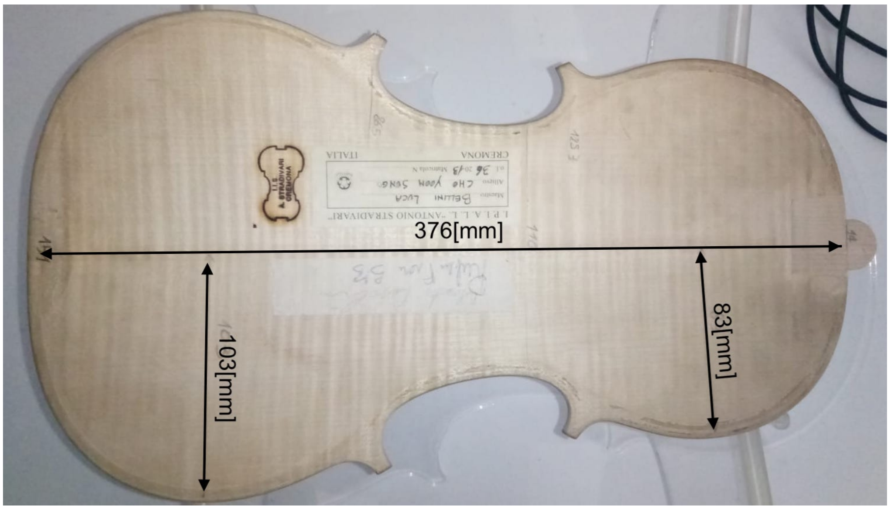


Figure 9.1: *The violin plate acquired by the 3D laser scanner.*

final acoustic simulation. We envision the development of a framework for the numerical simulation of violins, in which the proposed techniques can increase the accuracy of the final simulated model.

9.1 Predictive Simulation of Violin Plates

9.1.1 Reconstruction Methodology

The reconstruction of the 3D of the plate from the outer surface proceeds through modeling the inner surface of the plate as the composition of two separate regions. The region of the inner surface close to the edge is flat, and therefore it determines a plane on which the whole plate can lie. Conversely, the central part of the inner surface exhibits a nonuniform thickness. Starting from the reference points, the thickness is interpolated on a regular grid that covers the whole central area. Joining the outer and the reconstructed inner surface the 3D geometry of the plate is reconstructed.

Violin plate scanning and mesh generation

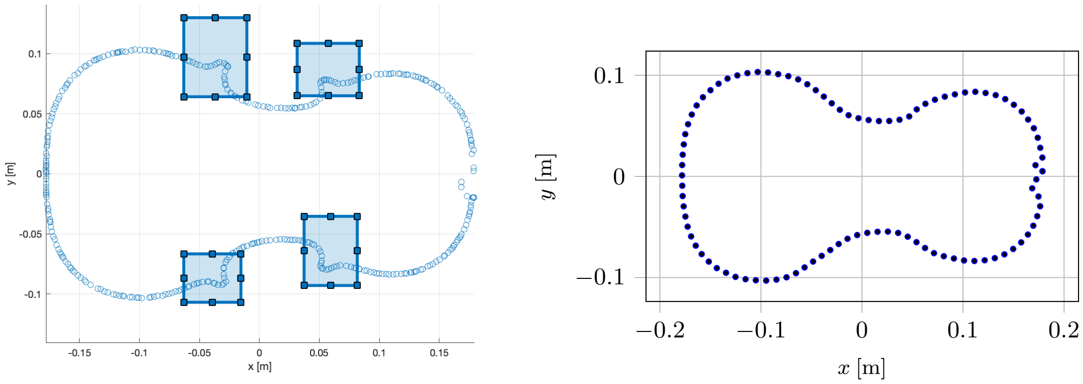
We briefly introduce the measurement process of the 3D geometry of the plate from laser scanning. A violin plate with definite geometrical properties (outline, arch and thickness) is employed (see Figure9.1). The violin plate has been measured before the final varnishing process and assembling on the instrument. The length of the plate is $376\,\mathrm{mm}$ and the upper and lower radii are approximately $83\,\mathrm{mm}$ and $103\,\mathrm{mm}$, respectively (see Figure 9.1). As regards the thickness reference samples, a thickness gauge (a tool typically available in a luthier workshop) has been adopted for the measurements.

The 3D geometry of the violin plate is acquired using a Romer ABSOLUTE ARM laser scanner with $0.01\,\mathrm{mm}$ resolution and Polyworks$^{\mathrm{TM}}$ acquisition software by InnovMetric. A scan of the entire plate has been performed and considered as the groundtruth in the evaluation of our reconstruction technique.

Algorithm 1 Violin Plate Surface Reconstruction

```
Input   : Outer Surface Plate Point Cloud
          Thickness Samples

Steps:
    1. External Boundary Estimation
    2. Inner Surface Plane Estimation
    3. Curved Region Boundary Estimation
    4. Thickness Interpolation
    5. Inner Surface Mesh Build
    6. Outer and Inner Surfaces Union
Output: Reconstructed Violin Plate Mesh
```



(a) (b)

Figure 9.2: *(a) The MATLAB graphical interface for the corner selection. (b) The fitted curve of the outline.*

Reconstruction algorithm

Our goal is to estimate the inner surface from the outer surface scan and then merge the two faces in one single mesh obtaining the reconstruction of the entire plate geometry.

In order to correctly reconstruct the structure of the violin plate, it is essential to model the inner surface carefully. According to its curvature profile, the inner face can be subdivided in two main regions: the flat one, which follows the plate border and a curved central area. The latter determines the nonuniform thickness profile of the plate. Our method determines the thickness of the central region interpolating the inner surface point cloud from a set of measurements given by the user. The reference values required by the interpolation process can be, when possible, directly measured on the violin plate, or for example given by the luthier.

The reconstruction procedure is summarized in Algorithm 1.

Step 1. External Boundary Estimation Given the point cloud of the outer surface mesh, the first step requires the identification of the contour of the outer surface. This is accomplished by means of the well-known alpha shapes algorithm [35]. The contour is used in Step 3 to determine the curved region boundary.

Step 2. Inner Surface Plane Estimation In the global three dimensional Cartesian coordinate system the outer surface ideally lies on the x, y plane looking toward the z^+ direction. Unfortunately, sampled plates may present deformations along the three axis. As a result, the plate lies on a different plane with respect to the x, y plane. In this step, we determine this plane employing a polynomial fitting of second degree in x and third degree in y. The point cloud is then projected onto the lying plane, and rigidly translated to obtain the inner surface plane with the required thickness along the edges. This step is essential to obtain the flat region along the edge of the inner surface.

Step 3. Curved Region Boundary Estimation In order to characterize the curved region of the inner surface, we manually remove the corners from the the contour (Step 1) and we fit a smooth curve on the remaining points using a spline kernel [69] (See Figure 9.2). The identified curve is used as a boundary between the flat and curved regions and its location is determined by the flat area extension. The flat region width differs slightly from plate to plate, according to the choices of the violin maker. In general, the width profile follows the plate design and it is wider in correspondence of the corners and on the upper and lower edges (see Figure 9.3(c)) Given the average, upper and lower flat region widths we can correctly identify the boundary of the curved surface area.

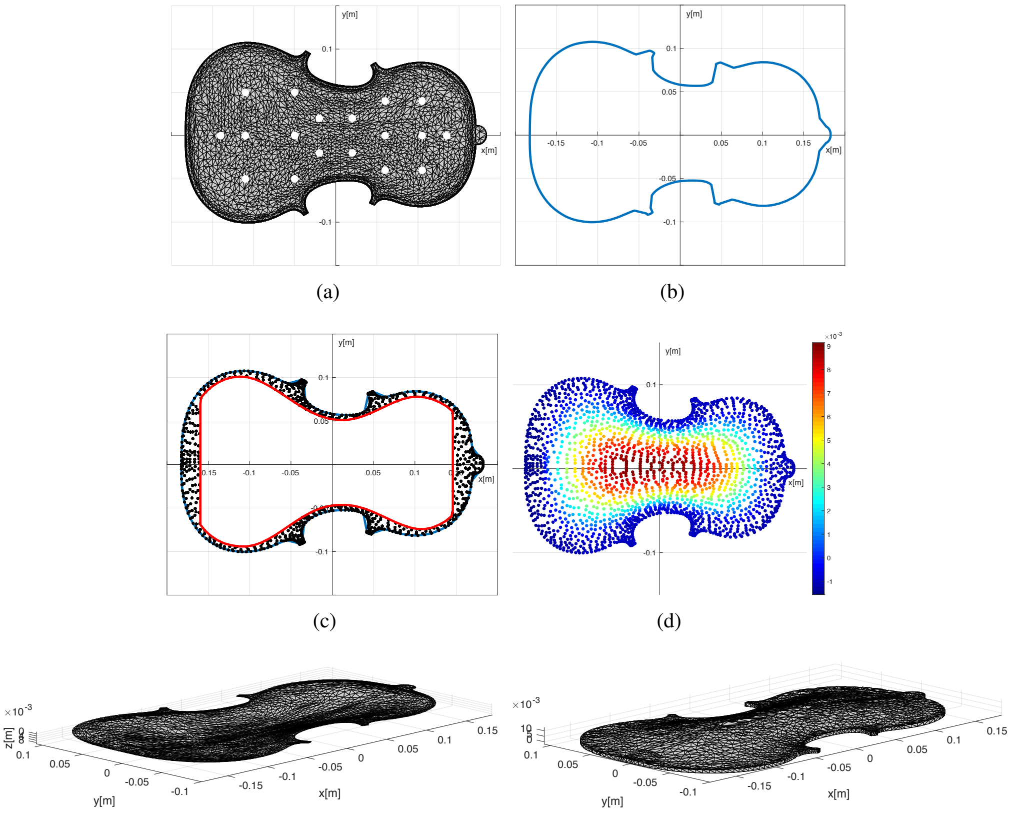
Step 4. Thickness Interpolation Once the curved central region is identified in (Step 3), we proceed to the interpolation of the thickness from the sampled points on a regular and dense grid. The measured values are used to drive an interpolation algorithm [15] on the surface points. This operation computes the thickness so that the transition from the flat region to the curved area is smooth. The interpolated thickness map is used for computing the actual z coordinate of the inner surface point cloud from the outer surface z coordinates and the normal directions, as shown in Figure 9.3(d).

Step 5. Inner Surface Mesh Build The output of Step 4 consists in the final inner surface point cloud determined by the thickness interpolation of the curved region. The data is used as a set of vertices and a triangulation algorithm [13, 14] determines the mesh of the inner surface, as shown in Figure 9.3(e). Note that after the triangulation of the vertices has been performed, a further cleaning of the output mesh may be required, based on the meshing process accuracy.

Step 6. Outer and Inner Surfaces Union Finally, the mesh of the entire plate is obtained merging the outer and inner surfaces. This can be easily done with standard 3D computer graphics software, such as *Blender* [43]. The selection of the mesh edge loops and the connection of the outer and inner edges can be accomplished with built in functions. The unified output mesh of our procedure, consisting in the merged outer and inner surfaces (see Figure 9.3(f)), represents a reconstruction of the entire violin plate geometry that can be employed for accurate simulations and analysis, as shown in the next section.

9.1.2 Validation and Results

In this section we discuss the results of a mechanical simulation of the reconstructed violin plate. Our reconstruction procedure is compared to the results of a *vanilla* recon-



(a) (b)

(c) (d)

(e) (f)

Figure 9.3: *(a) Thickness measurements sampling points (white) on the violin plate. (b) External boundary estimated in Step 1. (c) Curved region boundary (red) and flat region points (black) computed in Step 3. (d) Inner surface point cloud given by Step 4. (e) Inner surface mesh build in Step 5. (f) Reconstructed plate mesh generated in Step 6*

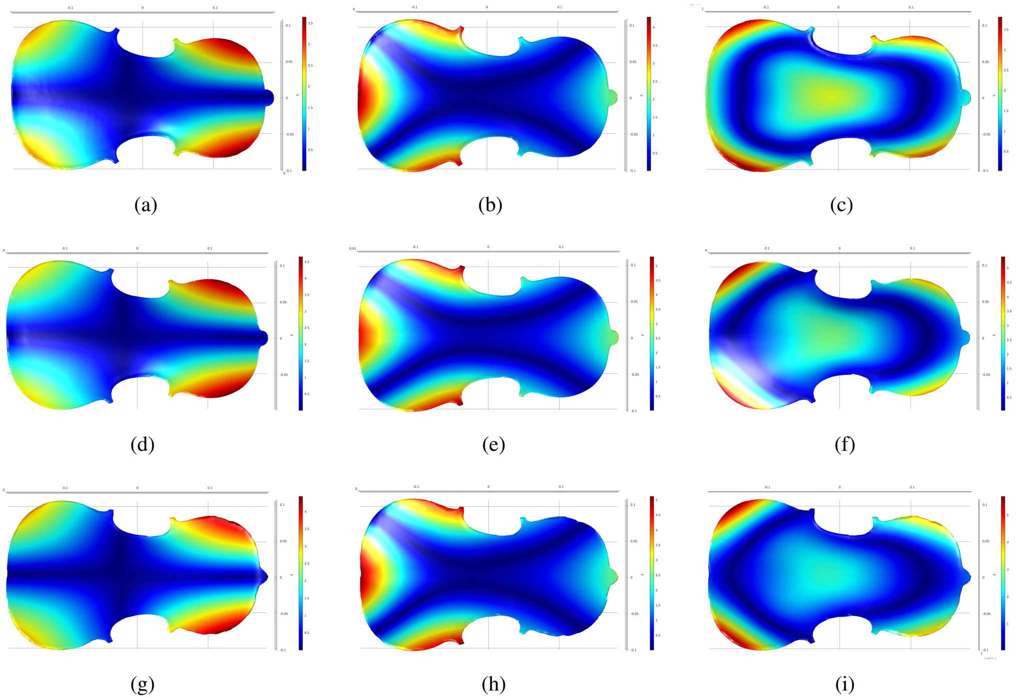
struction method, consisting in a plate with uniform thickness.

Simulation setup

We reconstruct the entire plate surface of Figure 9.1 using its outer surface. We follow the methodology described in 9.1.1, where the thickness of the curved region has been sampled in 20 points (see Figure 9.3(a)). The thickness has been measured on the actual violin plate (see Figure 9.1) by means of a thickness gauge. Hence, we aim at obtaining the same simulated mechanical behavior between the reconstructed 3D model and the actual scanned violin plate. The average thickness of the flat region is set to 3[mm] and as for the width we measured $30.4\,\mathrm{mm}$ and $16.9\,\mathrm{mm}$ for the upper and

Young's modulus	Shear modulus	Poisson's ratio
$E_x = 12.6[GPa]$	$G_{xy}/E_x = 0.111$	$\mu_{xy} = 0.424$
$E_y/E_x = 0.132$	$G_{yz}/E_x = 0.021$	$\mu_{yz} = 0.774$
$E_z/E_x = 0.065$	$G_{xz}/E_x = 0.063$	$\mu_{xz} = 0.476$

Table 9.1: *Values of the orthotropic properties of the simulated material.*



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 9.4: *(a) Scanned plate mode 1. (b) Scanned plate mode 2. (c) Scanned plate mode 5. (d) Uniform thickness reconstruction mode 1. (e) Uniform thickness reconstruction mode 2. (f) Uniform thickness reconstruction mode 5. (g) Proposed reconstruction mode 1. (h) Proposed reconstruction mode 2. (i) Proposed reconstruction mode 5.*

lower edges, respectively. The average width for the remaining part of the contour is set to $6\,\mathrm{mm}$. The plate with uniform thickness has been obtained by rigidly translating the outer surface by $3\,\mathrm{mm}$ along the z direction. We analyze the mechanical behavior of the three plates (reference, vanilla and proposed reconstructions) with free boundary conditions through an FEM simulation performed with *COMSOL Multiphysics*® [64] software. For each plate, we import the polygon mesh data and a tetrahedron mesh is automatically generated by the software. In order to accurately simulate the vibration behavior of the plate, the material properties of the object must be carefully set. In this case, the plate is made of spruce wood, whose elastic properties along the three axes are taken from [116] and shown in Table 9.1.

Results

As a first study, we compare the eigenfrequencies of three modes that are considered important by violin makers, which are namely, mode 1, mode 2 (the so called "cross-mode") and mode 5 (also known as "ring mode"). In Figure 9.4 the mode shapes of the eigenfrequencies in analysis are depicted for the scanned surface (Figure 9.4(a),(b),(c)), the uniform thickness reconstruction (Figure 9.4(d),(e),(f)) and for the surface reconstructed through the proposed methodology (Figure 9.4(g),(h),(i)).

In order to evaluate the effectiveness of the proposed reconstruction technique, we consider the absolute error in Hz computed as the absolute value of the difference between the eigenfrequencies obtained through the mechanical simulation of the scanned

	Scanned	Uniform	Proposed	E_u [Hz]	E_p [Hz]
Mode 1 [Hz]	96.5	90	93.2	6.5 (6.7 %)	3.3 (3.4 %)
Mode 2 [Hz]	158.8	144	152.2	14.8 (9.7 %)	6.7 (4.2 %)
Mode 5 [Hz]	337	371.8	337.3	34.8 (10.3 %)	0.3 (0.09 %)

Table 9.2: *Eigenfrequency values and the relative error E_u of the uniform thickness plate and the proposed methodology E_p.*

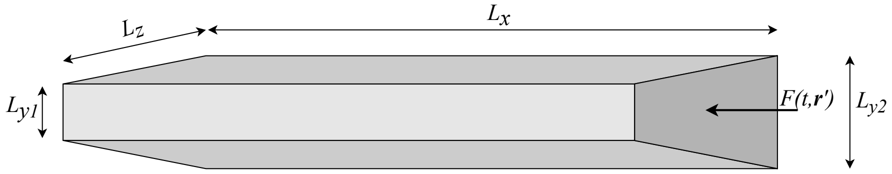


Figure 9.5: *The shape of a wood block used for building plates of violins or other string musical instruments.*

plate and the ones given by the reconstructed plates.

The eigenfrequencies of the three simulations are reported in Table 9.2, along with the relative and percentage error with respect to the reference plate. From the values reported in columns E_u and E_p of Table 9.2, it is possible to notice that the plate reconstructed with the proposed process is characterized by eigenfrequencies that are closer to the reference values with respect to a plate reconstructed with uniform thickness. In particular, the error obtained with our reconstruction technique E_p is reduced with respect to the error given by a plate of uniform thickness E_u for all the eigenfrequencies considered.

