POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A Deep Learning-based method for Multi-Zone Sound Field Synthesis

Tesi di Laurea Magistrale in
Music Engineering - Music and Acoustic Engineering

Author: **Roberto Alessandri**

Student ID: 964993
Advisor: Prof. Fabio Antonacci
Co-advisors: Luca Comanducci
Academic Year: 2021-22

Abstract

Multi-zone sound field synthesis is a branch of sound field synthesis in which we reproduce different pressure fields inside multiple regions. Is a complex and challenging problem in acoustic signal processing that is increasingly being required to be addressed. In this thesis, we propose a technique for multi-zone sound field synthesis based on a deep neural network. Most of the nowadays approaches focus on the reproduction of a desired sound field in a target *bright* region, while attenuating the acoustic potential energy in a second target *dark* region. One of the main issues of these methods is in their ability to accurately reproduce the bright zone, without failing on attenuating the second region. Acoustic Contrast Control has been demonstrated to be the best-performing technique in terms of attenuation of the dark zone, at the cost of a high error in the bright region. In the proposed technique we synthesise the estimated pressure field through a Uniform Linear Array of loudspeakers and follow the Pressure Matching and Amplitude Matching approaches, in which the driving signals to reproduce a sound field are retrieved by minimising the reproduction error at a discrete set of control points. Following deep learning's recent widespread adoption in the acoustic signal processing field we perform the minimisation by applying an encoder-decoder-structured Convolutional Neural Network. Through simulations and numerical experiments, we compare the performance of the aforementioned methods with the proposed technique and demonstrate how the latter can overcome the trade-off between the accuracy of the reproduction in the bright zone and the acoustic contrast between the two target regions.

Keywords: multizone soundfield synthesis, personal audio, pressure matching, deep learning, convolutional neural network

Abstract in lingua italiana

La riproduzione multizona del campo sonoro è un branca della sintesi dei campi sonori che si occupa di riprodurre diversi campi di pressione in regioni multiple dello spazio. È un problema complesso e impegnativo nell'elaborazione dei segnali acustici che sta diventando sempre più necessario da affrontare. In questa tesi, proponiamo una tecnica per la sintesi del campo sonoro multi-zona basata su una rete neurale profonda. La maggior parte degli approcci attuali si focalizzano sulla riproduzione di un campo sonoro desiderato in una regione *luminosa*, mentre attenuano l'energia potenziale acustica in una seconda regione *scura*. Uno dei problemi principali di questi metodi sta nella loro capacità di riprodurre con precisione la zona luminosa, senza mancare di attenuare la seconda regione. Acoustic Contrast Control ha dimostrato di essere la tecnica più performante in termini di attenuazione della zona scura, al costo di un errore elevato nella regione luminosa. Nella tecnica proposta sintetizziamo il campo di pressione stimato attraverso una serie lineare uniforme di altoparlanti e seguiamo gli approcci proposti nei metodi di Pressure Matching e Amplitude Matching, in cui i segnali di azionamento per riprodurre un campo sonoro sono ottenuti minimizzando l'errore di riproduzione in un insieme discreto di punti di controllo. In seguito alla recente adozione diffusa dell'apprendimento profondo nell'elaborazione del segnale acustico, eseguiamo la minimizzazione applicando una rete neurale convoluzionale basata sulla struttura encoder-decoder. Attraverso simulazioni ed esperimenti numerici, confrontiamo le prestazioni dei suddetti metodi con la tecnica proposta e dimostriamo come quest'ultima riesca superare il compromesso tra la precisione della riproduzione nella zona luminosa e il contrasto acustico tra le due regioni obiettivo.

Parole chiave: Sintesi acustica multizona, audio personale, Pressure Matching, apprendimento profondo, rete neurale convoluzionale

Contents

Introduction

Sound Field Synthesis (SFS) methods use multiple loudspeakers (secondary sources) to synthesize a desired pressure field in a target region of space. SFS has been an active research field in acoustics for several decades, with applications in immersive, virtual and augmented reality, telepresence, gaming, noise cancellation, and personal sound zone generation.

SFS methods can be classified into two categories. Classical approaches, such as Wave Field Synthesis and Ambisonics, are based on analytic methods derived from the Helmholtz equation and assume large continuous distributions of loudspeakers. Wave Field Synthesis (WFS) [9, 54] reproduces the desired sound field by leveraging on the Huygens principle and on large numbers of regularly displaced loudspeakers. Ambisonics [4, 24, 26, 60] is based on the analysis of the sound field in terms of first-order spherical harmonics and reproduces accurately the pressure field in a limited region of space denoted *sweet spot*. To achieve larger listening areas High Order Ambisonics (HOA) [48] considers also higher order spherical harmonics.
The other class of methods uses optimisation-based techniques that mathematically minimize the error between the synthesised and desired sound fields in a target region of space. Pressure Matching (PM) and (weighted) Mode Matching (MM) are examples of optimisation-based SFS methods. Pressure Matching techniques [32, 37, 44, 47] are based on the minimisation of the reproduction error at a fixed number of positions in the listening area, denoted as *control points* (CPs). Given the desired pressure in the CPs and the transfer functions - i.e. the acoustic functions that describe how sound propagates between a speaker and a microphone - the driving signal is obtained through regularised least squares. PM it's widely used in practice because of its simple implementation and flexibility on secondary sources placement [33]. Mode matching techniques [10, 35, 52] aim at minimising the difference between the modes of the wavefunction used to expand the sound field of the desired and reproduced pressure field at a single control point.
Since numerical-optimisation-based methods enable us to generate complex sound fields with a flexible array geometry of loudspeakers, they have a broad range of practical applications.