This simulation proves the importance of an accurate thickness reconstruction in the context of eigenfrequency analysis. More precisely, our methodology is able to better approximate the eigenfrequencies of an actual violin plate, improving significantly the simulation effectiveness.

9.2 Sound Wave Speed Estimation in Tone Wood

9.2.1 Problem Formulation and Data Model

Let us consider a tone wood block with length L_x. The cross-section is trapezioidal, with dimensions L_{y1}, L_{y2}, L_z, so that the cross section area is $A = \frac{(L_{y1}+L_{y2})\times L_z}{2}$, as depicted in Figure 9.5. We measure the signal of N accelerometers, placed at $\boldsymbol{r}_n = [x_n, y_n, z_n]^T$, with $n = 1, \dots, N$. The signal acquired by the nth sensor can be modelled, in absence of noise, as

$$s(t, \boldsymbol{r}_n) = h(t, \boldsymbol{r}', \boldsymbol{r}_n) * \eta(t, \boldsymbol{r}'), \tag{9.1}$$

where t is the time index, $\eta(t, \boldsymbol{r}')$ is the source signal given by an axial load $F(t, \boldsymbol{r}')$ placed at $\boldsymbol{r}' = [x', y', z']^T$, $h(t, \boldsymbol{r}', \boldsymbol{r}_n)$ is the impulse response (IR) of the block and $*$ is the linear convolution operator (2.4). The IR takes into account both the direct path from $\boldsymbol{r}'$ to $\boldsymbol{r}_n$ and the reflections given by the block boundaries. Therefore, the signal $s(t, \boldsymbol{r}_n)$ in (9.1) contains delayed and attenuated versions of $\eta(t, \boldsymbol{r}')$, whose delays are determined by the distance travelled by the wavefronts and the wave velocity. In solids

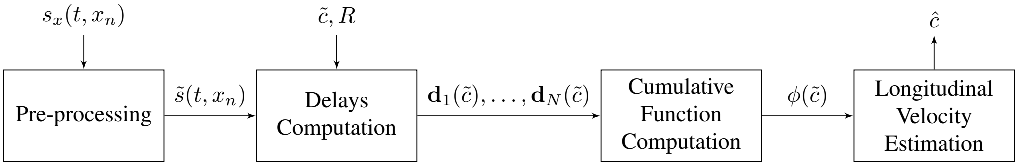

$s_x(t, x_n)$
Pre-processing
$\tilde{s}(t, x_n)$
$\tilde{c}, R$
Delays Computation
$\mathbf{d}_1(\tilde{c}), \ldots, \mathbf{d}_N(\tilde{c})$
Cumulative Function Computation
$\phi(\tilde{c})$
Longitudinal Velocity Estimation
$\hat{c}$

Figure 9.6: *The block diagram of the longitudinal velocity estimation procedure. The processing chain is divided in four stages: Pre-processing, Delays Computation, Cumulative Function Computation and Longitudinal Velocity Estimation.*

such as tone wood blocks, we can identify three types of waves (*longitudinal* or *axial*, *transverse* waves, and *bending* waves) defined according to the direction of displacement in the medium with respect to the wave propagation. In case of the longitudinal wave, the displacement in the medium is observed along the direction of propagation of the wave, i.e. longitudinal waves propagating along the x consist in local displacements of particles so that wavefronts are parallel to the yz plane. Here, we are interested in the estimation of the longitudinal wave velocity. Hence, we assume that the measurement and the excitation points are positioned at the endpoints of the wood block and aligned on the x axis, i.e. $x_n = \{0, L_x\}, y_n = y', z_n = z', \forall n = 1, \ldots, N$. Hence we define the longitudinal component of (9.1) as

$$s_x(t, x_n) = h(t, x', x_n) * \eta(t, x'). \tag{9.2}$$

The longitudinal wave equation [102] is

$$\frac{\partial^2 u(t,x)}{\partial t^2} = c^2 \frac{\partial^2 u(t,x)}{\partial x^2}, \tag{9.3}$$

where c is the longitudinal wave velocity and $u(t, x)$ is the longitudinal displacement field that corresponds to an elongation or contraction caused by the axial load $F(t, \mathbf{r}')$. The longitudinal velocity in (9.3) can be expressed in terms of the material characteristics [102]

$$c = \sqrt{\frac{EA}{m(1-\nu^2)}} = \sqrt{\frac{E}{\rho(1-\nu^2)}}, \tag{9.4}$$

where m is the mass of the block, while E, ν and $\rho = \frac{A}{m}$ are respectively the Young's modulus, the Poisson ratio and the density of the material. From (9.4), we can consequently provide an estimate of the Young's modulus E, assuming a suitable choice of the Poisson ratio ν.

9.2.2 Longitudinal Wave Speed Estimation

The proposed procedure for the estimation of the longitudinal wave velocity in tone wood blocks adopts a model fitting approach and given the measured signals, it is able to estimate in a rake receiver fashion, the longitudinal velocity exploiting a priori information provided by the signal model (9.2). The procedure receives as input the signals s_n with $n = 1, \ldots, N$ and an integer R that indicates the number of reflections to be considered during the estimation. The system provides as output an estimate $\hat{c}$ of the longitudinal velocity c by testing a set of hypothetical values of the velocity. The block diagram of the longitudinal wave velocity estimation procedure is depicted in Figure 9.6.

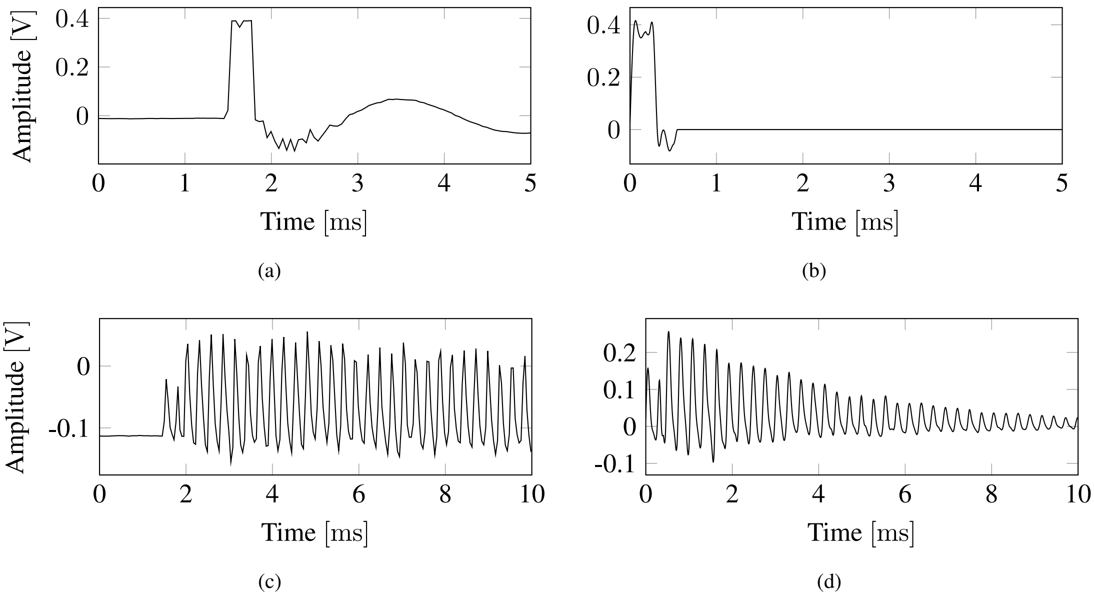


Figure 9.7: *The hammer signal before (a) and after (b) the pre-processing stage. The response measured by an accelerometer before (c) and after (d) the pre-processing stage.*

Pre-processing

In this step, we process the input signals (9.2) in order to improve the signal-to-noise ratio of the signals and avoid leakage phenomena [101]. In practice, the samples before the occurrence of the impulse are discarded and an exponential smoothing window $w(t)$ is applied to the signal captured by the nth accelerometer, yielding

$$\tilde{s}(t, x_n) = w(t) * s_x(t, x_n) \tag{9.5}$$

with $w(t) = e^{-\frac{t}{\tau}}$ and $w(t) = 0\,\forall t < t_i$ where t_i is the time instant at which the impulse occurs. In Figure 9.7, an excitation signal and a response are depicted both before and after the pre-processing stage.

Delays computation

Considering an hypothetical velocity $\tilde{c}$ and the required number of reflections R, a set of hypothetical delays from the impact point $\boldsymbol{r}'$ to the nth sensor location $\boldsymbol{r}_n$ are computed for all the N measurement points. Let us define the vector $\mathbf{d}_n$ of the delays associated to the nth sensor as

$$\mathbf{d}_n(\tilde{c}) = \left[\frac{l_0^n}{\tilde{c}}, \dots \frac{l_{R-1}^n}{\tilde{c}}\right] \in \mathbb{R}^{1\times R} \tag{9.6}$$

where $l_k^n = 2L_x k + l_0^n$ where $k = 0 \dots, R-1$ is the length of the path related to the kth reflection as seen by the nth sensor. We remark that, here, the length of the wood block L_x is assumed to be known. Commonly the wood block are cut by numerically controlled machine, therefore the accurate dimensions of the block are known or readily measurable. The index $k = 0$ refers to the direct path between $\boldsymbol{r}'$ and $\boldsymbol{r}_n$. This corresponds to the distance $l_0^n = \|\boldsymbol{r}' - \boldsymbol{r}_n\| = |x' - x_n|$. Each element of the vector

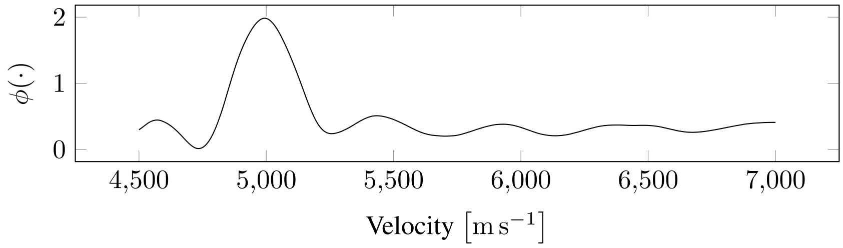


Figure 9.8: *An example of the cumulative function* $\phi(\cdot)$ (9.7) *evaluated at different velocity values.*

$\mathbf{d}_n(\tilde{c}) = l_k^n/\tilde{c}$ corresponds to the delay, expressed in seconds, given by the hypothetical velocity $\tilde{c}$ and the distance l_k^n traveled after k reflections.

Cumulative Function Computation

This step represents the core of the proposed method for the estimation of the longitudinal velocity. We exploit the delays computed in the previous step (9.6) in order to evaluate the fitness of the velocity $\tilde{c}$ with the data at hand. We test $\tilde{c}$ computing a cumulative function of the measurements defined as

$$\phi(\tilde{c}) = \sum_{n=1}^{N} \sum_{k=0}^{R-1} \tilde{s}(\mathbf{d}_{n,k}(\tilde{c}), x_n), \tag{9.7}$$

where $\mathbf{d}_{n,k}(\tilde{c})$ refers to the kth element in (9.6).

In practice, given a velocity, in (9.7) we sum the values of the N signals $\tilde{s}(t, x_n)$ in correspondence of the time instants in (9.6), which are determined by the combinations of the candidate velocity $\tilde{c}$ and the R reflection paths. Hence, the computation of the cumulative function (9.7) requires the computation of $N \times R$ summations for each velocity candidate $\tilde{c}$. It is worth noticing that in actual scenarios we work with discrete signals, consequently, the delay values in (9.6) may not perfectly correspond to sampled time instants. In order to compute (9.7) at the desired delays (9.6), a parabolic interpolation [8] is applied to the samples of $\tilde{s}(t, x_n)$.

Finally, in the last step an estimate of the longitudinal velocity is computed from (9.7). Inspecting Figure 9.8, we can notice that the graph presents a single prevailing peak and the cumulative function (9.7) attains its maximum where

$$\tilde{c}^{\star} = \arg\max_{\tilde{c}} \quad \phi(\mathbf{d}(\tilde{c})). \tag{9.8}$$

This can be interpreted as the fact that for a specific velocity value $\tilde{c}^{\star}$, the relative delay vectors match the actual reflection delays in the measurements. As a consequence, the values $\tilde{s}(\mathbf{d}_n(\tilde{c}^{\star}), x_n)$ in (9.7) will correspond to the peaks in the signals (see Figure 9.8). Therefore, we assume as an estimate of the longitudinal velocity c, the value $\tilde{c}^{\star}$ for which the cumulative function is maximized (9.8). In practice, we evaluate (9.8) on a discrete set of J candidate velocities such that

$$\hat{c} = \tilde{c}^{\star} \in \{\tilde{c}_1, \dots, \tilde{c}_J\}. \tag{9.9}$$

It is worth noting that from the inversion of (9.4), it is possible to exploit $\hat{c}$ for the estimation of the material properties e.g. the Young's modulus.

9.2.3 Validation and Results

In order to validate the proposed technique, we tested the longitudinal velocity estimation both on simulated synthetic data and signals measured from actual tone wood blocks. The whole estimation procedure described in Section 9.2.2 is implemented in *MATLAB* [164].

Simulation setup

In order to evaluate the performance of the proposed technique we simulated rectangular blocks of homogeneous isotropic material with length $L_x \in \{0.5, 1\}\,\mathrm{m}$, width $L_y = 0.15\,\mathrm{m}$ and height $L_z = 0.03\,\mathrm{m}$. The $N = 2$ impulse responses (9.2) were computed using the image source method [66, 118, 119, 121] at two different sampling frequencies $Fs_1 = 22.05\,\mathrm{kHz}$ and $Fs_2 = 44.1\,\mathrm{kHz}$. An additive sensor noise is simulated using a random white Gaussian noise, whose variance is set so that the desired signal to noise ratio at each sensor is $60\,\mathrm{dB}$. For each block we varied the wave speed in the range $c \in \{1000, 8000\}\,\mathrm{m\,s^{-1}}$ with a step of $500\,\mathrm{m\,s^{-1}}$. The estimation of the longitudinal wave velocity is evaluated in terms of the relative error

$$\varepsilon_{\mathrm{rel}}(c) = \left| \frac{c - \hat{c}}{c} \right|, \tag{9.10}$$

where c is the actual longitudinal velocity and $\hat{c}$ is the estimate given by (9.9) considering $J = 4096$ uniformly sampled candidates $\tilde{c} \in \{c/1.2, 1.2c\}$. As regards the algorithm parameters, in the pre-processing stage we adopted $\tau = 10\,\mathrm{ms}$ (9.5) while for the delay estimation (9.6) $R = 15$ reflections are considered.

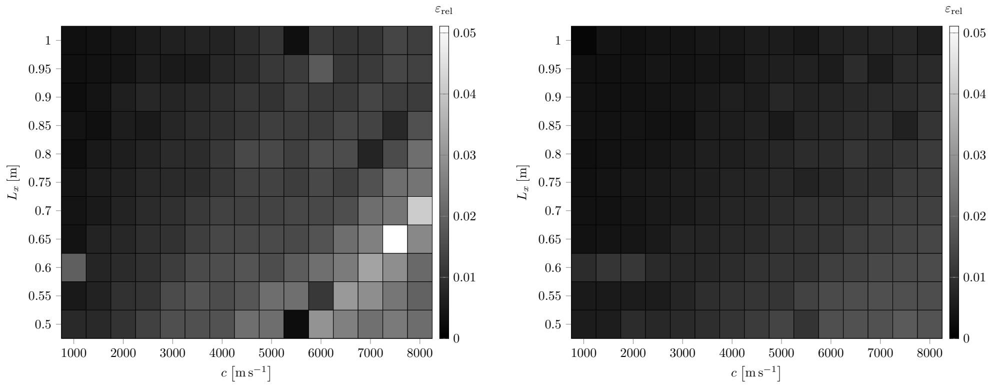
Measurement setup

The impulsive excitation and the response of the wood block (9.2) are recorded using $N = 2$ accelerometers *ADXL326* by *Analog Devices* [74]. The sensors are connected to the *Bela Mini* [30], an acquisition board that performs AD conversion with sampling rate $Fs = 22.05\,\mathrm{kHz}$. It is worth noticing that a calibration step is needed in order to guaranteed the synchronization of the sensors. The algorithm parameters were set as in the simulation setup (see Section 9.2.3). We measured the longitudinal wave velocity in $K = 7$ tone wood blocks made of red spruce coming from woods of Trentino South Tyrol, in Italy. The length of the K blocks are reported in the last column of Table 9.3. A total number of 5 measurements have been performed for each tone wood block and the average of the obtained values have been considered as the longitudinal wave velocity estimate. In order to asses the performance of the proposed procedure, we compare the obtained results with the tap tone and when available with the TOF methods.

The tap tone technique [128] estimates the longitudinal velocity $\hat{c}_{\mathrm{T}}$ as follows

$$\hat{c}_{\mathrm{T}} = \frac{0.973 \cdot f \cdot L_x}{h}, \tag{9.11}$$

where L_x and $h = \frac{L_{y1} + L_{y2}}{2}$ are the wood block length and the average thickness, respectively. The resonance frequency f is manually determined by the violin maker by tapping the block close to an antinode of the resonance mode, while holding lightly



(a)

(b)

Figure 9.9: *The relative error* $\varepsilon_{\rm rel}$ (9.10) *for the simulations with* $Fs_1 = 22.05\,\text{kHz}$ *(a) and* $Fs_2 = 44.1\,\text{kHz}$ *(b).*

Block	$\hat{c}[\mathrm{m\,s^{-1}}]$							[m]
	Rep. 1	**Rep. 2**	**Rep. 3**	**Rep. 4**	**Rep. 5**	**Average**	**Std. Deviation**	L_x
1	6236	6187	6220	6123	5990	6151	100	0.64
2	5806	5947	5831	5983	5891	5892	75	0.54
3	4937	4985	5083	5174	5217	5079	120	0.30
4	4932	5192	4879	5152	5283	5088	173	0.41
5	6212	6153	6359	5912	5960	6119	184	0.45
6	5025	5299	5245	5211	5274	5211	109	0.45
7	5888	5862	6026	5601	5511	5778	214	0.45

Table 9.3: *Longitudinal velocity estimates obtained from the measurements of the tone wood blocks.*

the wood on a nodal line. It is worth noting that (9.11), regarded by violin makers as a ground truth, is valid under the assumption that the tonewood block can be approximated by a bar. The TOF was measured using the Lucchi meter [281], an ultrasonic tester designed to measure the TOF in a piece of wood. Through the knowledge of L_x and the TOF it is possible to obtain the average longitudinal wave velocity.

Results

First we evaluate the performance of the sound wave velocity technique on a set of impulse responses obtained through simulations. In Figure 9.9 the relative error $\varepsilon_{\rm rel}$ (9.10) is reported considering the two different sampling frequencies. In general, the proposed technique provides a good estimation for both the sampling frequencies with $\varepsilon_{\rm rel} \leq 0.051\,(5.1\,\%)$. Inspecting both Figure 9.9(a) and Figure 9.9(b) we can observe that $\varepsilon_{\rm rel}$ tends to increase with the wave velocity c, while it decreases with the block length L_x. As expected, the estimation greatly improves at Fs_2, with $\varepsilon_{\rm rel} \leq 0.017\,(1.7\,\%)$ as depicted in Figure 9.9(b).

The second evaluation of the proposed sound wave velocity estimation technique employs actual impulse responses measured on tone wood blocks. In Table 9.3, the estimated velocities for each measurement are reported along with the average value over the repetitions and the standard deviation. It is worth noting that the consistency and repeatability of the measurements is confirmed by the standard deviations in Table 9.3, which present relatively small values with respect to the velocity magnitudes.

Block	$\hat{c}[\mathrm{m\,s^{-1}}]$			**% Error w.r.t. tap tone**	
	Tap Tone	**Lucchi meter**	**Proposed**	**Lucchi meter**	**Proposed**
1	6059	-	6151	-	1.5
2	5886	-	5892	-	0.1
3	5010	-	5079	-	1.4
4	4605	4946	5088	7.4	10.5
5	6420	5885	6119	8.3	4.7
6	5372	5127	5211	4.6	3.0
7	5763	5575	5778	3.3	0.3

Table 9.4: *Estimated velocities given by the considered techniques. The percentage error with respect to the tap tone is reported.*

In Table 9.4, the estimated velocities are reported along with the results obtained using the tap tone and the Lucchi meter. We choose to evaluate the estimation in terms of the percentage error with respect to (9.11) since the tap tone method is considered to be the standard by violin makers. Inspecting the third column of Table 9.4, we can notice that the estimates obtained with the proposed method are close to the reference ones.

Moreover the proposed estimation outperformed the estimates given by the TOF technique for all the blocks except for $k = 4$. A more detailed analysis on this specific case shows us that the block number 4 presented inhomogeneous wood grains with respect to the other samples. Hence, the results given by the tap tone (9.11) are less reliable due to the bar approximation. This hypothesis is confirmed by the fact that the estimates given both by the Lucchi meter and the proposed estimator are close to each other.

CHAPTER 10

Virtual Source Synthesis

The purpose of this chapter is to discuss the implementation of a full EAR model. We demonstrate the potential of the VS model introduced in Section 10.1 in the EAR framework through a proof-of-concept simulation.

Regarding the synthesis of VSs, in the previous chapters we discussed the derivation of VS models from measurements of actual acoustic sources or simulations of the source vibroacoustic behavior. The VS model described in Section 10.1 is borrowed from the parametric sound field reconstruction technique of Section 5.3, thus it requires the synthesis of direct and diffuse sound field components. The adoption of a parametric description of VSs provides two main advantages. On the one hand, we can describe the virtual source by means of few parameters that we can arbitrarily define. These parameters concern the location of VSs in the scene, the signal emitted and not least radiation characteristics. On the other hand, we maintain the comprehensibility of parametric sound field reconstruction techniques, providing a self-contained framework that considers both virtual and actual sound fields within the same description. While the definition of the direct component is limited to the characteristics of the VS only, and it does not require further information, the diffuse component is inherently related to the EAR setup. Hence, in Section 10.1, we discuss possible alternatives for devising the diffuse component in accordance with the available data.

In Section 10.2 a proof-of-concept simulation of the EAR framework is proposed. Our goal is to display the combination of both the parametric sound field reconstruction method, proposed in Section 5.3 and the virtual source model of Section 10.1 for the development of a complete EAR framework, in which virtual sources are added to a real acoustic environment.

10.1 EAR Signal Model

The signal of the VM computed in the EAR framework is a combination of the VM signal derived from the sound field reconstruction and the contribution given by the VSs. More specifically, we adopt the parametric VM model (5.2) discussed in Section 5.3 with the addition of the parametric VS signal model. Therefore, let us define the VM signal in the EAR scenario as

$$\hat{\hat{S}}(t,\omega,\check{\boldsymbol{r}}_v) = \hat{S}(t,\omega,\check{\boldsymbol{r}}_v) + \bar{S}(t,\omega,\check{\boldsymbol{r}}_v), \tag{10.1}$$

where $\hat{S}(t,\omega,\check{\boldsymbol{r}}_v)$ is the sound field reconstruction (5.2) and $\bar{S}(t,\omega,\check{\boldsymbol{r}}_v)$ is the additional contribution of the VSs. For the ease of the reader, here, we report the definition of $\hat{S}(t,\omega,\check{\boldsymbol{r}}_v)$

$$\hat{S}(t,\omega,\check{\boldsymbol{r}}_v) = C_v(\omega)\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v) + Q_v(\omega)\hat{S}_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v).$$

The contribution of the VSs $\bar{S}(t,\omega,\check{\boldsymbol{r}}_v)$ is given as the sum of each VS component

$$\bar{S}(t,\omega,\check{\boldsymbol{r}}_v) = \sum_{\bar{n}=1}^{\bar{N}} C_v(\omega)\bar{S}_{\bar{n},\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v) + Q_v(\omega)\bar{S}_{\bar{n},\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v). \tag{10.2}$$

where $\bar{S}_{\bar{n},\mathrm{dir}}(\cdot)$ is the direct signal virtually emitted by the $\bar{n}$th VS and $\bar{S}_{\bar{n},\mathrm{diff}}(\cdot)$ is the virtual diffuse sound component generated by the VS. The direct virtual sound $\bar{S}_{\bar{n},\mathrm{dir}}(\omega,t,\check{\boldsymbol{r}}_v)$ must accurately reflect the source acoustic characteristics. Therefore here, in order to describe the direct sound of a VS, we assume the parametric model from (5.3), as

$$\bar{S}_{\bar{n},\mathrm{dir}}(\omega,t,\check{\boldsymbol{r}}_v) = \bar{D}_{\bar{n}}(\omega,t,\check{\theta}_{v,\bar{n}},\check{\phi}_{v,\bar{n}})H(\omega,\check{\boldsymbol{r}}_v,\bar{\boldsymbol{r}}'_{\bar{n}})\bar{S}_{\bar{n}}(\omega,t,\bar{\boldsymbol{r}}'_{\bar{n}}) \tag{10.3}$$

where $H(\omega,\check{\boldsymbol{r}}_v,\bar{\boldsymbol{r}}'_{\bar{n}})$ is the Green's function (3.29), $\bar{S}_{\bar{n}}(\omega,t,\bar{\boldsymbol{r}}'_{\bar{n}})$ is the source signal and $\bar{D}_{\bar{n}}(\omega,t,\check{\theta}_{v,\bar{n}},\check{\phi}_{v,\bar{n}})$ is the directivity pattern function. It is worth underlining that we inherently assume the VS being in the far field with respect to the VM. Hence, the VS is considered as a point-like source with an arbitrary directivity pattern. We remark that since the propagation $H(\omega,\check{\boldsymbol{r}}_v,\bar{\boldsymbol{r}}'_{\bar{n},})$ is inversely proportional to the distance between the VSs and VMs it is limited to a maximum value $H_{\max} = -6\,\mathrm{dB}$. Therefore, the $\bar{n}$th VS can be compactly described using three parameters: the location $\bar{\boldsymbol{r}}_{\bar{n}}$, the source signal $\bar{S}_{\bar{n}}(\omega,t,\bar{\boldsymbol{r}}'_{\bar{n}})$, and the directivity pattern $\bar{D}_{\bar{n}}(\cdot)$. As a matter of fact, the VS source signal $\bar{S}_{\bar{n}}(\omega,t,\bar{\boldsymbol{r}}'_{\bar{n}})$ provides the characteristics of the source timbre, i.e., the time-frequency and loudness evolution of the radiated sound. The directivity pattern $\bar{D}_{\bar{n}}(\cdot)$, is a fundamental parameter for accurately rendering the acoustic source. In practice, the directivity pattern describes the acoustic energy radiation in space, a property that is strictly related to the physics of the source. In the context of modeling a virtual violin, we can define the directivity pattern both from measurements performed on actual instruments and from simulations of the instruments obtained through FEM. As regards the source signal, we could employ the data acquired by a microphone placed in the proximity of an actual instrument while being played, or we could consider the signal generated by a silent electric violin [157].

As far as the diffuse component $\bar{S}_{\bar{n},\mathrm{diff}}(\cdot)$ is concerned, the computation is directly related to the EAR setup. In fact, the available knowledge on the EAR setting can be exploited in order to model the diffuse component with different levels of accuracy. Therefore, we can identify different possible scenarios for the computation of the VS diffuse component accordingly to the EAR setup.

- The knowledge of the acoustic environment represents the most advantageous information that we could exploit in order to compute the VS diffuse component. In fact, if the geometry of the environment (typically a room) in the EAR system is known, we can straightforwardly compute the acoustic field generated by a VS. Usually, this is implemented computing the impulse response between the acoustic source (VS) and the receiver location (VM). Different algorithms for the computation of the acoustic field exist which can be mainly divided in two categories: numerical approaches where the wave equation is numerically solved and techniques based on geometrical acoustics. Although numerical methods provide an accurate computation of the acoustic field, they are computationally expensive, hence we usually employ faster, but less accurate geometrical acoustics methods. Such techniques rely on the assumption that the amplitude of a wavefront varies little over a distance that is comparable of the wavelength and that the radii of the curvature of the wavefront is larger than the wavelength [204, 231]. Hence, sound is assumed to travel along rays. An example of geometrical acoustics algorithms is the well-known image source method (ISM) [11], in which reflections are generated by artificial sources obtained "mirroring" the actual source against the room walls. Alternative approaches are represented by ray tracing [12, 144, 239] and the more efficient beam tracing [20, 22, 103, 104, 162, 251]. It is worth noting that geometrical acoustics methods are inherently unable to model diffraction and diffuse phenomena, hence extensions have been proposed in the literature [19, 182, 231, 238]. A different approach to the generation of RIRs is represented by feed-back delay network (FDN) [130, 249, 266]. FDN represents an efficient and parametric technique for the computation of impulse responses that is based on a set of delay lines interconnected by a feedback matrix. By controlling the parameters of the FDN, the computation of the RIR can be guided using isotropic assumption [168, 212, 235, 249] or directional information [9, 10, 234]. While in the context of VR, the environment is virtually defined by the system, in ER, we might infer the acoustic environment geometry from some measurements. As instance one can exploit acoustic sensors for both the sound field reconstruction and the room geometry inference (RGI). RGI has been widely studied in the literature and generally they concerns the estimation of the location and orientation of wall reflectors from the signal of microphones. Usually, arrays of microphones [21, 53, 97, 98, 223] or loudspeakers [84–86, 261] are employed and the geometry of the room is estimated from a set of measured RIR.

- A second scenario for the computation of the VS diffuse component concerns the knowledge of a set of measured RIRs. In particular, we can exploit a set of known RIR for modeling the impulse response and thus compute diffuse (isotropic) component of the VS source from such model. In [265] a parametric method for the modeling a RIR has been presented. The method relies on realizations of velvet noise signal [125, 129, 233] which are properly filtered in order to match the measured impulse responses. The adoption of velvet noise allows a computational efficient implementation of the RIR, while the parameters of the filters are used as a model to represent the RIR.

- Finally, when no information about the acoustic environment is given we can mainly follow two approaches. On the one hand, one could discard the VS diffuse component. Despite reducing the immersivity of the VS rendering, this solution might be desirable in applications when we are interested in distinguishing between VSs and actual acoustic sources in the scene. On the other hand, we could employ an arbitrary user-controlled reverberator that can be tuned in order to provide a realistic diffuse component to the VS.

In Figure 10.1 a block diagram of the overall EAR approach is shown. The main goal of our approach to EAR is to combine the sound field navigation, i.e., the sound field reconstruction at the VM location with the spatial rendering of VSs with arbitrary directivity patterns. Hence, Figure 10.1 can be divided in two main parallel blocks: the sound field reconstruction and the virtual source rendering parts.

On the one side, as described in Section 5.1, we process the signals of the arrays $X(\cdot)$ in order to estimate the parameters of the model. The estimated parameters are then adopted for synthesizing the VM signal $\hat{S}(\check{\boldsymbol{r}}_v)$ (5.2) at the vth VM location.

On the other side, the sound field contribution of the VS is computed. In particular, as previously discussed, we describe the $\bar{n}$ VS through its position $\bar{\boldsymbol{r}}'$, source signal $\bar{S}_{\bar{n}}$ and directivity pattern $\bar{D}_{\bar{n}}$. In the previous chapters, we discussed different approaches to the estimation or the prediction of the directivity patterns of sound sources with a focus on the violin. The VS contribution at the VM $\bar{S}(\check{\boldsymbol{r}}_v)$ (10.2) is computed as the superposition of the direct component $\bar{S}_{\mathrm{dir}}(\check{\boldsymbol{r}}_v)$ (10.3) and the diffuse component $\bar{S}_{\mathrm{diff}}(\check{\boldsymbol{r}}_v)$ which requires additional information for the RIR computation.

Finally, the vth VM parameters concerns its location $\check{\boldsymbol{r}}$, pick-up pattern $C_v(\cdot)$ and sensitivity $Q_v(\cdot)$ and they are adopted both for the sound field reconstruction and the the VS rendering. Lastly, the combination of the outputs of the two main blocks $\hat{S}(\check{\boldsymbol{r}}_v)$ and $\bar{S}_{\mathrm{dir}}(\check{\boldsymbol{r}}_v)$ provides the EAR signal $\hat{\bar{S}}(\check{\boldsymbol{r}}_v)$ (10.1) comprised of the reconstructed sound field and the additional VSs.

10.2 EAR Proof of Concept

In this section, we show the parametric EAR approach through a simulation in which VSs are inserted in an actual sound field with active acoustic sources. The aim of this simulation is to present the possibility of the EAR adopting the VS model introduced in the previous section. As a proof of concept we adopt a string trio scenario composed of a cello and two violins. While the cello is considered as an acoustic source actually present in the environment, the remaining two instruments are included in the scene as VSs. As VSs, we consider a violin whose directivity pattern has been estimated using the technique in Chapter 7 and a violin whose pattern has been predicted using a FEM simulation of a simplified model of a violin body. Therefore, one instrument provides a virtual version of an historic violin and the second VS represents a "dummy" model of a generic violin whose acoustic behavior is computed through FEM.

Therefore, in this proof of concept, we adopt both "measured" and "simulated" directivity patterns. As regards the diffuse component, its is computed through the convolution of the source signal with late part of the RIR given by [121]. Hence, here we are

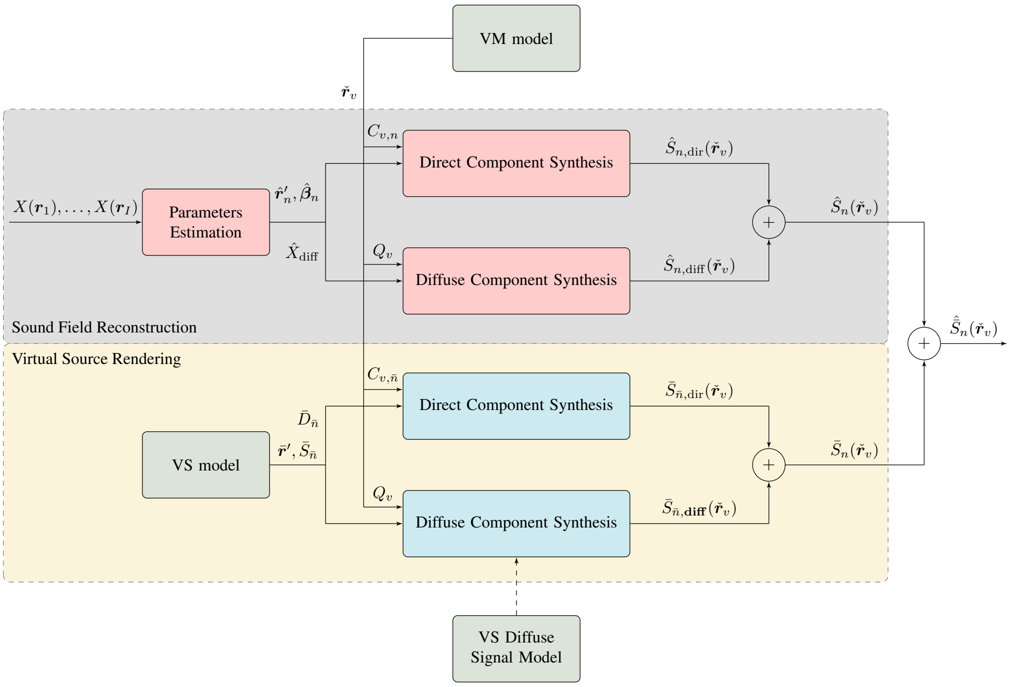

VM model
$\check{\boldsymbol{r}}_v$
$C_{v,n}$
Direct Component Synthesis
$\hat{S}_{n,\mathrm{dir}}(\check{\boldsymbol{r}}_v)$
$X(\boldsymbol{r}_1),\dots,X(\boldsymbol{r}_I)$
Parameters Estimation
$\hat{\boldsymbol{r}}'_n, \hat{\boldsymbol{\beta}}_n$
$\hat{X}_{\mathrm{diff}}$
$\hat{S}_n(\check{\boldsymbol{r}}_v)$
Q_v
Diffuse Component Synthesis
$\hat{S}_{n,\mathrm{diff}}(\check{\boldsymbol{r}}_v)$
Sound Field Reconstruction
$\hat{\bar{S}}_n(\check{\boldsymbol{r}}_v)$
Virtual Source Rendering
$C_{v,\bar{n}}$
Direct Component Synthesis
$\bar{S}_{\bar{n},\mathrm{dir}}(\check{\boldsymbol{r}}_v)$
$\bar{D}_{\bar{n}}$
VS model
$\bar{\boldsymbol{r}}', \bar{S}_{\bar{n}}$
$\bar{S}_n(\check{\boldsymbol{r}}_v)$
Q_v
Diffuse Component Synthesis
$\bar{S}_{\bar{n},\mathbf{diff}}(\check{\boldsymbol{r}}_v)$
VS Diffuse Signal Model

Figure 10.1: *The EAR rendering block diagram. Two main macro blocks are present: the sound field reconstruction (light grey) and the virtual source rendering (light yellow) blocks. Light red blocks refer to the sound field reconstruction steps. In particular, the microphone signals are exploited for the estimation of the sound field parameters. The estimated parameters provides the sound field reconstruction at the VM location. The light blue blocks, instead concerns the computation of the VS sound field contribution at the VM. Light green blocks are referred to the models of both the VM and the VS. The parameters defining these models can be arbitrarily specified according to the EAR scenario.*

exploiting the full knowledge of the acoustic environment as discussed in the previous section.