In personal audio applications [11], it is often necessary to synthesise different sound fields inside multiple regions. This problem is knwon as the *multizone sound field synthesis* (MZ-SFS) problem [29, 56, 61, 64]. For example, an MZ-SFS problem could be the synthesis of plane waves with different propagation angles inside two regions. However, it is sometimes difficult to achieve accurate synthesis in such a control problem because its physical feasibility becomes significantly low depending on the desired propagation angles of the plane waves and the chosen system layout. Furthermore, in some applications, it is desired to generate sound fields of certain acoustic power levels, e.g. it could be necessary to have the acoustic power distribution high in one region and suppressed in another region [11]. In previous studies, this kind of MZ-SFS problem has been addressed by the Acoustic Contrast Control (ACC) and Amplitude Matching (AM) approaches.
ACC [15, 17], aims to maximise the ratio between the acoustic potential energy between two regions. However, the synthesised power distribution by ACC cannot be guaranteed to be flat inside the target region, as only the total energy is taken into account. Another drawback of ACC is that it can't be applied to generating sound fields in more than two regions at different acoustic potential energy. Furthermore, it completely discards any kind of information about the two desired sound fields; it only focuses on maximising the acoustic contrast by just considering the transfer functions between loudspeakers and control points in the two regions.
Amplitude Matching [1, 34], following the same rationale of PM, is based on a numerical optimisation that aims to minimise the error between amplitude distributions of synthesised and desired sound fields. Being a non-linear optimisation problem it is necessary to adopt a non-linear optimisation algorithm to obtain the optimal solution. However, to be able to represent a desired sound field over the target region, and not only its power distribution, the algorithm is also usually initialised with the optimal solution of the PM method, as otherwise the phase information would be completely discarded. PM-initialised AM is slightly less accurate in the reproduction w.r.t. PM but is capable of achieving a larger improvement in terms of acoustic contrast. However, it doesn't reach the upper bound of ACC in terms of acoustic contrast.

More recently, following its widespread adoption in the acoustic signal processing field [14], deep learning techniques have also been applied to sound field synthesis. In [43] mono audio recordings are converted into First-Order Ambisonics (FOA) signals by taking advantage of a 360° video camera through a NN. In [25] HOA encoding process' frequency is expanded through a learning-based model while in [50] Ambisonics signals are upscaled through a NN. Also in [63] spherical harmonic coefficients in sound field recording are estimated using feed-forward NN. In [46] the optimal number of driving signals are calculated

through a NN. In [39] has been proposed a method to localise sound sources' Direction Of Arrival (DOA) using a neural network and two differential microphone arrays. A neural network is also been applied in [16] for the inference of a Room Impulse Response (RIR) to an audio signal, by just using photos of the environment. Deep Learning methods were demonstrated to be successful also in tasks such as the characterisation of reverberant environments [18] and echo cancellation [38]. Recently, it was proposed a technique for sound field synthesis using irregular arrays, that is based on the compensation of driving signals [20]. All the above studies demonstrate how methods based on Deep Learning can outperform classical methods. The only drawback of these techniques is that there is a need of dealing with a large amount of data, real or simulated.

Recently in [51] a multi-zone environment in harsh conditions has been synthesised using a neural network. Their method can achieve a high acoustic contrast while keeping at low values the reproduction error. However, to be able to counter the harsh conditions, in their layout they surrounded with a Uniform Circular Array the bright zone with secondary sources. It's an effective technique, but in most practical scenarios the target regions are on the same side w.r.t. the loudspeaker distribution, i.e. both zones are inside the UCA or both zones are outside the UCA. In their setup, most of the loudspeakers are used in order to reproduce the desired pressure fields, while only the secondary sources near the quiet zone focus on generating disruptive interferences.
Our approach will follow the procedure shown in [19], in which they estimate the driving signals directly through a Convolutional Neural Network from the sound field measured at a set of control points.
Encouraged by their results, in this manuscript we propose a Deep Learning-based Pressure Matching technique for the synthesis of Multi-Zones (MZ-DLPM), i.e. we set two different target zones with different acoustic potential energy. Another main difference from their work is in the layout. We used a Uniform Linear Array (ULA) instead of surrounding the controlled environment with a UCA. The main reason behind this choice is that ULAs are already commercially available as soundbars, while UCA's set nowadays are mainly used for research purposes. Since no ground truth for the desired driving signals is available, the network is optimised by computing the loss between the desired and estimated sound field, obtained by convolving the estimated driving signals with the appropriate Green's function. Through simulations, we compare the performances of the proposed technique with AC, the original PM approach and its AM variant, and find out that the proposed method can overcome the trade-off between accuracy and acoustic contrast.

The rest of this thesis is organised as follows. In *Chapter 1* prior works on sound field

synthesis methods are briefly summarised. We provide a list of the main SFS techniques used over the years, addressing their advantages and drawbacks. We will also stress how the research is focusing on the application of optimisation-based methods when it comes to MZ-SFS. *Chapter 2* is dedicated to the coverage of the theoretical background used throughout the thesis. We introduce the main concepts of acoustics applied in the optimisation-based algorithms and then we make a brief introduction to Deep Learning techniques, focusing on Convolutional Neural Networks. The problem statement on MZ-DLPM is described in *Chapter 3* along with the proposed algorithm implementation. Experimental results are presented and discussed in *Chapter 4*. We provide simulation and evaluation results to validate our method, analysing how it performs when compared with the state-of-the-art methods. Finally, *Chapter 5* concludes this thesis and proposes future works.