10.2.1 Violin Model FEM Simulation

As discussed in the previous chapters, the parameters required by the VS model can be obtained from simulations of acoustic sources. Here, the simulation is performed with *COMSOL Multiphysics*®. Clearly, the performance of the VS is directly related to the simulated model accuracy. In the case of violins, an accurate model of the instrument geometry and material properties is still an open challenge. Therefore, here we adopt a simplified model of the violin body as shown in Figure 10.2. The maximum dimension of the violin body is $L_y = 0.32\,\mathrm{m}$, the lower bout dimension is $L_x = 0.21\,\mathrm{m}$ and the height of the full body is $L_z = 0.041\,\mathrm{m}$. The simplified violin model adopts a flat design for both the top and bottom plates with a thickness of $2.5\,\mathrm{mm}$ and $3\,\mathrm{mm}$, respectively. The bass bar is $0.25\,\mathrm{m}$ long with a square section of $3\,\mathrm{mm}$ width. The sound post, connecting the two violin plates has a radius of $3.2\,\mathrm{mm}$. Finally the bridge feet position is indicated by the two small rectangles aligned at $y = 0$ (see Figure 10.2). As regards the material properties of the model, in order to obtain a close simulation of the wood behavior, we adopt orthotropic definition of Sitka spruce with the same parameters given in Table 8.1.

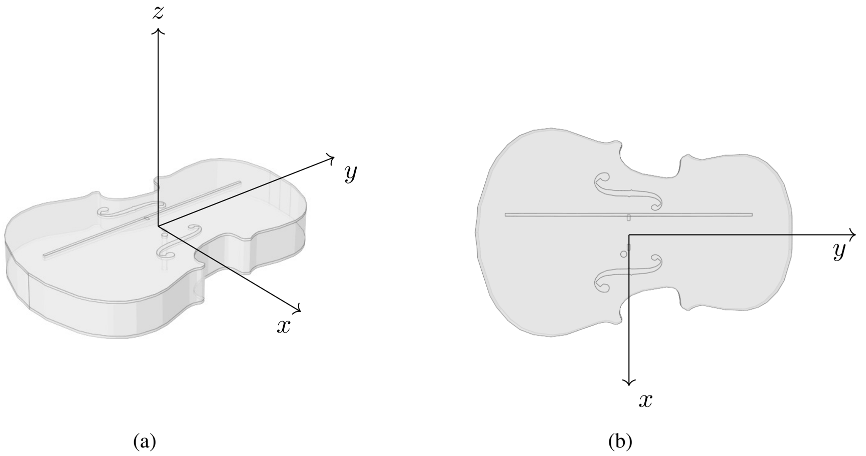


Figure 10.2: *(a) 3D graphical representation of the "dummy" violin model adopted in the FEM simulation. (b) Top view of the violin model.*

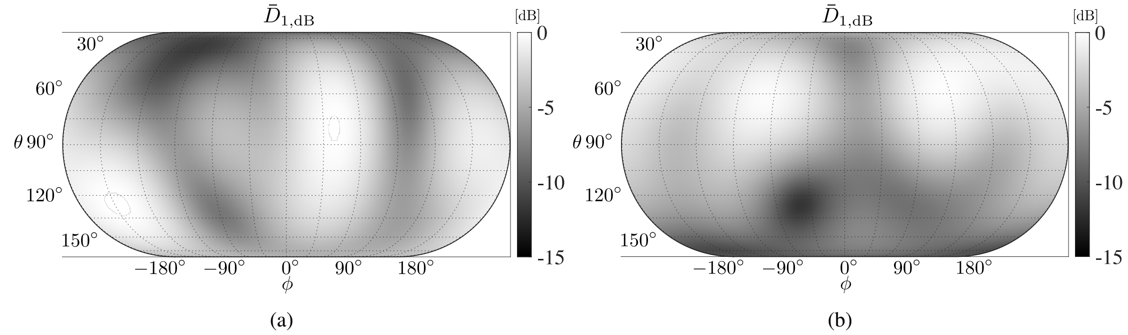


Figure 10.3: *Directivity pattern of the VS obtained from the FEM simulation of a simplified violin model. The directivity pattern are expressed in* dB *and computed at* $880\,\mathrm{Hz}$ *(a) and* $1760\,\mathrm{Hz}$ *(b) with order* $L = 4$.

In order to compute the response of the violin model, we performed a frequency domain study in *COMSOL Multiphysics®* varying the harmonic load frequency in the range $[196, 4000]\mathrm{Hz}$. A unitary load is applied at the bridge location in order to approximate the excitation provided by a played string. The radiated acoustic pressure is then computed over a sphere surrounding the instrument, whose radius is varied according to the analyzed frequency in order to preserve the far field condition.

The acoustic pressure data is sampled adopting spiral sampling [243] on a total of $B = 1024$ points on the sphere from which directivity pattern $\hat{\bar{\mathbf{d}}}_1$ is expressed using the spherical harmonic representation of (7.5). Hence, the spherical harmonics coefficients of the simulated violin model are obtained similarly to (7.6) as

$$\hat{\bar{\mathbf{c}}}_1^{(L)}(\omega) = \mathbf{Y}^\dagger \hat{\bar{\mathbf{d}}}_1. \tag{10.4}$$

It follows that we can compute the VS directivity pattern for arbitrary VM locations in

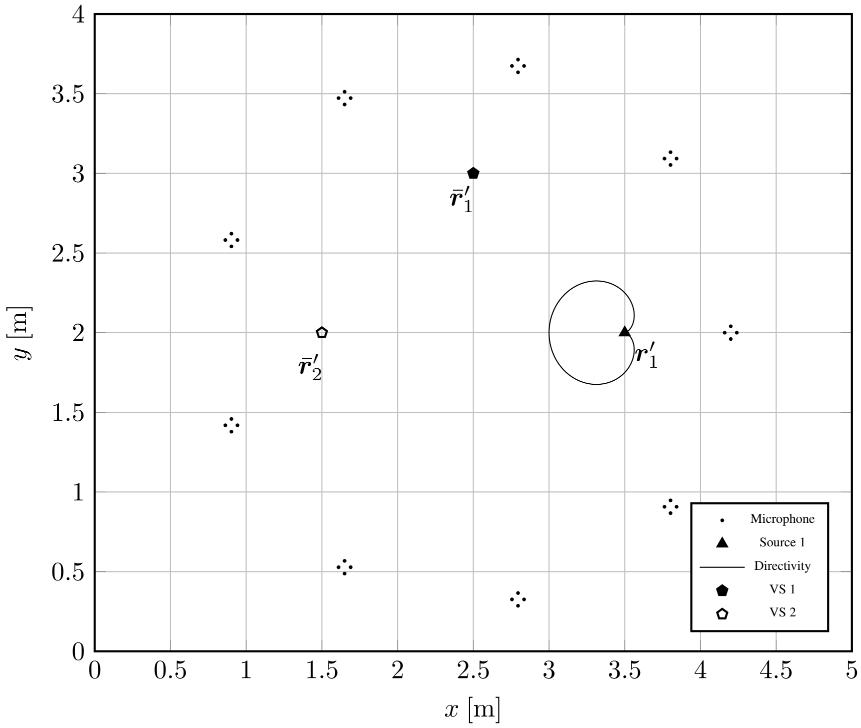


Figure 10.4: *2D graphical representation of the EAR setup. An acoustic source is positioned in the scene at* $\boldsymbol{r}'_1 = [3.5, 2]^T\mathrm{m}$*, while two VS are included as located at* $\bar{\boldsymbol{r}}'_1 = [2.5, 3]^T\mathrm{m}$ *and* $\bar{\boldsymbol{r}}'_2 = [1.5, 2]^T\mathrm{m}$

(10.3) using the spherical harmonics model of (7.4) as

$$\bar{D}_1(\omega, \check{\theta}_{v,1}, \check{\phi}_{v,1}) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \hat{\bar{C}}_{lm,1}(\omega) Y_{lm}(\check{\theta}_{v,1}, \check{\phi}_{v,1}), \tag{10.5}$$

where $\hat{\bar{C}}_{lm,1}(\omega)$ are the coefficients estimated in (10.4) and the dependency on time is omitted for simplicity. In Figure 10.3 we report examples of the VS directivity pattern (10.5) obtained from the FEM simulation of the violin model.

10.2.2 EAR Setup and Parameters

The EAR setup is illustrated in Figure 10.4. Similarly to the virtual miking setup described in Section 5.3.6, we employ $A = 9$ circular microphone arrays with radius $0.04\,\mathrm{m}$, composed on $M = 4$ omnidirectional microphones each. Therefore, the number of deployed sensors is $I = A \times M = 36$. The room has dimensions $L_x = 5\,\mathrm{m}$, $L_y = 4\,\mathrm{m}$ and $L_z = 3\,\mathrm{m}$. The real acoustic source ($N = 1$), representing a cello, is located at $\boldsymbol{r}'_1 = [3.5, 2]^T\mathrm{m}$, while the $\bar{N} = 2$ violin VSs are positioned at $\bar{\boldsymbol{r}}'_1 = [2.5, 3]^T\mathrm{m}$ and $\bar{\boldsymbol{r}}'_2 = [1.5, 2]^T\mathrm{m}$, respectively. The source signals for both the actual source and the VSs are $5\,\mathrm{s}$ melodic extracts taken from [264]. For simplicity, the acoustic source (cello) presents a first-order cardioid directivity pattern with looking direction equal to 180°. Nevertheless, such assumption on its directivity pattern seems to be reasonable from the analysis in [170], where cello present a rather steady principal radiation region towards direction of sight of the player.

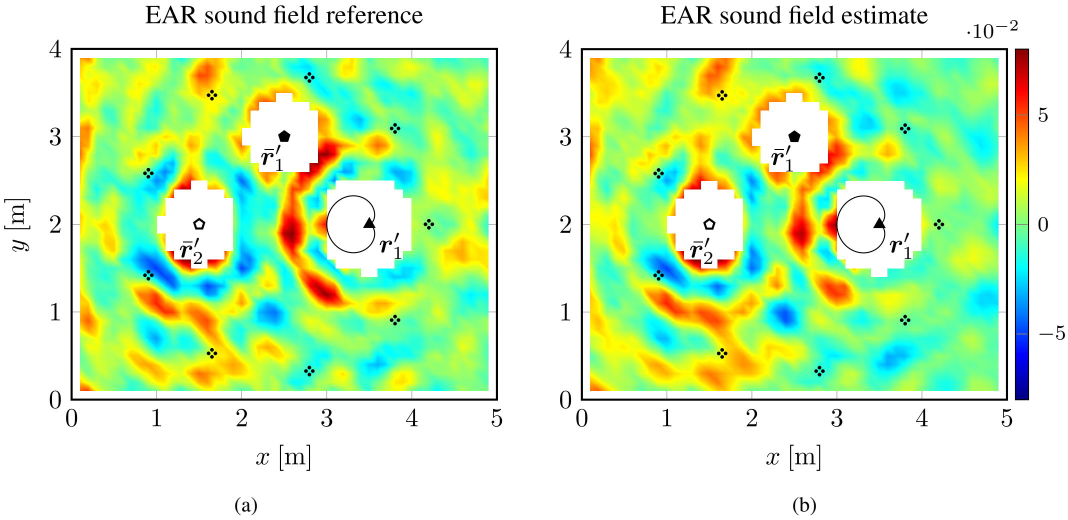


Figure 10.5: *Example of the actual sound field considering the VSs (a) and the estimate provided by the sound field reconstruction and the VSs (b). The time-domain sound field is shown at instant* $t = 1.5\,\mathrm{s}$.

The first VS is derived from the simulated violin model (see Section 10.2.1), while the second VS implements a virtual replica of the prestigious *Il Cremonese* violin by *Antonio Stradivari* whose directivity pattern is analyzed in Section 7.2. Therefore, the VSs present frequency dependent directivity patterns computed adopting a spherical harmonics expansion of order $L = 4$. It is worth underlining that the implemented VS models provide the spatial sound radiation of the two instruments, while the spectral sound characteristics are inherently related to the adopted source signal in (10.3) that is taken from [264]. We set the looking directions equal to $-90°$ for both the VSs. In order to evaluate the EAR sound field on a large area of the environment, we define the VMs sampling the plane where VSs, source and arrays lie. We define rectangular grid of VMs with a spatial sampling step $d = 0.1\,\mathrm{m}$ obtaining $V_x = \frac{L_x}{d} - 1$ and $V_y = \frac{L_y}{d} - 1$ points along the x and y axis, respectively. Hence, the locations of the $V = V_x \times V_y = 49 \times 39 = 1911$ VMs are given as

$$\check{\boldsymbol{r}}_v = [dv_x, dv_y]^T, \quad v_x = 1, \dots, V_x, \quad v_y = 1, \dots, V_y. \tag{10.6}$$

The aim of VM positioning is to cover a wide area of the room capturing the spatial sound characteristics of the signals of the virtual sources with respect to the actual source present in the scene. The signal at the microphones generated by the actual acoustic source is computed convolving the source signal with the source associated RIRs obtained through ISM [121].

As previously discussed, we assume that the environment of the EAR setup is known. Therefore, also the contributions of the VSs at the VM is estimated adopting the RIRs obtained through ISM [121]. The signals are processed at $16\,\mathrm{kHz}$ sampling rate and the STFT is performed with 1024 points and $64\,\mathrm{ms}$ Hamming window with $75\,\%$ overlap for both the analysis and the synthesis phase. As regards the rest of parameters required by the parametric sound field reconstruction technique, we adopted the same values of Section 5.3.6.

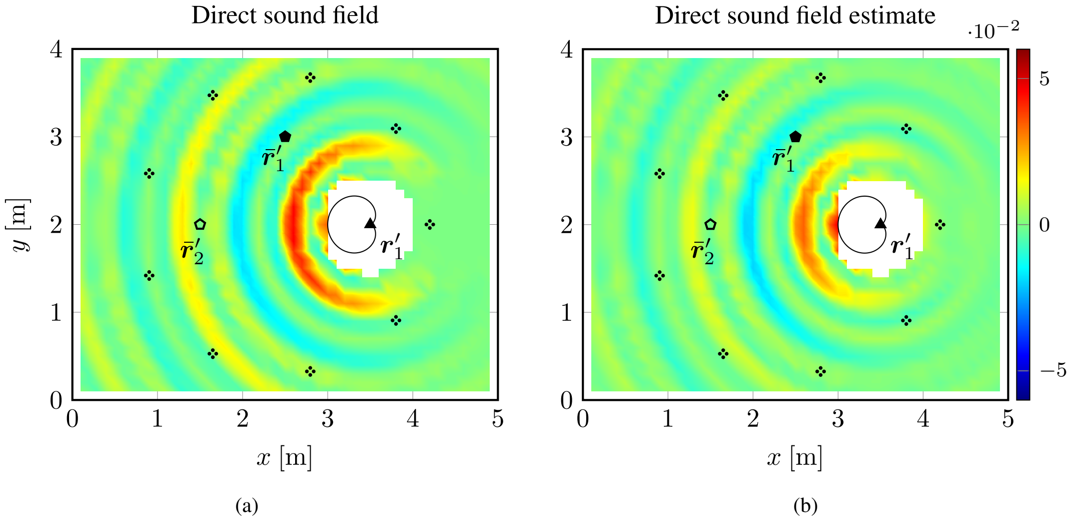


Figure 10.6: *Example of the direct sound field of the source (a) and the estimate provided by the parametric sound field reconstruction (b). The time-domain sound field is shown at instant $t = 1.5$ s.*

10.2.3 Results

In Figure 10.5, a snapshot of the EAR sound field at time $t = 1.5$ s is depicted. The time domain signal has been computed over the rectangular grid of VMs defined as (10.6). Although, in general, the spatial pattern of the EAR sound field is similar to the reference, it is possible to note small differences between the sound field in Figure 10.5(a) and the EAR signal in Figure 10.5(b). In particular, the EAR sound field appears as underestimated in the region “behind” the real source in $\boldsymbol{r}'_1$. This observation is consistent with the sound field reconstruction results in Section 5.3.6, for which the acoustic field estimate deviates for directions where the acoustic radiation given by the ideal cardiod pattern is null.

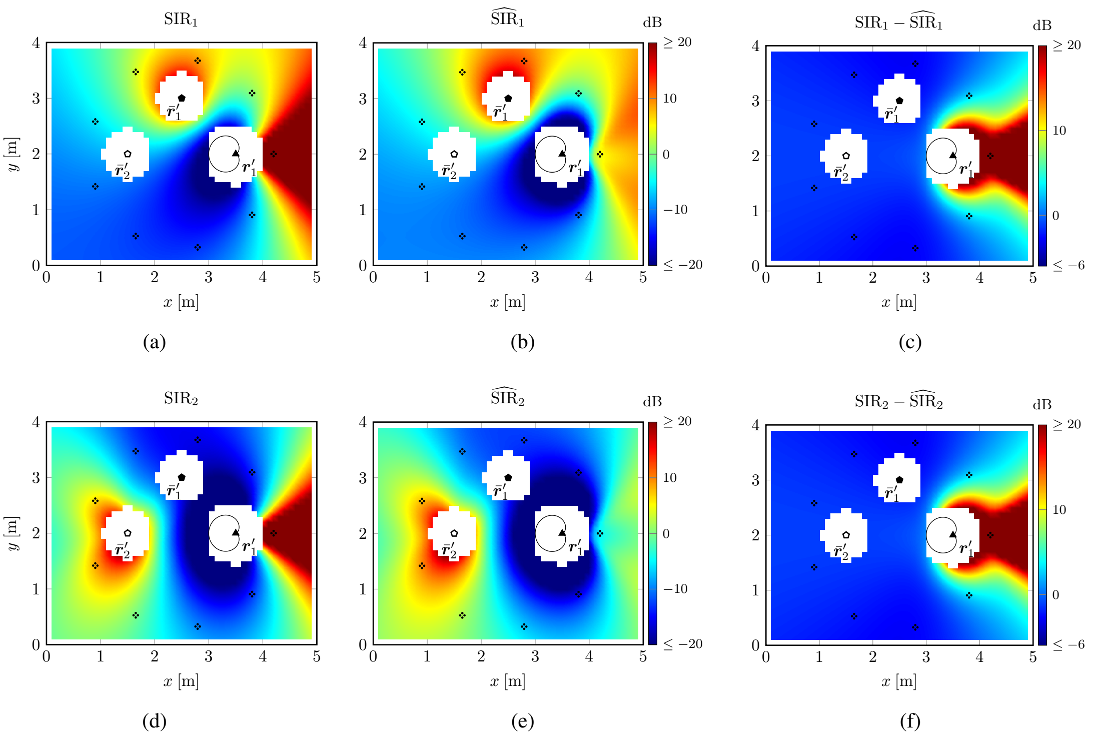
As far as the source localization is concerned, the estimated source position, (see Section 5.3.2) is $\hat{\boldsymbol{r}}'_1 = [3.518, 1.994]^T \mathrm{m}$, resulting in a localization error of $\|\boldsymbol{r}'_1 - \hat{\boldsymbol{r}}'_1\| = 0.019\,\mathrm{m}$.

In order to disclose the performance of the EAR framework in rendering VSs in an actual acoustic environment, we evaluate $\widehat{\mathrm{SIR}}$ (5.73) between each VS and the actual source active in the scene.

We choose to consider the $\widehat{\mathrm{SIR}}$ since it is directly related to the directivity pattern of the sources. The energy ratio between a real source and a VS is inherently influenced by the energy emission of the sources towards the VM locations. It follows that an accurate estimation of the real source directivity pattern is required as describe in Chapter 5, while as regards the VS, the modeling of the directivity pattern is directly reflected in the metrics. Hence, we can compare the $\widehat{\mathrm{SIR}}$ obtained with the EAR to the SIR that one would obtain if the VSs were actually present in the environment.

Moreover, in order to inspect performance of the SIR at different frequencies we introduce the $\widehat{\mathrm{SIR}}$ between the VSs and the real source as

$$\widehat{\mathrm{SIR}}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega) = \frac{\sum_t |\bar{S}_{\bar{n},\mathrm{dir}}(t, \omega_w, \check{\boldsymbol{r}}_v)|^2}{\sum_t |\hat{S}_{n,\mathrm{dir}}(t, \omega_w, \check{\boldsymbol{r}}_v)|^2}. \tag{10.7}$$



(a) (b) (c)

(d) (e) (f)

Figure 10.7: *Wideband* SIR *of to the first (a) and second VSs (d). The estimated* $\widehat{\mathrm{SIR}}$ *for the first source and the second VS are depicted in (b) and (e), respectively. The difference between the reference* SIR *and its estimate* $\widehat{\mathrm{SIR}}$ *for the first (c) and second VSs (f).*

	VS 1	VS 2
NMSE [dB]	−12.1	−18.2

where the dependency on frequency of the $\widehat{\mathrm{SIR}}$ is highlighted since the VSs provide a frequency-dependent directivity. We indicate with the subscript $\bar{n}$ the $\widehat{\mathrm{SIR}}$ computed with respect to the $\bar{n}$th VS. The VS direct signal $\bar{S}_{\bar{n},\mathrm{dir}}(\cdot)$ is defined according to (10.3) and it is reported here for the reader convenience

$$\bar{S}_{\bar{n},\mathrm{dir}}(\omega, t, \check{\boldsymbol{r}}_v) = \bar{D}_{\bar{n}}(\omega, t, \check{\theta}_{v,\bar{n}}, \check{\phi}_{v,\bar{n}}) H(\omega, \check{\boldsymbol{r}}_v, \bar{\boldsymbol{r}}'_{\bar{n},}) \bar{S}_{\bar{n}}(\omega, t, \bar{\boldsymbol{r}}'_{\bar{n}}),$$

with $\bar{n} = 1, 2$ for the simulated violin model and the violin virtual replica, respectively. Therefore, the directivity pattern $\bar{D}_{\bar{n}}(\omega, t, \check{\theta}_{v,\bar{n}}, \check{\phi}_{v,\bar{n}})$ has been derived from the violin body simulation for $\bar{n} = 1$ (see Section 10.2.1), while for the violin virtual replica $\bar{n} = 2$ it has been measured as described in Section 7.2. It is worth noting that here a 2D setup is assumed, hence the inclination angle is fixed as $\check{\theta}_{v,\bar{n}} = \pi/2$, $\forall v = 1, \ldots, V$, $\bar{n} = 1, \ldots, \bar{N}$.

We remark that $\hat{S}_{n,\mathrm{dir}}(\cdot)$ in (10.7) represents the source direct signal estimate given by the parametric sound field reconstruction method in (5.53). Figure 10.6 shows a snapshot of both the direct sound field and the estimate provided by the parametric sound field reconstruction (5.53). The time domain signals are obtained as the inverse STFT of $S_{n,\mathrm{dir}}(\cdot)$ and $\hat{S}_{n,\mathrm{dir}}(\cdot)$ for the reference and the estimate, respectively.

Similarly to what is done in Section 5.3.6, we evaluate $\widehat{\mathrm{SIR}}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$ at the VMs through $\mathrm{NMSE}_{\mathrm{SIR}}^{(\bar{n})}$ (8.8) computed with respect to the reference $\mathrm{SIR}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$. The reference $\mathrm{SIR}(\eta^{(\bar{n})}, \omega)$ is obtained adopting the actual source direct signal $S_{n,\mathrm{dir}}(\cdot)$ in (10.7) instead of the estimate $\hat{S}_{n,\mathrm{dir}}(\cdot)$ provided by the sound field reconstruction technique.

In Figure 10.7, the reference wideband SIR and its estimate $\widehat{\mathrm{SIR}}$ are depicted for both the VSs. In general, we can observe that the estimate follows the pattern of SIR in the environment. As expected, a deviation in $\widehat{\mathrm{SIR}}$ can be noted in the region shadowed by the cardiod pattern. This deviation agrees with the results in Figure 10.5 and it can be explained by the limits of the sound field reconstruction (see Section 5.3.6). In fact, inspecting Figure 10.7(c) and Figure 10.7(f), we observe a difference between SIR and $\widehat{\mathrm{SIR}}$ below 3 dB, in absolute value for the whole room, excluding the region "behind" the cardioid source. In this area, the absence of the real source acoustic emission produces SIR $\rightarrow \infty$ in the reference values. The $\mathrm{NMSE}_{\mathrm{SIR}}$, computed excluding the points for which SIR tends to infinity, corresponds to $-12.1\,\mathrm{dB}$ and $-18.2\,\mathrm{dB}$ for the first and the second VS, respectively.

In Figure 10.8 and Figure 10.9 examples of the narrowband $\widehat{\mathrm{SIR}}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$ are reported along with the reference values and their differences are depicted for the first and the second VS, respectively.

We can note that the trend of $\widehat{\mathrm{SIR}}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$ follows its reference value. In particular, the first rows in Figure 10.8 and Figure 10.9, show the $\mathrm{SIR}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$ and $\widehat{\mathrm{SIR}}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$ evaluated at $\omega_1 = 2\pi f_1$ with $f_1 = 500\,\mathrm{Hz}$. The $\mathrm{NMSE}_{\mathrm{SIR}}^{(1)}$ is equal to $-25.9\,\mathrm{dB}$ for the first VS, while $\mathrm{NMSE}_{\mathrm{SIR}}^{(2)}$ is equal to $-49.5\,\mathrm{dB}$.

In the second row of Figure 10.8 and Figure 10.9, $\widehat{\mathrm{SIR}}_1(\check{\boldsymbol{r}}_v, \omega_2)$ and $\widehat{\mathrm{SIR}}_2(\check{\boldsymbol{r}}_v, \omega_2)$ with $\omega_2 = 2\pi f_2$, $f_2 = 1500\,\mathrm{Hz}$ are depicted for the simulated violin model and the violin virtual replica, respectively. At frequency $f_2 = 1500\,\mathrm{Hz}$, we report $\mathrm{NMSE}_{\mathrm{SIR}}^{(1)} = -22.8\,\mathrm{dB}$ for the first VS and $\mathrm{NMSE}_{\mathrm{SIR}}^{(2)}$ is equal to $-21.5\,\mathrm{dB}$ for the second VS.

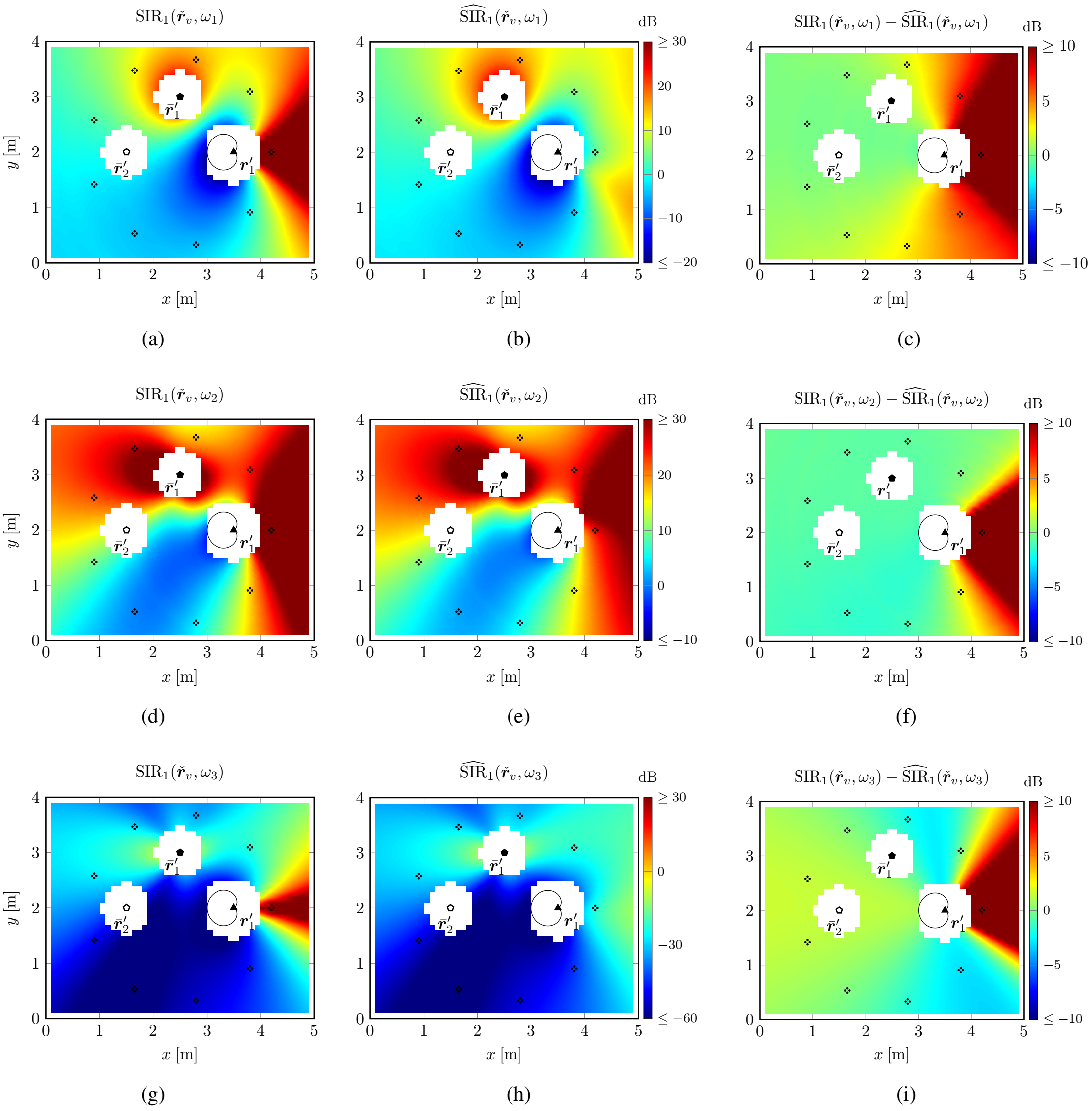
Finally, $\widehat{\mathrm{SIR}}_1(\check{\boldsymbol{r}}_v, \omega_3)$ and $\widehat{\mathrm{SIR}}_2(\check{\boldsymbol{r}}_v, \omega_3)$ with $\omega_3 = 2\pi f_3$, $f_3 = 2500\,\mathrm{Hz}$ are shown in the third rows of Figure 10.8 and Figure 10.9, respectively. As regards the $\mathrm{NMSE}_{\mathrm{SIR}}^{(\bar{n})}$, at frequency $f_3 = 2500\,\mathrm{Hz}$ we obtain $\mathrm{NMSE}_{\mathrm{SIR}}^{(1)} = -29.1\,\mathrm{dB}$ for the first VS and $\mathrm{NMSE}_{\mathrm{SIR}}^{(2)} = -14.9\,\mathrm{dB}$ for the second VS. We recall that every $\mathrm{NMSE}_{\mathrm{SIR}}$ is obtained excluding the locations $\check{\boldsymbol{r}}_v$ for which SIR $\rightarrow \infty$ due to the null in the ideal cardioid pattern.

Overall, we observe a slight overestimation of the $\widehat{\mathrm{SIR}}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$ with respect to its reference value. We interpret this deviation as the effect of small deviations of the real source direct signal estimates provided by the sound field reconstruction technique. Nevertheless, the pattern of the metrics in space is matched for a wide region of the room. As a matter of fact, inspecting $\mathrm{SIR}_{\bar{n}}(\check{\boldsymbol{r}}_v, \omega)$ in Figure 10.8 and Figure 10.9, we immediately identify the influence of the frequency dependent directivity pattern of the sources on the spatial pattern of the metrics. This behavior shows the relevant influence of the directional sound radiation of the sound field and the advantage of the proposed EAR model that includes directivity patterns into the signal model for both sources and VSs.

In conclusion, additional multimedia material[1] regarding simulations of the proposed EAR approach is provided online. In order to show further examples of the EAR simulation, the available additional material comprises different simulations with

[1] https://youtube.com/channel/UC2Vo9hlO283hGWZCOIe6kUw

both stereo and binaural rendering. We adopted the very same setup described in Section 10.2.2, therefore the number of employed microphones and the locations of the source and VSs are fixed. Nonetheless, we employed different source signals taken from [258]. The instruments in [258] are recorded in an anechoic room, and the isolated tracks of each instrument are available. Among the musical pieces of [258], we adopted 30 s extracts of the *RV 315 Opus 8* by *A. Vivaldi*. The recordings include two violins, a viola and a cello, therefore, according to our setup, we discarded the viola track from the set. Multimedia materials provide both *three-degree-of-freedom* (rotation) and *six-degree-of-freedom* (navigation) examples. Both stereo and binaural rendering are available for each simulation. The binaural VM is synthesized by employing HRTFs as pick-up patterns of the VM (see Section 5.3.5). The HRTF adopted in the simulations are taken from the FABIAN dataset [50]. In addition to the EAR signal (including the VSs), the sound field reconstruction (without VSs) is available and for each example the reference sound field rendering is reported for comparison.



(a)

(b)

(c)

(d)

(e)

(f)

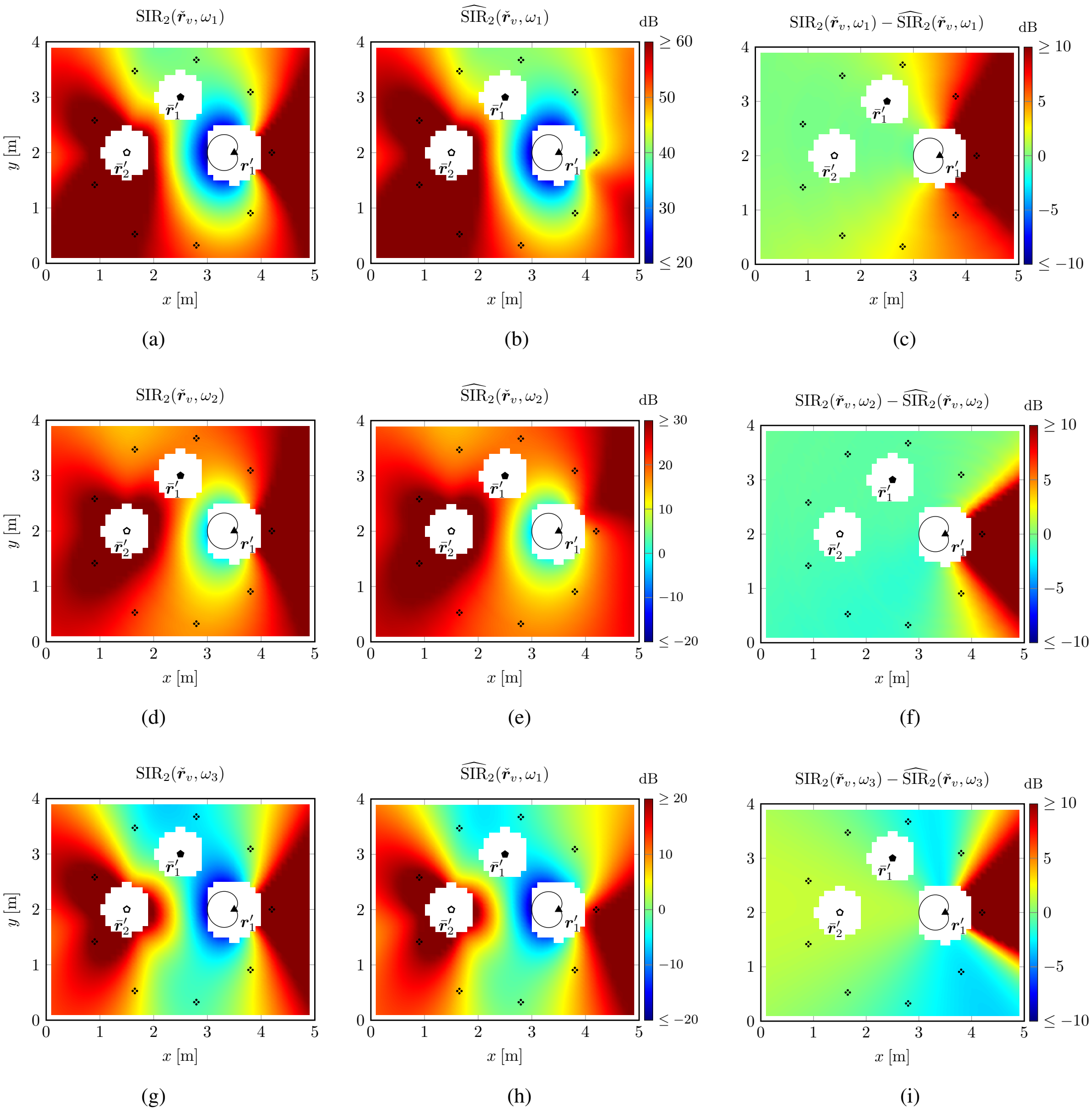
(g)