1 | State of the Art

This chapter reports the most recent and successful techniques for Sound Field Synthesis (SFS), with a particular focus on their applications to the creation of Personal Sound Zones (PSZ). We talk about SFS when we aim to use a superposition of sound fields emitted by a combination of elementary sound sources to create a sound field with desired properties over a given area. While we refer to PSZs sound field synthesis, multi-zone sound field synthesis or sound field control (SFC) when the objective is to create multiple sound zones with different sound content inside the same acoustic environment using an array of loudspeakers.

We will use the term *secondary sources* to define these elementary sources. Loudspeakers are used when it comes to real-life experimentations. The term secondary source represents the fact that such a sound source is not the *primary* source of the auditory event which is desired to be evoked in the listener. Another expression that will be frequently used in this thesis is *virtual sources*. We refer to a virtual source when we have the sound field generated by the source, but not the source itself. Therefore, to easily describe a relation between secondary and virtual sources, we can say that secondary sources, are the emitting sources used to synthesize the soundfield that would be generated by the virtual sources if they were present, as described in figure 1.1.

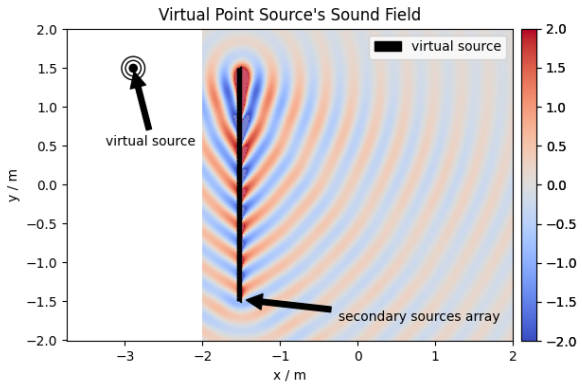


Figure 1.1: Secondary sources synthesising a sound field emitted by a virtual source.

In the following, we will first give an overview to the approaches that rely on a physical model to obtain the desired sound field.
In our context, physical models are used to describe the spatial characteristics of an object (e.g., a sound source). This includes object's location and radiation properties. For example, with a physical model, a virtual source can be defined using an appropriate coordinate system, specifying both its location and directionality. In SFS, the model is used to define an acoustic scene and properly describe its properties.

In our investigation we adopt a physical model that represents sounds fields using the Single Layer Potential (SLP), as introduced in [2]. This representation, relies only on two key factors, a spatio-temporal impulse and a driving function: here spatio-temporal impulses are to be considered as the emission of a spherical or circular wave in a free field (i.e. a field free from reflections); while the driving functions, sometimes called *densities* can be defined as the weights to be applied to each source to render the desired sound field.
Within this representation, spatio-temporal impulses are easily describable. What techniques that rely on the SLP focus on is to obtain the appropriate driving function. Many of the methods that will be described, are based on this representation.
Specifically, we will introduce Vector-Based Amplitude Panning (VBAP) [49], Arbitrary Order Ambisonics [24], Wave Field Synthesis (WFS) [9] and Plenacoustic [12] rendering in this sense.

Then, we will describe how the research is trying to be independent from physical models, with methods for which there is parameters optimisation. For this purpose, Pressure Matching (PM) [44], Mode-Matching (MM) [10, 48, 58] and Deep Learning-based methods are presented and analysed, and the main advantages of not relying on physical models are shown.

Ultimately, Multizone Sound Field methods are introduced, as they are the main focus of this research. It is shown how the most used methods behave in this context and which are their advantages and disadvantages. As it will emerge from this chapter, when it comes to PSZs, optimisation methods are used.
In recent years, following the success achieved in other topics concerning acoustics such as speech recognition [40], characterisation of reverberant environments [18], Room Impulse Response (RIR) inference [16], echo cancellation [38] and Direction of Arrival (DOA) estimation [39], a DL approach have been proposed to address the aforementioned task in a harsh environment [51].
The success of Neural Networks (NN) in all these fields and in particular of Convolutional Neural Networks (CNNs) is related to their ability to find patterns and local correlation in

their inputs: this characteristic has been particularly successfull when using spectrograms of speech signals in input [6], since there are local correlation both in time and frequency [7]. Due to their spatial correlation also Sound Fields are appropriate to be used as inputs to CNNs. Furthermore, NNs have the ability to learn non-linear behaviours, and this makes it convenient to use them to learn driving functions, that are non-linear.
However, to the best of our knowledge, not much research has been done with NNs, and for all of the reasons described above, our work tries to fit in this is gap in the literature. The work done in [51] focus on rendering of the sound field in a harsh environment, and to achieve this result in their layout they've surrounded the bright zone with secondary sources.
Our approach tries to synthesize a Multizone environment in an ideal anechoic environment and without surrounding the bright zone. For our purpose, we used a ULA and we placed the two zones one the same side with respect to the secondary sources. Despite they explored the ability to learn non-linearities of NNs by using a Multi-Layer Perceptron, in their system the local spatial correlation of Sound Fields are not investigated. For our work, we considered more appropriate to use a CNN, since it's able to recognize local correlations of its input.

1.1. Sound Field Synthesis

In recent times, audiophiles and researchers have been interested in what is called "spatial audio". The goal of spatial audio techniques is to recreate sound and timbre qualities of acoustic scenes, along with the position, orientation of sound sources, and shape and characteristics of the environment. In this sense, complex techniques have been developed that can evoke the impression of being "immersed" in a reproduced acoustic scene.