(h)

(i)

Figure 10.8: *The narrowband* SIR *evaluated with respect to the first VS. The first row is computed at* $\omega_1 = 500\,\mathrm{Hz}$ *and (a), (b), (c) refer to* $\mathrm{SIR}_1(\check{\boldsymbol{r}}_v, \omega_1)$, $\widehat{\mathrm{SIR}}_1(\check{\boldsymbol{r}}_v, \omega_1)$, *and their difference, respectively. Figures in the second row are computed at* $\omega_2 = 1500\,\mathrm{Hz}$ *and (d), (e), (f) refer to* $\mathrm{SIR}_1(\check{\boldsymbol{r}}_v, \omega_1)$, $\widehat{\mathrm{SIR}}_1(\check{\boldsymbol{r}}_v, \omega_1)$, *and their difference, respectively. The third row is evaluated at* $\omega_3 = 2500\,\mathrm{Hz}$ *and (g), (h), (i) refer to* $\mathrm{SIR}_1(\check{\boldsymbol{r}}_v, \omega_1)$, $\widehat{\mathrm{SIR}}_1(\check{\boldsymbol{r}}_v, \omega_1)$, *and their difference, respectively.*

	ω_1	ω_3	ω_3
$\mathrm{NMSE}^{(1)}_{\mathrm{SIR}}$ [dB]	-25.9	-22.8	-29.1



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 10.9: *The narrowband* SIR *computed with respect to second VS. The first row is computed at* $\omega_2 = 500\,\mathrm{Hz}$ *and (a), (b), (c) refer to* $\mathrm{SIR}_2(\check{\boldsymbol{r}}_v, \omega_1)$, $\widehat{\mathrm{SIR}}_2(\check{\boldsymbol{r}}_v, \omega_1)$ *and their difference, respectively. The second row considers* $\omega_2 = 1500\,\mathrm{Hz}$ *and (d), (e), (f) refer to* $\mathrm{SIR}_2(\check{\boldsymbol{r}}_v, \omega_1)$, $\widehat{\mathrm{SIR}}_2(\check{\boldsymbol{r}}_v, \omega_1)$ *and their difference, respectively. The third row is computed at* $\omega_2 = 2500\,\mathrm{Hz}$ *and (g), (h), (i) refer to* $\mathrm{SIR}_2(\check{\boldsymbol{r}}_v, \omega_1)$, $\widehat{\mathrm{SIR}}_2(\check{\boldsymbol{r}}_v, \omega_1)$ *and their difference, respectively.*

	ω_1	ω_3	ω_3
$\mathrm{NMSE}_{\mathrm{SIR}}^{(2)}$ [dB]	-49.5	-21.5	-14.9

CHAPTER 11

Conclusions and Future Developments

This thesis proposed a parametric approach to Extended Audio Reality (EAR), that concerns the interaction between real and virtual acoustic sources (VSs).

In Chapter 5, we proposed a fundamental block for EAR: the sound field reconstruction. A novel parametric model for sound field reconstruction allowed us to represent the acoustic scene with few intuitive parameters.

The sound scene is analyzed by means of distributed compact microphone arrays and the parameters are estimated from their signals. Thanks to the spatially distributed sensors, the proposed model is able to include the directivity of the acoustic sources, providing more accurate results with respect to the omnidirectional radiation model commonly adopted. As a matter of fact, acoustic sources usually present a space dependent sound radiation due to their inherent physical characteristics. This information is relevant for a correct rendering of the sound scene, since the sound field perception varies with the user perspective, e.g., the acoustic field is different in front of or behind a loudspeaker. Moreover, the parametric description of acoustic sources adopted for the sound field reconstruction, provides a suitable model for the implementation of VSs within the EAR framework.

In Chapter 6, we introduced a novel technique for performing multichannel blind source separation (BSS), one of the possible sound field processings that can be performed in the context of EAR.

We adopted a sound field representation, introduced in [37] and known as ray space, as the domain for extracting the source signal through the well-known multichannel nonnegative matrix factorization (MNMF) technique.

The ray space transform (RST) [37] is adopted in order to map the signal acquired by a uniform linear microphone array onto the ray space domain, where the location of acoustic sources is conveniently displayed thanks to the parametrization of the directional components of the sound field as a function of the analysis location. Therefore, we took advantage of the inherent representation of the location of the sources in the ray space, in order to improve the performance of multichannel NMF. Additionally, we introduced a computationally efficient implementation of the RST [37].

In Part III, we discussed the implementation of VSs adopting the violin as a case study. When it comes to the estimation of the parameters required for implementing a violin VS, we can rely on different strategies. We group the approaches according to their invasiveness.

On the one side, we can create VS replicas of actual violins retrieving the parameters from a set of measurements. Thanks to the collaboration with *Museo del Violino* in *Cremona*, we had the possibility of measuring, for the first time, the directivity pattern of a relevant number of valuable historical violins. The measurement methodology and the results are described in Chapter 7. Hence, we can derive VSs that replicate their directional radiation characteristics. Here, we review the violin acoustics that shows the importance of the directional characteristics of the instrument radiation, underlining the need of directional VS models. Therefore, in Chapter 7, we defined a set of tools for the quantitative description of different instruments in terms of their spatial radiation behavior useful for comparing the violins. This made clear evidence of the directional properties of each individual instrument which in principle should be mimicked by VSs.

A less invasive analysis is represented by nearfield acoustic holography (NAH). In Chapter 8 we explored the application of deep learning to the analysis of sound sources through the introduction of a novel data-driven approach to NAH. In particular, we employed a convolutional neural network (CNN) architecture for performing NAH of rectangular and violin plates. The CNN is trained using datasets of synthetic data generated through FEM simulations, in order to estimate the velocity field of the vibrating surface of the object from the acoustic pressure acquired in the proximity of the source. We showed that the proposed CNN is robust against noisy data, sampling position errors and missing data during the training.

On the other side, one can imagine to simulate the vibroacoustic dynamic behavior of a violin by means of Finite Element Method (FEM) simulations and successively, to estimate the VS parameters from the synthetic data. Effective simulations require an accurate model of the violin in terms of geometry and material properties. Hence, in Chapter 9 we approached the development of violin model introducing practical techniques for both the estimation of the sound wave speed in wood and the 3D geometry of violin plates from laser scans.

Finally, in Chapter 10 we provide a first proof of concept simulation of an EAR system. We discussed the requirements for rendering VSs in different EAR scenarios and we implemented a proof-of-concept simulation. More specifically, we simulated a EAR scenario of a string trio. The setup comprised a cello simulated as a sound source actually present in the environment and two violins rendered in the sound scene as VSs. We provided two different violin models: the first was derived by the FEM simulation of a simplified violin body model, while the second is a virtual replica of the prestigious *Il Cremonese* violin by *Antonio Stradivari* derived from the measurements of Chapter 7.

11.1 Future Developments

Future works unfold both sound field processing and virtual source modeling. In the context of sound field reconstruction, we will investigate the implementation of hybrid models that consider both parametric and non-parametric approaches in order to improve the reconstruction performance maintaining the interpretability of parametric techniques. In the light of the promising results obtained with the proposed data-driven NAH, we envision the development of full deep-learning-based sound field reconstruction. We expect that neural networks could potentially perform the sound field reconstruction exploiting the features learned during the training phase from the signals of the sensors. Moreover, we can interpret the sound field reconstruction as a spatial interpolation problem in which we exploit the known signals for estimating the missing information (acoustic field at arbitrary locations). Similar problems have been recently tackled with deep prior approaches in the field of geophysics and computer vision. Deep prior techniques are able to reconstruct missing information in the data e.g., holes in images or unknown seismic data, exploiting the inherent model of the neural network without relying on training data. Therefore, the application of such techniques in the context of sound field reconstruction is appealing due to the implicit difficulties in realizing proper datasets for the training of traditional deep learning architectures. As regards the BSS, we aim at extending the ray-space-based BSS algorithm from the current single array setup, to a distributed arrays setting. The adoption of properly designed ray space parametrization can be exploited in order to merge the multiple views of the acoustic scene given by the distributed sensors in a single representation suitable for the application of BSS algorithms.

As regards the modeling of VSs, the development of effective diffuse sound field rendering is a relevant aspect to investigate. By means of a further sound field analysis, it would be interesting to derive suitable parameters for the computation of the VS diffuse component. This will improve the rendering of VSs in scenarios where no a priori information about the acoustic environment are given allowing a “blind” implementation of VSs in an acoustic scene. Focusing on the simulation of violins, we aim at parameterizing the instrument geometry, in order to provide improved 3D models that can be easily simulated through FEM. Moreover, we envision that the availability of more and better models will allow us to explore data-driven approaches to the analysis and simulation of the instrument.

Finally, the development of computationally efficient implementations of the algorithms is fundamental in order to move towards real-time EAR applications.

APPENDIX $\mathcal{A}$

Demonstration of equality between estimated and actual power of direct and diffuse components

In this appendix we will demonstrate that

$$\begin{aligned} E\{|\hat{X}_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)|^2\} &= E\{|X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)|^2\} \\ E\{|\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)|^2\} &= E\{|X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)|^2\}, \end{aligned} \tag{A.1}$$

under the assumption that the oversubtraction factor $\nu = 1$, the gain floor $G_{\min} = 0$ and that the direct, diffuse and noise power at the ith and at the i'th microphone are equal.

In order to demonstrate the first equality, we make use of the filter definition given in (5.52) and the definition of the CDR in (5.44). It follows that

$$\begin{aligned} E\{|\hat{X}_{n,\mathrm{dir}}(\boldsymbol{r}_i)|^2\} &= |G_{\mathrm{dir}}(\boldsymbol{r}_i)|^2 E\{|U(\boldsymbol{r}_i)|^2\} \\ &= \frac{\mathrm{CDR}(\boldsymbol{r}_i)}{\mathrm{CDR}(\boldsymbol{r}_i)+1} E\left\{\frac{Z\left(\boldsymbol{r}_i\right)+Z\left(\boldsymbol{r}_{i'}\right)}{2}\right\} \\ &= \frac{\Phi_{\mathrm{dir},ii}}{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}} \times \frac{1}{2}[2E\left\{|X_{n,\mathrm{dir}}(\boldsymbol{r}_i)|^2\right\} + \\ 2E\left\{|X_{\mathrm{diff}}(\boldsymbol{r}_i)|^2\right\} &+ 2E\left\{|N(\boldsymbol{r}_i)|^2\right\} - 2E\left\{|N(\boldsymbol{r}_i)|^2\right\}] \\ &= \frac{\Phi_{\mathrm{dir},ii}}{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}} \frac{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}}{1} = E\{|X_{n,\mathrm{dir}}(\boldsymbol{r}_i)|^2\}, \end{aligned} \tag{A.2}$$

where the dependences on the time frame index t and the radial frequency ω have been omitted for the sake of readability.

In order to demonstrate the second equality instead, we make use of the filter definitions given in (5.56) and (5.52) and the definition of the CDR in (5.44). It follows that

$$
\begin{aligned}
E\{|\hat{X}_{\text{diff}}(\boldsymbol{r}_i)|^2\} &= |G_{\text{diff}}(\boldsymbol{r}_i)|^2 E\{|U(\boldsymbol{r}_i)|^2\} \\
&= \left(1 - \frac{\text{CDR}(\boldsymbol{r}_i)}{\text{CDR}(\boldsymbol{r}_i)+1}\right) E\left\{\frac{Z\left(\boldsymbol{r}_i\right) + Z\left(\boldsymbol{r}_{i'}\right)}{2}\right\} \\
&= \frac{\Phi_{\text{diff},ii}}{\Phi_{\text{diff},ii} + \Phi_{\text{dir},ii}} \times \frac{1}{2}[2E\left\{|X_{n,\text{dir}}(\boldsymbol{r}_i)|^2\right\} + \\
2E\left\{|X_{\text{diff}}(\boldsymbol{r}_i)|^2\right\} &+ 2E\left\{|N(\boldsymbol{r}_i)|^2\right\} - 2E\left\{|N(\boldsymbol{r}_i)|^2\right\}] \\
&= \frac{\Phi_{\text{diff},ii}}{\Phi_{\text{diff},ii} + \Phi_{\text{dir},ii}} \frac{\Phi_{\text{diff},ii} + \Phi_{\text{dir},ii}}{1} = E\{|X_{\text{diff}}(\boldsymbol{r}_i)|^2\}.
\end{aligned}
\tag{A.3}
$$

Bibliography

[1] T. D. Abhayapala and A. Gupta. Spherical harmonic analysis of wavefields using multiple circular sensor arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1655–1666, 2009.

[2] T. D. Abhayapala and A. Gupta. Higher order differential-integral microphone arrays. *The Journal of the Acoustical Society of America*, 127:EL227–EL233, May 2010.

[3] T. D. Abhayapala and D. B. Ward. Theory and design of high order sound field microphones using spherical microphone array. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1949. IEEE, 2002.

[4] M. Acerbi, R. Malvermi, M. Pezzoli, F. Antonacci, A. Sarti, and R. Corradi. Interpolation of irregularly sampled frequency response functions using convolutional neural networks. In *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*. IEEE, 2021.

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

[6] J. Ahrens. *Analytic Methods of Sound Field Synthesis*. Springer, Berlin, DE, 2012.

[7] T. Ajdler, L. Sbaiz, and M. Vetterli. The plenacoustic function and its sampling. *IEEE transactions on Signal Processing*, 54(10):3790–3804, 2006.

[8] H. Akima. A new method of interpolation and smooth curve fitting based on local procedures. *J. ACM*, 17(4):589–602, Oct. 1970.

[9] B. Alary and A. Politis. Frequency-dependent directional feedback delay network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 176–180. IEEE, 2020.

[10] B. Alary, A. Politis, S. Schlecht, and V. Välimäki. Directional feedback delay network. *Journal of the Audio Engineering Society*, 67(10):752–762, 2019.

[11] 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.

[12] J. C. Allred and A. Newhouse. Applications of the monte carlo method to architectural acoustics. *The Journal of the Acoustical Society of America*, 30(1):1–3, 1958.

[13] N. Amenta. The crust algorithm for 3 d surface reconstruction. In *Symposium on Computational geometry*, pages 423–424, 1999.

[14] N. Amenta, M. W. Bern, M. Kamvysselis, et al. A new voronoi-based surface reconstruction algorithm. In *Siggraph*, volume 98, pages 415–421, 1998.

[15] I. Amidror. Scattered data interpolation methods for electronic imaging systems: a survey. *Journal of electronic imaging*, 11(ARTICLE):157–76, 2002.

[16] M. E. Anderson and G. E. Trahey. The direct estimation of sound speed using pulse–echo ultrasound. *The Journal of the Acoustical Society of America*, 104(5):3099–3106, 1998.

[17] P. Annibale, J. Filos, P. A. Naylor, and R. Rabenstein. Tdoa-based speed of sound estimation for air temperature and room geometry inference. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):234–246, Feb 2013.

[18] P. Annibale and R. Rabenstein. Closed-form estimation of the speed of propagating waves from time measurements. *Multidimensional Systems and Signal Processing*, 25(2):361–378, 2014.

[19] L. Antani, A. Chandak, M. Taylor, and D. Manocha. Efficient finite-edge diffraction using conservative from-region visibility. *Applied Acoustics*, 73(3):218–233, 2012.

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

[21] F. Antonacci, A. Sarti, and S. Tubaro. Geometric reconstruction of the environment from its response to multiple acoustic emissions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2822–2825. IEEE, 2010.

[22] F. Antonacci, A. Sarti, and S. Tubaro. Two-dimensional beam tracing from visibility diagrams for real-time acoustic rendering. *EURASIP Journal on Advances in Signal Processing*, 2010(1):642316, 2010.

[23] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot. Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1929–1941, 2017.

[24] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot. Joint source localization and dereverberation by sound field interpolation using sparse regularization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6892–6896. IEEE, 2018.

[25] S. Araki, T. Nakatani, H. Sawada, and S. Makino. Stereo source separation and source counting with map estimation with Dirichlet prior considering spatial aliasing problem. In T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, editors, *Independent Component Analysis and Signal Separation*, pages 742–750, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[26] S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Transactions on Signal Processing*, 58(1):121–133, 2010.

[27] M. R. Azimi-Sadjadi, A. Pezeshki, L. L. Scharf, and M. E. Hohil. Wideband doa estimation algorithms for multiple target detection and tracking using unattended acoustic sensors. In *Unattended/Unmanned Ground, Ocean, and Air Sensor Technologies and Applications VI*, volume 5417, pages 1–11. International Society for Optics and Photonics, 2004.

[28] P. K. Banerjee and R. Butterfield. *Boundary element methods in engineering science*, volume 17. McGraw-Hill London, 1981.

[29] N. Barrett and S. Berge. A new method for b-format to binaural transcoding. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.

[30] Bela. *Bela website `https://bela.io`*. Queen Mary University of London, Augmented Instruments Laboratory, 2020.

[31] Y. Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[32] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[33] S. Berge and N. Barrett. High angular resolution planewave expansion. In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May*, pages 6–7, 2010.

[34] R. Berkun, I. Cohen, and J. Benesty. Combined beamformers for robust broadband regularized superdirective beamforming. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(5):877–886, 2015.

[35] F. Bernardini and C. L. Bajaj. Sampling and reconstructing manifolds using alpha-shapes. In *9th Canadian Conference Computational Geometry*. CCCG, 1997.

[36] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend. Binaural reproduction of plane waves with reduced modal order. *Acta Acustica united with Acustica*, 100(5):972–983, 2014.