These representations are categorized as model-based or data-driven depending on whether the acoustic scene has been characterized through a physical model as stated above or whether displayed objects already contain all the spatial information. With the second category the system used has not only to extract the driving signals but also the spatial information of objects present in the environment, e.g. sound sources and microphones. Most of the following methods have been developed in both data-driven and model-based directions. For our scope, we concentrate in the following description in the model-based scenarios as the technique proposed in this work rely on a physical model.

1.1.1. Model-based Methods

We talk of model-based SFS methods when we use a physical model to describe all the information about virtual or secondary sources. Typically, the type of information encoded in the model refers to the location of the object and the type of radiation, which in most straightforward cases are plane or spherical waves.

The simplest known method with wide commercial use is Vector-Based Amplitude Panning (VBAP) [49]. Vector-Based Amplitude Panning is the evolution of the standard Stereophony panning, from two to multiple channels. In its simplest form, i.e. in a 2D horizontal plane reproduction case, behaves as a pair-wise Stereophony panning technique: each phantom source[1] is reproduced only by the two speakers closest to its position. VBAP can also be used to reproduce sources in a 3D space using three speakers for each source instead of one, where the third loudspeaker is not on the same plane of the other two, as shown in Fig 1.2.

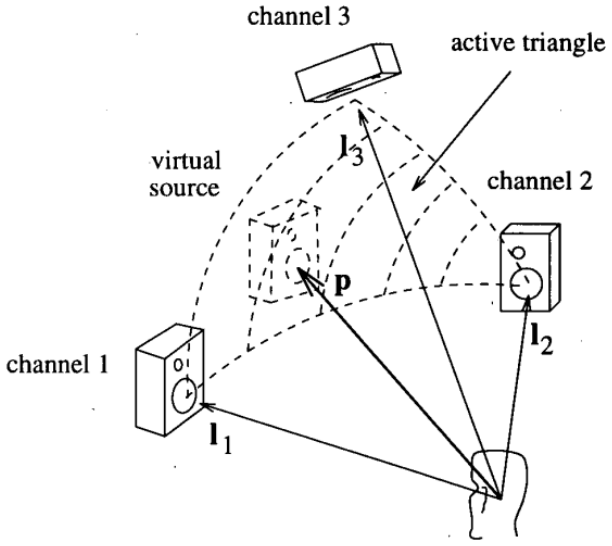


Figure 1.2: Configuration for 3D VBAP. The phantom source can be placed in the triangle formed by the loudspeakers. Image taken from [49].

The techniques based on this concept, however, are not flexible. Firstly, are able to reproduce the desired environment only in the so-called *sweet spot*, that is in the position equidistant from all the secondary sources; as the listener moves away from this sweet spot, each phantom source will be perceived as coming from the closest loudspeaker - for each couple of secondary sources - to the listener. Secondly, regarding reproduction, the

[1] Is inappropriate to talk about virtual sources in the context of Amplitude Panning as the sound field created by two loudspeakers is generally very different from that of a real source[3]

same loudspeaker layout must be set up to reproduce a specific environment, and this layout must have a regular geometry. At last, is not possible to render phantom sources in position closer to the surface occupied by the loudspeakers, only in between and/or farther. From the above description we can deduce that is not fitting to talk about sound field synthesis when we talk about VBAP, because even if the result obtained is similar, its goal is to place sources in the environment and not to render a specific sound field. Still it was necessary to describe it, since when it comes to spatialisation, is the most widespread configuration both in houses and theaters due to its simplicity, computational efficiency, and the ability to achieve good results for stationary listeners.

Another classic technique for which much research has been done, is Ambisonics [24]. The authors of this technique, noticed that in those years (70's) there wasn't any system capable of encoding (and also decoding) the directionality of a sound and in particular the information concerning the relationship between a direct sound and its early-and-late reflections [24]. Ambisonics was born as a data-driven technique, and in fact its focus was to be able to reconstruct a defined ambience after an acquisition with specifically designed microphones, as shown in Fig 1.3.

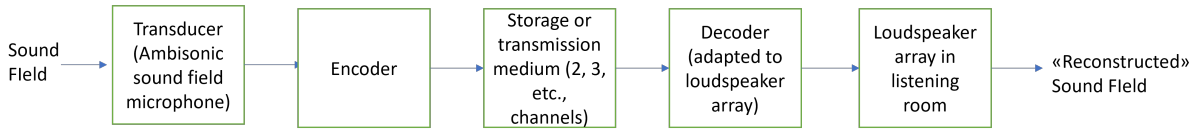


Figure 1.3: Basic features of ambisonic reproduction. As described in [24]

Since it relies on a characterization of the sound field in terms of spherical (or circular) harmonics, this technique can be used to reconstruct a virtual venue described through a physical model. The main difference with standard panning methods is that for the reproduction of the sound field, there isn't a direct one-to-one correlation channel-to-speaker, but in each channel are encoded information about physical properties of the acoustic field. From this we can derive that the sound field is represented by a set of signals that is independent from the setup of loudspeakers. In fact, in Ambisonics systems the number of channels is lower than the number of loudspeakers used to reproduce the sound field. From this characteristic it follows that not only secondary sources closest to the desired source direction are used, but all loudspeakers contribute to the rendering of each virtual source, as shown in Fig 1.4.