[37] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro. The ray space transform: A new framework for wave field processing. *IEEE Transactions on Signal Processing*, 64(21):5696–5706, Nov. 2016.

[38] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5):3590–3628, 2019.

[39] G. Bissinger. A0 and a1 coupling, arching, rib height, and f-hole geometry dependence in the 2 degree-of-freedom network model of violin cavity modes. *The Journal of the Acoustical Society of America*, 104(6):3608–3615, 1998.

[40] G. Bissinger. Structural acoustics model of the violin radiativity profile. *The Journal of the Acoustical Society of America*, 124(6):4013–4023, 2008.

[41] G. Bissinger and J. Keiffer. Radiation damping, efficiency, and directivity for violin normal modes below 4 khz. *Acoustics Research Letters Online*, 4(1):7–12, 2003.

[42] J. Bitzer and K. U. Simmer. Superdirective microphone arrays. In *Microphone arrays*, pages 19–38. Springer, 2001.

[43] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2018.

[44] F. Borra, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, S. Tubaro, and A. Sarti. A fast ray space transform for wave field processing using acoustic arrays. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 186–190. IEEE, 2020.

[45] C. Borß. A polygon-based panning method for 3d loudspeaker setups. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[46] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[47] S. Boyd, N. Parikh, and E. Chu. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[48] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, MA, USA, seventh edition, 2009.

[49] P. Brémaud. *Mathematical principles of signal processing: Fourier and wavelet analysis*. Springer Science & Business Media, 2013.

[50] F. Brinkmann, A. Lindau, S. Weinzierl, M. Müller-Trapet, R. Opdam, M. Vorländer, et al. A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations. *Journal of the Audio Engineering Society*, 65(10):841–848, 2017.

[51] H. Buchner, R. Aichner, and W. Kellermann. TRINICON: A versatile framework for multichannel blind signal processing. In *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, volume 3, pages iii–889. IEEE, 2004.

[52] C. Campagnoli, M. Pezzoli, F. Antonacci, and A. Sarti. Vibrational modal shape interpolation through convolutional auto encoder. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 261, pages 5619–5626. Institute of Noise Control Engineering, 2020.

[53] A. Canclini, F. Antonacci, J. Filos, A. Sarti, and P. Naylor. Exact localization of planar acoustic reflectors in three-dimensional geometries. In *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pages 1–4. VDE, 2012.

[54] A. Canclini, F. Antonacci, S. Tubaro, and A. Sarti. A methodology for the robust estimation of the radiation pattern of acoustic sources. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:211–224, 2020.

[55] A. Canclini, L. Mucci, F. Antonacci, A. Sarti, and S. Tubaro. A methodology for estimating the radiation pattern of a violin during the performance. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1546–1550, 2015.

[56] A. Canclini, M. Varini, F. Antonacci, and A. Sarti. Dictionary-based equivalent source method for near-field acoustic holography. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 166–170. IEEE, 2017.

[57] J. J. Carabias-Orti, J. Nikunen, T. Virtanen, and P. Vera-Candeas. Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1512–1527, 2018.

[58] G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin, and R. Gribonval. Near-field acoustic holography using sparse regularization and compressive sampling principles. *The Journal of the Acoustical Society of America*, 132(3):1521–1534, 2012.

[59] G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin, and R. Gribonval. Nachos database and toolbox. `http://echange.inria.fr/nah/`, 2013.

[60] A. Cheng and D. Cheng. Heritage and early history of the boundary element method. *Engineering Analysis with Boundary Elements*, 29:268–302, 03 2005.

[61] F. Chollet et al. Keras. `https://keras.io`, 2015.

[62] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee. A survey of sound source localization methods in wireless acoustic sensor networks. *Wireless Communications and Mobile Computing*, 2017, 2017.

[63] L. Comanducci, F. Borra, P. Bestagini, F. Antonacci, S. Tubaro, and A. Sarti. Source localization using distributed microphones in reverberant environments based on deep learning and ray space transform. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2238–2251, 2020.

[64] COMSOL®. *COMSOL Multiphysics® v. 5.4.* COMSOL®, COMSOL AB, Stockholm, Sweden, 2019.

[65] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

[66] J. Cuenca, F. Gautier, and L. Simon. The image source method for calculating the vibrations of simply supported convex polygonal plates. *Journal of Sound and Vibration*, 322(4):1048 – 1069, 2009.

[67] J. Curtin. Measuring violin sound radiation using an impact hammer. *Journal of Violin Society of America VSA Papers*, XXII(1):186–209, 2009.

[68] L. S. Davis, R. Duraiswami, E. Grassi, N. A. Gumerov, Z. Li, and D. N. Zotkin. High order spatial audio capture and its binaural head-tracked playback over headphones with hrtf cues. In *Audio Engineering Society Convention 119*. Audio Engineering Society, 2005.

[69] C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, New York, NY, USA, 1978.

[70] G. Del Galdo, F. Kuech, M. Prus, and O. Thiergart. Three-dimensional sound field analysis with directional audio coding based on signal adaptive parameter estimators. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.

[71] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki. The diffuse sound field in energetic analysis. *The Journal of the Acoustical Society of America*, 131(3):2141–2151, 2012.

[72] G. Del Galdo, O. Thiergart, T. Weller, and E. A. P. Habets. Generating virtual microphone signals using geometrical information gathered by distributed arrays. In *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pages 185–190. IEEE, 2011.

[73] M. Delcroix, T. Hikichi, and M. Miyoshi. Precise dereverberation using multichannel linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):430–440, 2007.

[74] A. Devices. Adxl326 datasheet. https://www.analog.com/en/products/adxl326.html.

[75] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, Feb 2016.

[76] P. Duhamel and M. Vetterli. Fast Fourier transforms: A tutorial review and a state of the art. *Signal Processing (Elsevier)*, 19(4):259–299, 1990.

[77] A. J. W. Duijndam and M. A. Schonewille. Nonuniform fast Fourier transform. *Geophysics*, 64(2):539–551, 1999.

[78] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[79] N. Q. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010.

[80] R. Duraiswami, Z. Li, D. N. Zotkin, E. Grassi, and N. A. Gumerov. Plane-wave decomposition analysis for spherical microphone arrays. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pages 150–153. IEEE, 2005.

[81] F. Durup and E. V. Jansson. The quest of the violin bridge-hill. *Acta acustica united with acustica*, 91(2):206–213, 2005.

[82] A. Dutt and V. Rokhlin. Fast Fourier transforms for nonequispaced data. *SIAM Journal on Scientific computing*, 14(6):1368–1393, 1993.

[83] A. Dutt and V. Rokhlin. Fast Fourier transforms for nonequispaced data, ii. *Applied and Computational Harmonic Analysis*, 2(1):85–100, 1995.

[84] Y. El Baba, A. Walther, and E. A. P. Habets. Reflector localization based on multiple reflection points. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1458–1462. IEEE, 2016.

[85] Y. El Baba, A. Walther, and E. A. P. Habets. 3d room geometry inference based on room impulse response stacks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5):857–872, 2017.

[86] Y. El Baba, A. Walther, and E. A. P. Habets. Room geometry inference using sources and receivers on a uniform linear array. In *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICSA 2019; 5th International Conference on Spatial Audio*, pages 115–121, Ilmenau, Germany, September 2019.

[87] G. W. Elko. Spatial coherence functions. In M. Brandstein and D. Ward, editors, *Microphone arrays: signal processing techniques and applications*, chapter 4, pages 61–85. Springer-Verlag, New York, NY, USA, 2001.

[88] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala. Sound field separation in a mixed acoustic environment using a sparse array of higher order spherical microphones. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pages 151–155. IEEE, 2017.

[89] C. Faller. Parametric coding of spatial audio. Technical report, EPFL, 2004.

[90] A. Farina. Software implementation of b-format encoding and decoding. In *Audio Engineering Society Convention 104*. Audio Engineering Society, 1998.

[91] A. Farina. Advancements in impulse response measurements by sine sweeps. In *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

[92] X. Feng, Y. Zhang, and J. Glass. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1759–1763. IEEE, 2014.

[93] D. Fernandez Comesana, T. Takeuchi, S. Morales Cervera, and K. R. Holland. Measuring musical instruments directivity patterns with scanning techniques. In *19th International Congress on Sound and Vibration*, pages 1–8. International Institute of Acoustics & Vibration, 2012.

[94] E. Fernandez-Grande, A. Xenaki, and P. Gerstoft. A sparse equivalent source method for near-field acoustic holography. *The Journal of the Acoustical Society of America*, 141(1):532–542, 2017.

[95] J. A. Fessler and B. P. Sutton. Nonuniform fast Fourier transforms using min-max interpolation. *IEEE transactions on signal processing*, 51(2):560–574, 2003.

[96] P. Filippi, D. Habault, J. P. Lefebvre, and Bergassoli. *Acoustics: Basic Physics, Theory and Methods*. Academic Pres, London, UK, 1998.

[97] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. A. Naylor. Localization of planar acoustic reflectors from the combination of linear estimates. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1019–1023. IEEE, 2012.

[98] J. Filos, E. A. P. Habets, and P. A. Naylor. A two-step approach to blindly infer room geometries. In *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC), Tel Aviv, Israel*. Citeseer, 2010.

[99] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[100] D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *IEE conference publication*, volume 511, page 8. Citeseer, 2005.

[101] W. A. Fladung. Windows used for impact testing. In *International Society for Optical Engineering (SPIE)*, pages 1662–1666. International Society for Optical Engineering, 1997.

[102] N. H. Fletcher and T. D. Rossing. *The physics of musical instruments*. Springer Science & Business Media, New York, NY, USA, 2012.

[103] 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*, pages 21–32, 1998.

[104] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, and A. Ngan. A beam tracing method for interactive architectural acoustics. *The Journal of the acoustical society of America*, 115(2):739–756, 2004.

[105] C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[106] D. Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[107] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.*, 21(1):185–194, 1999.

[108] L. Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246. IEEE, 2016.

[109] S. Gonzalez, D. Salvi, D. Baeza, F. Antonacci, and A. Sarti. A data-driven approach to violin making. *arXiv preprint arXiv:2102.04254*, 2021.

[110] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[111] M. Goodwin and J.-M. Jot. Spatial audio scene coding. In *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[112] C. E. Gough. Measurement, modelling and synthesis of violin vibrato sounds. *Acta Acustica united with Acustica*, 91(2):229–240, 2005.

[113] C. E. Gough. A violin shell model: Vibrational modes and acoustics. *The Journal of the Acoustical Society of America*, 137(3):1210–1225, 2015.

[114] C. E. Gough. Violin acoustics. *Acoust. Today*, 12(2):22–30, 2016.

[115] E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley. Single-channel audio source separation using deep neural network ensembles. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[116] D. W. Green, J. E. Winandy, and D. E. Kretschmann. Mechanical properties of wood. wood handbook: wood as an engineering material. *Forest Products Laboratory*, 1999.

[117] T. Grill and J. Schluter. Music boundary detection using neural networks on spectrograms and self-similarity lag matrices. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1296–1300. IEEE, 2015.

[118] R. Gunda, S. Vijayakar, and R. Singh. Method of images for the harmonic response of beams and rectangular plates. *Journal of Sound and Vibration*, 185(5):791 – 808, 1995.

[119] R. Gunda, S. Vijayakar, R. Singh, and J. Farstad. Harmonic green's functions of a semi-infinite plate with clamped or free edges. *The Journal of the Acoustical Society of America*, 103(2):888–899, 1998.

[120] B. Günel, H. Hacihabiboglu, and A. M. Kondoz. Plane wave decomposition with regularization using a single rotating microphone. In *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society, 2007.

[121] E. A. P. Habets. Room impulse response generator. Technical Report 2.4, Technische Universiteit Eindhoven, Tech. Rep, 2006.

[122] P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-Curve. *SIAM Review*, 34(4):561–580, 1992.

[123] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[124] J. Herre, C. Falch, D. Mahane, G. Del Galdo, M. Kallinger, and O. Thiergart. Interactive teleconferencing combining spatial audio object coding and dirac technology. *Journal of the Audio Engineering Society*, 59(12):924–935, 2012.

[125] B. Holm-Rasmussena, H.-M. Lehtonenb, and V. Välimäkib. A new reverberator based on variable sparsity convolution. In *16th Int. Conference on Digital Audio Effects (DAFx-13)*, volume 5, pages 7–8, 2013.

[126] M. T. Isik and O. B. Akan. A three dimensional localization algorithm for underwater acoustic sensor networks. *IEEE Transactions on Wireless Communications*, 8(9):4457–4463, Sep. 2009.

[127] N. Ito and T. Nakatani. Fastmnmf: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 371–375, 2019.

[128] E. Jansson. *Acoustic for violin and guitar makers - Chapter V: Vibration properties of the wood and tuning of violin plates*. KTH Royal Institute of Technology, 2002.

[129] H. Järveläinen and M. Karjalainen. Reverberation modeling using velvet noise. In *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society, 2007.

[130] J.-M. Jot and A. Chaigne. Digital delay networks for designing artificial reverberators. In *Audio Engineering Society Convention 90*. Audio Engineering Society, 1991.

[131] F. Katzberg, R. Mazur, M. Maass, M. Böhme, and A. Mertins. Spatial interpolation of room impulse responses using compressed sensing. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 426–430. IEEE, 2018.

[132] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones. Intrinsic limits of dimensionality and richness in random multipath fields. *IEEE Transactions on Signal processing*, 55(6):2542–2556, 2007.

[133] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[134] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders. *Fundamentals of acoustics*. Wiley, fourth edition edition, December 1999.

[135] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1626–1641, 2016.

[136] P. Kleinschmidt and V. Magori. Ultrasonic remote sensors for noncontact object detection. *Siemens Forschungs und Entwicklungsberichte*, 10:110–118, 1981.

[137] F. Kong, V. Lipari, F. Picetti, P. Bestagini, and S. Tubaro. A deep prior convolutional autoencoder for seismic data interpolation. In *82nd EAGE Annual Conference and Exhibition Workshop Programme*, pages 1–5, 2020.

[138] Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft. Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, 90(1):3–14, 2019.

[139] G. H. Koopmann, L. Song, and J. B. Fahnline. A method for computing acoustic fields based on the principle of wave superposition. *Journal of the Acoustical Society of America*, 86(6):2433–2438, 1989.

[140] J. Kornycky, B. Gunel, and A. Kondoz. Comparison of subjective and objective evaluation methods for audio source separation. *Proceedings of Meetings on Acoustics*, 4(1):050001, 2008.

[141] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets. Parametric spatial sound processing: a flexible and efficient solution to sound scene acquisition, modification, and reproduction. *IEEE Signal Processing Magazine*, 32(2):31–42, 2015.

[142] S. Koyama and L. Daudet. Sparse representation of a spatial sound field in a reverberant environment. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):172–184, 2019.

[143] S. Koyama and H. Saruwatari. Sound field decomposition in reverberant environment using sparse and low-rank signal models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 395–399. IEEE, 2016.

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

[145] H. Kuttruff. *Room acoustics*. CRC Press, 2016.

[146] M.-V. Laitinen and V. Pulkki. Binaural reproduction for directional audio coding. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 337–340. IEEE, 2009.

[147] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, July 2017.

[148] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[149] S. Lee. Review: The use of equivalent source method in computational acoustics. *Journal of Computational Acoustics*, page 1630001, 2016.

[150] S. Lee, S. H. Park, and K. Sung. Beamspace-domain multichannel nonnegative matrix factorization for audio source separation. *IEEE Signal Processing Letters*, 19(1):43–46, 2011.

[151] Q. H. Liu and N. Nguyen. An accurate algorithm for nonuniform fast Fourier transforms (NUFFT's). *IEEE Microwave and guided wave letters*, 8(1):18–20, 1998.

[152] W. Liu and S. Weiss. *Wideband beamforming: concepts and techniques*, volume 17. John Wiley & Sons, 2010.

[153] F. Lluís, P. Martínez-Nuevo, M. Bo Møller, and S. Ewan Shepstone. Sound field reconstruction in rooms: Inpainting meets super-resolution. *The Journal of the Acoustical Society of America*, 148(2):649–659, 2020.

[154] B. Loesch and B. Yang. Source number estimation and clustering for underdetermined blind source separation. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2008.

[155] X. Lu, Y. Tsao, S. Matsuda, and C. Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, volume 2013, pages 436–440, 2013.

[156] M. D. Macleod. Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones. *IEEE Transactions on Signal Processing*, 46(1):141–148, Jan 1998.

[157] E. Maestre and G. Scavone. Creating virtual acoustic replicas of real violins. *International Symposium on Musical Acoustics (ISMA)*, 2019.

[158] D. Marioli, C. Narduzzi, C. Offelli, D. Petri, E. Sardini, and A. Taroni. Digital time-of-flight measurement for ultrasonic sensors. *Transactions on Instrumentation and Measurement*, 41(1):93–97, 1992.

[159] D. Marković, F. Antonacci, L. Bianchi, S. Tubaro, and A. Sarti. Extraction of acoustic sources through the processing of sound field maps in the ray space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2481–2494, 2016.

[160] D. Markovic, F. Antonacci, A. Sarti, and S. Tubaro. Soundfield imaging in the ray space. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2493–2505, 2013.

[161] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro. Multiview soundfield imaging in the projective ray space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6):1054–1067, 2015.

[162] D. Marković, A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro. Visibility-based beam tracing for soundfield rendering. In *2010 IEEE International Workshop on Multimedia Signal Processing*, pages 40–45. IEEE, 2010.

[163] P. A. Martin. *Multiple scattering: interaction of time-harmonic waves with N obstacles*, volume 107. Cambridge University Press, 2006.

[164] MATLAB. *version 9.6.0 (R2019a)*. The MathWorks Inc., Natick, Massachusetts, 2019.

[165] J. D. Maynard, E. G. Williams, and Y. Lee. Nearfield acoustic holography: I. theory of generalized holography and the development of NAH. *The Journal of the Acoustical Society of America*, 78(4):1395–1413, 1985.

[166] I. A. McCowan and H. Bourlard. Microphone array post-filter based on noise field coherence. *IEEE Transactions on Speech and Audio Processing*, 11(6):709–716, 2003.

[167] A. McKeag and D. S. McGrath. Sound field format to binaural decoder with head tracking. In *Audio Engineering Society Convention 6r*. Audio Engineering Society, 1996.

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

[169] J. Meyer. Directivity of the bowed stringed instruments and its effect on orchestral sound in concert halls. *The Journal of the Acoustical Society of America*, 51(6B):1994–2009, 1972.