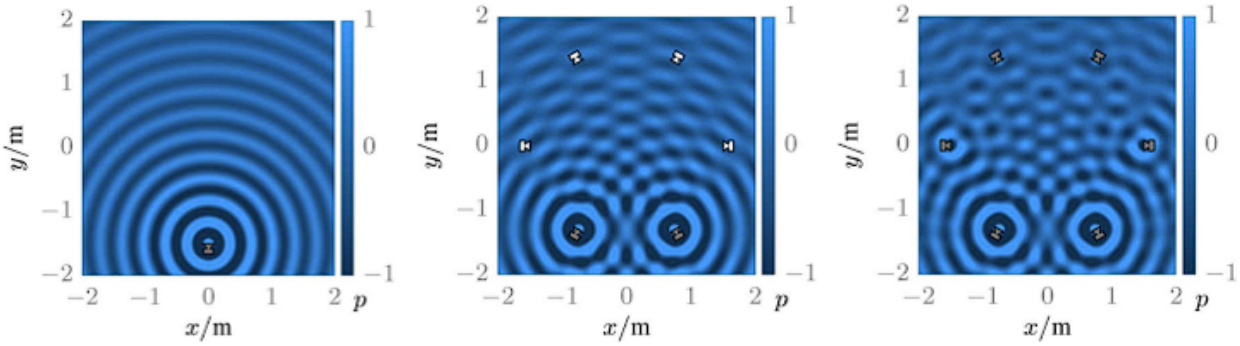


Figure 1.4: The colored speakers denote the active speakers. (a) The virtual source constitutes a point source. (b) Stereo amplitude panning of virtual source. (c) Ambisonic amplitude panning of virtual source. Image from [55]

Even though a sound field at any point can be characterised with a spherical harmonics expansion [59], is also true that there is a direct correlation between the number of orders n of the expansion used to describe the sound field and the number of secondary sources necessary to decode the information, i.e. $(n+1) * *2$. For this reason, in practical implementations, the order of the expansion is truncated. And depending on the order used we talk about Ambisonics if we use just the $0th$ and $1st$ orders or Higher Order Ambisonics, if also higher orders are used. Thankfully zeroth and first order spherical harmonics already contain respectively information about pressure field and the three components of particle velocity. Higher orders, represent non-redundant combinations of higher gradients.

The main idea behind this method is to use a set of loudspeakers to reproduce a desired sound field in a specific region inside the volume delimited by the speakers, assuming that the secondary sources are at an infinite distance. A consequence of this last assumption is that in standard Ambisonics systems, loudspeakers are modeled as plane waves emitters. This is a model mismatch leads to artefacts since actual loudspeakers behave similarly to point sources [3].

In Ambisonics, depending the order to which the spherical harmonic expansion is truncated, there is a strong correlation between the radius of the listening area and the maximum frequency that can be correctly reproduced. This correlation is expressed in the Relative Truncation Error Bound Theorem [30], from which we can deduce that in first-order Ambisonics to have a listening area with a radius of 1 meter, the maximum reproducible frequency without artefacts - to be precise, with a reproduction error upper

bounded to 0.16127 - is around 40 Hz; vice-versa if our goal is to reproduce a sound wave at high frequency like for example 7 kHz, the radius of the listening area cannot exceed 6 millimetres. These values, are to be considered valid just for First-Order Ambisonics. In fact, the main advantage of HOA is that as we increase the number of the order, these constraints are relaxed.

The major advantages of this method are its low complexity and the low minimum number of speakers needed to reproduce the field, which is only four.

Many variants address the minor reproduction zone issue; Near-field Compensated Higher Order Ambisonics (NFC-HOA) is the most successful [22]. The *near-field* term means that the secondary sources are assumed to be at a finite distance, i.e. monopoles. While the *higher-order* term, as described above, is connected to the mathematical model used for Ambisonics. For NFC-HOA, only circular or spherical secondary sources distribution are used since these are the only distributions for which there is a closed-form solution of the driving function.

Wave Field Synthesis (WFS) is one of the best-known methods for sound field reproduction [9]. One of the main differences with the NFC-HOA approach lies in the different choice of mathematical model. This method is based on Huygens's Principle, which states that any point on a propagating wavefront can be a point source for producing spherical secondary waves, as shown in Fig 1.5 .

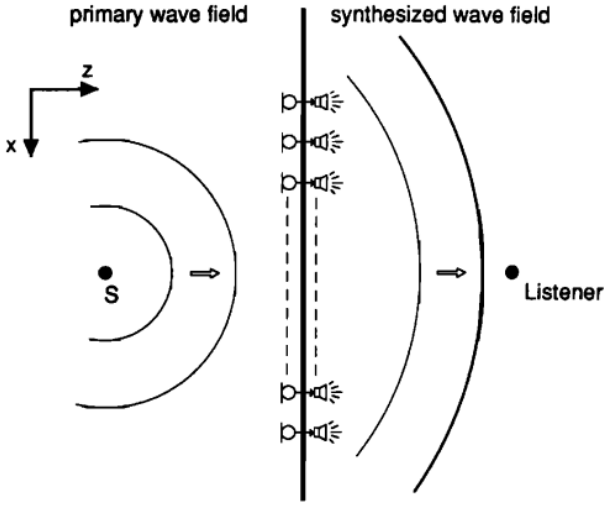


Figure 1.5: Simple WFS implementation following Huygens Principle. Image taken from [9]

One of the essential advantages of WFS with respect to NFC-HOA is its independence from secondary source distribution. It is being demonstrated to work with planar, linear,

spherical, circular and irregular distributions. Another significant advantage is that the reproduction region is the whole area delimited by the speakers. However, in WFS is not possible to enlarge the area bounded by the loudspeakers without increasing their number; for NFC-HOA systems, it does not matter what the size of the bounded area is, the number of secondary sources needed to reproduce the sound field in the desired region does not change.

Mathematical models give an optimal solution for a continuous distribution of secondary sources in all the SFS techniques. It is essential to underline that this is an ideal case and that we must deal with a discrete and limited distribution of loudspeakers in practical cases. With this constraint, solutions provided by these methods are valid up to a specific frequency, depending on the distance between loudspeakers. Once we go higher this frequency, artefacts arise. This phenomenon, called *spatial aliasing*, leads to different effects depending on the method. For example, in NFC-HOA, there is a considerable reduction of the maximum listening area dimensions, while in the WFS, the artefacts are distributed homogeneously in the listening area. Is noteworthy that spatial aliasing is one of the major drawbacks when it comes to use irregular set-ups for WFS [21].

The spacing between secondary sources is the last consideration regarding the differences between the two model-based methods. While for WFS, there is a straight correlation with the minimum wavelength of the reproduced field and the spacing between secondary sources, in NFC-HOA is also essential to consider the ratio between the region bounded by loudspeakers and the maximum reproduction area; as a rule of thumb, once this ratio is above 1.4, NFC-HOA systems are more flexible in terms of loudspeakers spacing w.r.t. WFS.

A last and more recent technique is the Plenacoustic rendering method [12, 13]. This technique is drawn from on the Plenacoustic Function (PAF) [5], which was introduced to determine the whole sound field inside the studied volume by measuring it in specific control points. This method was further improved in order to be able to give also directional information.

As described in [5] the field is modelled in terms of acoustic rays since has been demonstrated to be more efficient than previous models when it comes to the description of complex fields, even though simple fields may lack accuracy. As control points are used a Uniform Linear Array (ULA); in this way, with spatial filtering, it is possible to determine the acoustic ray passing through the centre of the array. Finally, sub-arrays are used to identify the acoustic ray passing through the centre of each of them. In this way, at the end of the procedure, is obtained a sampling of all the acoustic rays passing through the

ULA.

From this analysis approach, it was developed a synthesis technique following the same principle.
It starts with a plane wave decomposition of the sound field to be rendered. Then a ULA of loudspeakers is used and to its sub-arrays are given specific wavefront components to be rendered. It is like having a set of acoustic beams that braid to give the desired acoustic field. It is essential to underline that with this method, we include the directional information of the virtual source already in the rendering framework without the need for a previous analysis of the sound field to be rendered, like in NFC-HOA and WFS. It is being demonstrated [12] that this method is more accurate and computationally more efficient than previous methods. Another advantage is that to reconstruct the desired field is optional to surround the area of interest. As in WFS, it does not have the constraint of NFC-HOA to define a listening region. A simple set of soundbars could commercially implement this method.

1.1.2. Optimisation-based Methods

Optimisation-based methods are not to be considered an alternative to the model-based methods but as a sub-category. These techniques always start from a physical model and, through an optimisation process, e.g. minimisation of the error produced and desired pressure field values or modes, lead to the reproduction of the desired sound field. One of the main advantages of all these approaches is that they are independent of the system layout since the synthesis process consists only of the minimisation or maximisation of some parameters unrelated to the physical model. Thanks to this characteristic, these methods are widely studied when it comes to the synthesis of complex sound fields, and are also being used in research for the synthesis of multi-zones.

In recent years mode matching-based SFS have gained considerable attention [10, 48], due to the fact that it requires only one control point. In mode matching-based methods, the desired sound field is first described in terms of circular or spherical harmonics expansion, depending if we are in the 2D or 3D case. Then the same is done with the reproduced sound field. Finally, the solution is obtained by matching the desired and reproduced sound field coefficients with mode-by-mode matching.

It should be clear that a arbitrary order Ambisonics method is a mode-matching method. The particularity of Ambisonics is that it is applicable only with circular and spherical configurations because these are the only configurations for which exists a definitive solution for each mode.

A significant advantage of this method is that since it is based on the calculation of the modes, it needs to use only a control point to confront the desired sound field, since from that control point is done the spherical harmonics expansion.

Pressure Matching [44] is the first proposed technique that is based on the concept of optimisation-based methods. This procedure searches for the driving function that minimises the difference between the desired and synthesised pressure fields in specific Control Points (CPs). The main advantage of this method is that there is a closed-form solution for the estimation of the driving function, which depends only on the transfer function between the secondary sources and the control points and, of course, on the desired sound field to be reproduced.

In its original form, the optimisation procedure is performed through Least Squares (LS), and the error is measured on the pressure values assumed by the field in the CPs. Alternatively, the error is measured on the amplitude of the sound field without considering the phase (Amplitude Matching) [34].

More recently, also Deep Learning (DL) methods have been applied to the problem of sound field synthesis. These inherently fall in the category of optimisation-based approaches due to the functioning of Deep Neural Networks (DNN). DNNs are a valuable tool for describing non-linear behaviour, and the filtering applied by the driving function is non-linear for each speaker and between speakers.

Even though the results in standard conditions are not much different from other optimisation methods, these techniques were robust when applied in different acoustic problems in reverberant or noisy environments [39].

1.2. Multi-Zone Sound Field Generation

We talk about Multi-Zone SFS methods when we try to render simultaneously two or more sound fields with different characteristics in different zones inside the same environment, as shown in Fig 1.6.

Sound zones permit to enjoy different audio scenes in the same acoustic environment without disturbing each other. The research is focused on transmitting specific information to a region of interest by leaving also a quite zone as in classic Active Noise Control [41], i.e. synthesising zones with a high difference in acoustic energy [17]. In other words, the techniques proposed in the literature aim to maximize the sound field energy in one zone of the space and to minimize it in another one. In these cases, it is frequently used to call *bright* and *dark* zones to define the two areas, respectively.