[170] J. Meyer. *Acoustics and the performance of music: Manual for acousticians, audio engineers, musicians, architects and musical instrument makers*. Springer Science & Business Media, 2009.

[171] J. Meyer and G. Elko. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1781. IEEE, 2002.

[172] Y. Mitsufuji, S. Uhlich, N. Takamune, D. Kitamura, S. Koyama, and H. Saruwatari. Multichannel non-negative matrix factorization using banded spatial covariance matrices in wavenumber domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:49–60, 2020.

[173] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2):145–152, 1988.

[174] A. Moiola, R. Hiptmair, and I. Perugia. Plane wave approximation of homogeneous helmholtz solutions. *Zeitschrift für angewandte Mathematik und Physik*, 62(5):809, 2011.

[175] P. Morse and H. Feshbach. *Methods of theoretical physics*, volume I. McGraw-Hill, New York, NY, USA, 1953.

[176] S. Müller and P. Massarani. Transfer-function measurement with sweeps. *J. Audio Eng. Soc*, 49(6):443–471, 2001.

[177] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[178] A. Nehorai and E. Paldi. Acoustic vector sensor array processing. In *26th Asilomar Conference on Signals, Systems & Computers, (ACSSC)*, pages 192–198. IEEE, 1992.

[179] N. Nguyen and Q. H. Liu. The regular Fourier matrices and nonuniform fast Fourier transforms. *SIAM Journal on Scientific Computing*, 21(1):283–293, 1999.

[180] J. Nikunen and T. Virtanen. Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):727–739, 2014.

[181] M. P. Norton and D. G. Karczub. *Fundamentals of Noise and Vibration Analysis for Engineers*. Cambridge University Press, 2 edition, 2003.

[182] E.-M. Nosal, M. Hodgson, and I. Ashdown. Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms. *The Journal of the Acoustical Society of America*, 116(2):970–980, 2004.

[183] A. A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664, 2016.

[184] A. Nuttall. Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(1):84–91, February 1981.

[185] M. Olivieri, M. Pezzoli, R. Malvermi, F. Antonacci, and A. Sarti. Near-field acoustic holography analysis with convolutional neural networks. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 261, pages 5607–5618, Seoul, Korea, 2020. Institute of Noise Control Engineering.

[186] N. Ono, Z. Koldovskỳ, S. Miyabe, and N. Ito. The 2013 signal separation evaluation campaign. In *2013 IEEE International workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2013.

[187] B. Ottersten, M. Viberg, and T. Kailath. Performance analysis of the total least squares ESPRIT algorithm. *IEEE Transactions on Signal Processing*, 39(5):1122–1135, 1991.

[188] G. Oy. Genelec 8020C. `https://www.genelec.com/previous-models/8020c`, 2020.

[189] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2009.

[190] M. Park and B. Rafaely. Sound-field analysis by plane-wave decomposition using spherical microphone array. *The Journal of the Acoustical Society of America*, 118(5):3094–3103, 2005.

[191] R. M. Parry and I. Essa. Estimating the spatial position of spectral components in audio. In *International Conference on Independent Component Analysis and Signal Separation*, pages 666–673. Springer, 2006.

[192] S. Pasha, J. Donley, and C. Ritz. Blind speaker counting in highly reverberant environments by clustering coherence features. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1684–1687, 2017.

[193] J. Pätynen and T. Lokki. Directivities of symphony orchestra instruments. *Acta Acustica united with Acustica*, 96(1):138–167, 2010.

[194] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris. Source counting in real-time sound source localization using a circular microphone array. In *Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 521–524. IEEE, 2012.

[195] Y. Peled and B. Rafaely. Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field. *IEEE transactions on audio, speech, and language processing*, 21(12):2532–2540, 2013.

[196] A. Pérez Carrillo, J. Bonada, J. Pätynen, and V. Välimäki. Method for measuring violin sound radiation based on bowed glissandi and its application to sound synthesis. *The Journal of the Acoustical Society of America*, 130(2):1020–1029, 2011.

[197] M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, and S. Tubaro. Estimation of the sound field at arbitrary positions in distributed microphone networks based on distributed ray space transform. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 186–190. IEEE, 2018.

[198] M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, and S. Tubaro. Reconstruction of the virtual microphone signal based on the distributed ray space transform. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1537–1541. IEEE, 2018.

[199] M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, and A. Sarti. A parametric approach to virtual miking for sources of arbitrary directivity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2333–2348, 2020.

[200] M. Pezzoli, A. Canclini, F. Antonacci, and A. Sarti. A comparative analysis of the directional sound radiation of historical violins. *The Journal of the Acoustical Society of America*, (submitted).

[201] M. Pezzoli, J. J. Carabias Orti, M. Cobos, F. Antonacci, and A. Sarti. Ray-space-based multichannel nonnegative matrix factorization for audio source separation. *IEEE Signal Processing Letters*, 28:369–373, 2021.

[202] M. Pezzoli, L. Comanducci, J. Waltz, A. Agnello, L. Bondi, A. Canclini, and A. Sarti. A dante powered modular microphone array system. In *Audio Engineering Society Convention 145*. Audio Engineering Society, Oct 2018.

[203] M. Pezzoli, R. R. De Lucia, F. Antonacci, and A. Sarti. Predictive simulation of mechanical behavior from 3D laser scans of violin plates. In *Proceedings of the 23rd International Congress on Acoustics*, Berlin, Germany, Sep 2019. 23rd International Congress on Acoustics, Aachen (Germany), Deutsche Gesellschaft für Akustik.

[204] A. D. Pierce. *Acoustics: an introduction to its physical principles and applications*. Springer, 2019.

[205] F. Pinto and M. Vetterli. Space-time-frequency processing of acoustic wave fields: Theory, algorithms, and applications. *IEEE Transactions on Signal Processing*, 58(9):4608–4620, 2010.

[206] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets. Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[207] M. Poletti. Unified description of ambisonics using real and complex spherical harmonics. In *Ambisonics Symp*, volume 1, pages 2–2, 2009.

[208] A. Politis, J. Vilkamo, and V. Pulkki. Sector-based parametric sound field reproduction in the spherical harmonic domain. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):852–866, 2015.

[209] D. Potts, G. Steidl, and M. Tasche. *Fast Fourier transforms for nonequispaced data: a tutorial*, pages 247–270. Birkhäuser Boston, Boston, MA, 2001.

[210] R. Price and P. E. Green. A communication technique for multipath channels. *Proceedings of the IRE*, 46(3):555–570, March 1958.

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

[212] V. Pulkki, S. Delikaris-Manias, and A. Politis, editors. *Parametric time-frequency domain spatial audio*. Wiley, Hoboken, NJ, USA, first edition edition, 2018.

[213] V. Pulkki, A. Politis, G. Del Galdo, and A. Kuntz. Parametric spatial audio reproduction with higher-order b-format microphone input. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[214] R. Rabenstein and P. Annibale. Acoustic source localization under variable speed of sound conditions. *Wireless Communications and Mobile Computing*, 2017, 2017.

[215] B. Rafaely. Plane-wave decomposition of the sound field on a sphere by spherical convolution. *The Journal of the Acoustical Society of America*, 116(4):2149–2157, 2004.

[216] B. Rafaely. Spatial alignment of acoustic sources based on spherical harmonics radiation analysis. In *4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pages 1–5. IEEE, 2010.

[217] B. Rafaely. *Fundamentals of spherical array processing*, volume 8. Springer, New York, NY, USA, 2015.

[218] B. Rafaely and A. Avni. Interaural cross correlation in a sound field represented by spherical harmonics. *The Journal of the Acoustical Society of America*, 127(2):823–828, 2010.

[219] J. Rahola, F. Belloni, and A. Richter. Modelling of radiation patterns using scalar spherical harmonics with vector coefficients. In *2009 3rd European Conference on Antennas and Propagation*, pages 3361–3365. IEEE, 2009.

[220] E. Ravina, P. Silvestri, P. Montanari, and G. De Vecchi. Spherical mapping of violins. *Journal of the Acoustical Society of America*, 123(5):3659, 2008.

[221] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[222] S. J. Record. *The mechanical properties of wood: including a discussion of the factors affecting the mechanical properties, and methods of timber testing*. J. Wiley & Sons, Incorporated, New York, US, 1914.

[223] L. Remaggi, P. J. Jackson, W. Wang, and J. A. Chambers. A 3d model for room boundary estimation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 514–518. IEEE, 2015.

[224] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[225] R. J. Ross et al. Wood handbook: wood as an engineering material. usda forest service, forest products laboratory. *General Technical Report FPL-GTR-190*, 509(5), 2010.

[226] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4501–4510, Oct 2017.

[227] D. Salvi, S. Gonzalez, F. Antonacci, and A. Sarti. Parametric optimization of violin top plates using machine learning. *arXiv preprint arXiv:2102.07133*, 2021.

[228] P. N. Samarasinghe, T. Abhayapala, and M. Poletti. Wavefield analysis over large areas using distributed higher order microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):647–658, 2014.

[229] P. N. Samarasinghe, T. D. Abhayapala, and M. A. Poletti. 3d spatial soundfield recording over large regions. In *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pages 1–4. VDE, 2012.

[230] P. N. Samarasinghe, H. Chen, A. Fahim, and T. D. Abhayapala. Performance analysis of a planar microphone array for three dimensional soundfield analysis. In *Workshop on Applications of Signal Processing to Audio and Acoustics, (WASPAA)*, pages 249–253. IEEE, 2017.

[231] L. Savioja and U. P. Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015.

[232] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):971–982, 2013.

[233] S. J. Schlecht, B. Alary, V. Välimäki, E. A. P. Habets, et al. Optimized velvet-noise decorrelator. In *Proc. Int. Conf. Digital Audio Effects (DAFx-18), Aveiro, Portugal*, 2018.

[234] S. J. Schlecht and E. A. P. Habets. Sign-agnostic matrix design for spatial artificial reverberation with feedback delay networks. In *Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science*. Audio Engineering Society, 2018.

[235] S. J. Schlecht and E. A. P. Habets. Dense reverberation with delay feedback matrices. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 150–154, 2019.

[236] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[237] R. Scholte, I. Lopez, N. B. Roozen, and H. Nijmeijer. Wavenumber domain regularization for near-field acoustic holography by means of modified filter functions and cut-off and slope iteration. *ACTA Acustica united with Acustica*, 94(3):339–348, 2008.

[238] D. Schröder, P. Dross, and M. Vorländer. A fast reverberation estimator for virtual environments. In *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society, 2007.

[239] M. R. Schroeder. Digital simulation of sound transmission in reverberant spaces. *The Journal of the acoustical society of america*, 47(2A):424–431, 1970.

[240] A. Schuhmacher, J. Hald, K. Rasmussen, and P. Hansen. Sound source reconstruction using inverse boundary element calculations. *The Journal of the Acoustical Society of America*, 113:114–27, 02 2003.

[241] A. Schwarz and W. Kellermann. Coherent-to-diffuse power ratio estimation for dereverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(6):1006–1018, 2015.

[242] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara. Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1, 2020.

[243] A. Semechko. S2 sampling toolbox. `https://github.com/AntonSemechko/S2-Sampling-Toolbox`, 2020.

[244] X. Sheng and Y.-H. Hu. Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *IEEE Transactions on Signal Processing*, 53(1):44–53, 2005.

[245] H. Shiba. Layered model sound speed profile estimation. In *MTS/IEEE OCEANS - Bergen*, pages 1–7. IEEE Oceanic Engineering Society, June 2013.

[246] E. K. Skarsoulis and G. S. Piperakis. Use of acoustic navigation signals for simultaneous localization and sound-speed estimation. *The Journal of the Acoustical Society of America*, 125(3):1384–1393, 2009.

[247] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[248] T. Sreenivas and P. Rao. FFT algorithm for both input and output pruning. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(3):291–292, June 1979.

[249] J. Stautner and M. Puckette. Designing multi-channel reverberators. *Computer Music Journal*, 6(1):52–65, 1982.

[250] G. Steidl. A note on fast Fourier transforms for nonequispaced grids. *Advances in computational mathematics*, 9(3-4):337–352, 1998.

[251] U. M. Stephenson. Quantized pyramidal beam tracing-a new algorithm for room acoustics and noise immission prognosis. *Acta Acustica united with Acustica*, 82(3):517–525, 1996.

[252] P. Stoica and R. Moses. *Spectral Analysis of Signals*. Prentice Hall, Upper Saddle River, NJ, USA, 2004.

[253] P. Stoica and K. C. Sharman. Maximum likelihood methods for direction-of-arrival estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(7):1132–1143, 1990.

[254] N. Sturmel and L. Daudet. Informed source separation using iterative reconstruction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 21:178–185, Jan 2013.

[255] F. Stöter, S. Chakrabarty, B. Edler, and E. A. P. Habets. Countnet: Estimating the number of concurrent speakers using supervised learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):268–282, 2019.

[256] Y. Takida, S. Koyama, and H. Saruwataril. Exterior and interior sound field separation using convex optimization: Comparison of signal models. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2549–2553. IEEE, 2018.

[257] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the /spl beta/-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1592–1605, 2013.

[258] D. Thery and B. Katz. Anechoic audio and 3d-video content database of small ensemble performances for virtual concerts. In *Proceedings of the 23rd International Congress on Acoustics*, Berlin, Germany, Sep 2019. 23rd International Congress on Acoustics, Aachen (Germany), Deutsche Gesellschaft für Akustik.

[259] O. Thiergart, G. Del Galdo, and E. A. P. Habets. On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation. *The Journal of the Acoustical Society of America*, 132(4):2337–2346, 2012.

[260] O. Thiergart, G. Del Galdo, M. Taseska, and E. A. P. Habets. Geometry-based spatial sound acquisition using distributed microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2583–2594, 2013.

[261] C. Tuna, A. Canclini, F. Borra, P. Götz, F. Antonacci, A. Walther, A. Sarti, and E. A. P. Habets. 3d room geometry inference using a linear loudspeaker array and a single microphone. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1729–1744, 2020.

[262] N. Ueno, S. Koyama, and H. Saruwatari. Sound field recording using distributed microphones based on harmonic analysis of infinite order. *IEEE Signal Processing Letters*, 25(1):135–139, 2017.

[263] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[264] E. B. Union. Sound quality assessment material recording for subjective tests. Technical report, European Broadcasting Union, 2008.

[265] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen. Late reverberation synthesis using filtered velvet noise. *Applied Sciences*, 7(5):483, 2017.

[266] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1421–1448, 2012.

[267] C. Van Loan. *Computational frameworks for the fast Fourier transform*, volume 10. Siam, 1992.

[268] H. L. Van Trees. *Optimum array processing*, volume 1. Wiley Online Library, 2002.

[269] S. V. Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, Chirchester, U.K., fourth edition edition, 2008.

[270] W. A. Veronesi and J. D. Maynard. Digital holographic reconstruction of sources with arbitrarily shaped surfaces. *The Journal of the Acoustical Society of America*, 85(2):588–598, Feb. 1989.

[271] M. Vetterli, J. Kovačević, and V. K. Goyal. *Foundations of signal processing*. Cambridge University Press, 2014.

[272] J. Vilkamo, T. Lokki, and V. Pulkki. Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *Journal of the Audio Engineering Society*, 57(9):709–724, 2009.

[273] L. Villa, M. Pezzoli, F. Antonacci, and A. Sarti. A methodology for the estimation of propagation speed of longitudinal waves in tone wood. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 66–70. IEEE, 2020.

[274] E. Vincent. *Contributions to audio source separation and content description*. Habilitation à diriger des recherches, Université Rennes 1, Nov. 2012.

[275] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1669, July 2006.

[276] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[277] O. Walter, L. Drude, and R. Haeb-Umbach. Source counting in speech mixtures by nonparametric bayesian estimation of an infinite gaussian mixture model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 459–463. IEEE, 2015.

[278] L. M. Wang and C. B. Burroughs. Acoustic radiation from bowed violins. *The journal of the acoustical society of america*, 110(1):543–555, 2001.

[279] Z.-Q. Wang and D. Wang. Combining spectral and spatial features for deep learning based blind speaker separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):457–468, 2018.

[280] A. F. Ware. Fast approximate Fourier transforms for irregularly spaced data. *SIAM review*, 40(4):838–856, 1998.

[281] L. M. website. *http://www.lucchimeter.com*. LucchiCremona, 2020.

[282] U. G. K. Wegst. Wood for sound. *American Journal of Botany*, 93(10):1439–1448, 2006.

[283] G. Weinreich. Sound hole sum rule and the dipole moment of the violin. *The Journal of the Acoustical Society of America*, 77(2):710–718, 1985.

[284] G. Weinreich. Directional tone color. *The Journal of the Acoustical Society of America*, 101(4):2338–2346, 1997.

[285] G. Weinreich and E. B. Arnold. Method for measuring acoustic radiation fields. *The Journal of the Acoustical Society of America*, 68(2):404–411, 1980.

[286] E. G. Williams. *Fourier Acoustics*. Academic Press, London, UK, 1999.

[287] E. G. Williams. Regularization methods for near-field acoustical holography. *The Journal of the Acoustical Society of America*, 110(4):1976–1988, 2001.

[288] E. G. Williams, J. D. Maynard, and E. Skudrzyk. Sound source reconstructions using a microphone array. *The Journal of the Acoustical Society of America*, 68(1):340–344, 1980.

[289] J. Woodhouse. The acoustics of the violin: a review. *Reports on Progress in Physics*, 77(11):115901, 2014.

[290] J. Woodhouse and R. Langley. Interpreting the input admittance of violins and guitars. *Acta Acustica united with Acustica*, 98(4):611–628, 2012.

[291] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y.-F. Chen, J. Li, and B. Firner. Crowd++: Unsupervised speaker count with smartphones. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp 13, pages 43–52, New York, NY, USA, 2013. Association for Computing Machinery.

[292] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem. Optimal modal beamforming for spherical microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):361–371, Feb 2011.

[293] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[294] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, volume 5, pages 2578–258. IEEE, 1988.

[295] Y.-B. Zhang, F. Jacobsen, C.-X. Bi, and X.-Z. Chen. Near field acoustic holography based on the equivalent source method and pressure-velocity transducers. *The Journal of the Acoustical Society of America*, 126(3):1257–1263, 2009.

[296] M. D. Zoltowski. Beamspace root-music. *IEEE Trans. Signal Processing*, 41(1):344–364, 1993.