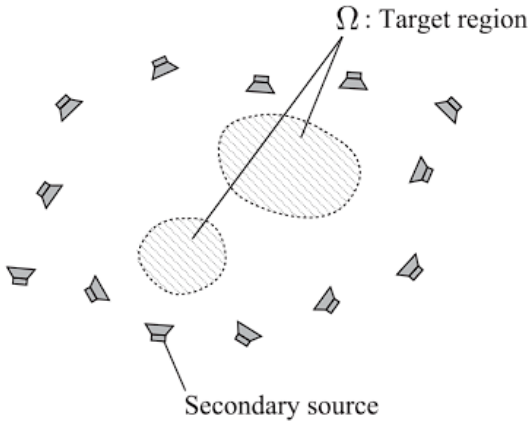


Figure 1.6: Personal Sound Zones. Image taken from [1]

The same Pressure Matching method itself, when presented, can also be applied to the problem of multi-zone rendering [47].

An alternative for PM is the Acoustic Contrast Control (ACC) method [17]. This technique is not well suited for scenarios different from the bright-vs-dark framework or when the controlled zones are generally more than two. In this approach, the energy ratio between the two zones is maximised. The comparison between the ACC and PM methods revealed that they perform best in their respective designated purposes. While PM excels in reproducing the bright zone due to its minimization of reproduction error, ACC is better suited for attenuating the dark zone [53].

A framework based on Variable Span Linear Filters (VSLF) [8] used to describe speech enhancement algorithms has been proposed. Variable Span Linear Filters are filters formed from linear combinations of the eigenvectors from the joint diagonalisation of the noise and desired signal correlation matrices. Its fundamental parameters are the number of eigenvectors used, and the rank of the correlation matrix of the desired field. By properly designing these filters, it is possible balance as desired by the user the trade-off between signal distortion and noise suppression.

Following the same logic, Variable Span Trade-off filters (VAST) are proposed for the problem of sound zone design in [37]. Here the filters are obtained from the joint diagonalisation of the spatial correlation matrices of the bright and dark zone, and the trade-off in designing the filters is between the reproduction error in the bright zone and the acoustic contrast. There is also a direct parallelism between the "special" cases of VSLF and VAST filters: for example, by choosing only the eigenvector corresponding to the highest eigenvalue we derive respectively a filter that gives the highest SNR and the

highest AC (ACC); in their respective field these are both approaches with the highest distortion. With the aforementioned framework, of which PM and ACC are special cases, it is possible to get intermediate results that can represent the sound zone control analougus of speech enhancement or noise suppression algorithms like the Minimum Variance Distortionless Response (MVDR), the Wiener Filter etc.

Amplitude Matching is a variant of a PM that has been investigated in recent years. It has been noticed that ACC, even if it can achieve excellent results in terms of acoustic contrast, cannot match the energy distribution of the desired field inside the two zones, which should be homogenous. The difference with respect to the PM method is that, by taking the absolute value of the desired and reproduced field in the optimisation process, it completely discards the phase and introduces a nonlinearity w.r.t. PM. Unfortunatelly, the Least Squares (LS) optimisation algorithm has a closed form solution only for linear problems, and for this reason, Amplitude Matching has no closed-form solution. However, the fields obtained with AM have a homogeneous energy distribution inside the two zones. For what concerns the main metrics used to evaluate multi-zone systems - i.e. the acoustic contrast and reproduction error - AM has intermediate results between PM and ACC.

1.3. Conclusive Remarks

In this chapter, we have presented various techniques that have been used for the synthesis of sound fields. Classic techniques, WFS and Ambisonics, have been described in detail. Even though these approaches are not recent, they are still widely studied nowadays, and many variants have been made to overcome their limitations. Limitations that in both cases can be summed up in a necessity of an enormous amount of loudspeakers to correctly render a sound field that can reach high frequencies in a sufficiently-wide reproduction area.

Based on the PAF, it has presented a more recent method that can render broad sound fields, but has yet to be explored much since it still needs many speakers to render complex sound fields.

Since these limitations are constantly present in all these model-based techniques, the research moved towards optimisation methods. For these cases, the parameters optimised are independent of the physical model and there is also an independency of the optimal solution from the layout. Still, physical constraints persist, and as for other techniques these methods achieve better results with more loudspeakers. However, their advantage is in the capacity to not degenerate too much as we decrease the number of sources or use irregular configurations. Furthermore, we can find the first experimentation with Neural

Networks within these methods.

The last part of this chapter describes current methods for the synthesis of personal sound zones, which is also the focus of this manuscript. Due to the complexity of the sound fields to be rendered, most methods are optimisation-based for these cases. The two methods that achieved better results have been presented - PM and ACC - and some of their variants.

It is noticeable how multi-zone sound field synthesis can be a critical aspect in the creation of immersive and personalized audio experiences in complex acoustic spaces. Conventional sound field synthesis techniques often rely on acoustic models, which may not accurately capture the complex behaviour of sound waves in such environments. Still, there is a need to improve the effectiveness and accuracy of these techniques. Deep Learning is a promising approach for enhancing multi-zone sound field synthesis. Due to their ability to learn complicated behaviours, deep neural networks could be used to model and reproduce complex sound fields more accurately. The next chapter will describe the mathematical theory and physics constraints in acoustics and some of the statistics behind Deep Learning to fully understand the method proposed in this thesis. In later chapters, it will be described the presented method and it will be shown results compared with some of the most current techniques.

2 | Theoretical Background

The first part of this chapter will introduce the theoretical aspects necessary to properly describe a sound field. The obtained model will then be applied to the multi-zone synthesis problem. We'll represent aforementioned problem into its main configurations. It is shown how the Sound Field Control problem can be formulated in terms of linear and non-linear optimisation, despite the original physical model used: in fact an optimisation procedure is shown both for methods that rely on a model based on the Kirchoff-Helmholtz Integral 2.7 and on spherical harmonics 2.22.

In the second part of this chapter it'll be shown the rationale behind deep learning theory. After a brief introduction to Machine Learning, a detailed description of Deep Learning algorithms, which are also used in this thesis, is presented. The description will be based on the standard and oldest configuration - i.e. Feed Forward Networks - and then extended to the Convolutional Neural Networks.

As it will be described in the chapter, acoustic fields are complex systems, and traditional methods for obtaining informative features can be challenging. Therefore, the choice of using a DNN in this work is also guided by the fact that it can automatically extract relevant features that reflect the characteristics of the data for a specific problem.

2.1. Sound Field Control

Before defining the Sound Field Control problem, we first describe some preliminary concepts of acoustics and spatial audio.

An acoustic field is a real-valued scalar function that represent the pressure, here defined as $p(\mathbf{r}, t)$, with $t \in \mathbb{R}$, denoting the time and $s \in \mathbb{R}^3$, denoting the space. Acoustic fields in volumes where active sources are present, satisfy the Inhomogeneous Wave Equation

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = -\frac{\partial q(\mathbf{r}, t)}{\partial t}, \tag{2.1}$$

being $q(\mathbf{r}, t)$ flow per unit volume or *excitation* caused by a source, and c the speed of

sound. Due to its time-harmonic behaviour the acoustic field can be redefined by applying the Fourier Transform as $P(\mathbf{r},\omega)e^{-j\omega t}$, with ω being the angular frequency.

In this domain, acoustic fields of (2.1) are described by the Inhomogeneous Helmholtz Equation

$$\nabla^2 P(\mathbf{r},\omega) + \left(\frac{\omega}{c}\right)^2 P(\mathbf{r},\omega) = -j\omega Q(\mathbf{r},\omega), \tag{2.2}$$

In the following we'll use apexs $'$ to refer to points in the space and the time instants of secondary sources. By considering a spatio-temporal impulse at $(\mathbf{r}', t')$, as excitation term we can model the secondary source as

$$q(\mathbf{r},t) = \delta(\mathbf{r}-\mathbf{r}')\delta(t-t'), \tag{2.3}$$

being $\delta(\cdot)$ Dirac's delta. As described in [59], by replacing the exctitation term of (2.1) with (2.3) we therefore obtain

$$g(\mathbf{r}|\mathbf{r}',t) = \frac{1}{\|\mathbf{r}-\mathbf{r}'\|}\delta(t-\frac{\|\mathbf{r}-\mathbf{r}'\|}{c}), \tag{2.4}$$

$$G(\mathbf{r}|\mathbf{r}',\omega) = \frac{e^{-j(\frac{\omega}{c})\|\mathbf{r}-\mathbf{r}'\|}}{4\pi\|\mathbf{r}-\mathbf{r}'\|}, \tag{2.5}$$

as solutions to the wave and Helmholtz equation, respectively.

The equation (2.5) is called Green's function and represents in the frequency domain the sound field in $\mathbf{r}$ resulting from a spatio-temporal impulse in $\mathbf{r}'$. However, usally we have a complex excitation term. In these cases, the Green function can be used to describe arbitrary solutions of the wave equation through the Single Layer Potential

$$P(\mathbf{r},\omega) = \oint_{\partial\Omega} G(\mathbf{r}|\mathbf{r}',\omega)D(\mathbf{r}',\omega)d\mathbf{r}', \mathbf{r}' \in \partial\Omega, \tag{2.6}$$

being $\partial\Omega$ the surface of the volume under consideration.

The following introduction will be based on [2]. The SLP has a direct derivation from the Kirchoff-Helmholtz (K-H) integral that is one of the essential theorems in acoustics. It states that the sound pressure is completely determined within a volume free of sources, if sound pressure and velocity are determined in all points on its surface

$$a(\mathbf{r})P(\mathbf{r},\omega) = \oint_{\partial\Omega} (G(\mathbf{r}|\mathbf{r}',\omega)\langle\nabla p(\mathbf{r},\omega), \hat{\mathbf{n}}(\mathbf{r}')\rangle|_{\mathbf{r}=\mathbf{r}'} + \\ -P(\mathbf{r}',\omega)\langle\nabla G(\mathbf{r}|\mathbf{r}',\omega), \hat{\mathbf{n}}(\mathbf{r}')\rangle)\, dA(\mathbf{r}'), \mathbf{r}' \in \partial\Omega \quad (2.7)$$

with $a(\mathbf{r})$ being the discrimination term, defined as

$$a(\mathbf{r}) = \begin{cases} 1, & \text{if } \mathbf{r} \in \Omega_i \\ \frac{1}{2}, & \text{if } \mathbf{r} \in \partial\Omega \\ 0, & \text{if } \mathbf{r} \in \Omega_e \end{cases}. \quad (2.8)$$

In the 2.7 $\partial\Omega$ denotes a surface enclosing the source-free volume Ω_i, $A(\mathbf{r}')$ an infinitesimal surface element of $\partial\Omega$, $\mathbf{r}'$ a point on $\partial\Omega$; while Ω_e denotes the domain outside $\partial\Omega$; $\hat{\mathbf{n}}$ is the unit vector pointing in direction inward the surface normal.

Green's function directional gradient $\langle\nabla G(\mathbf{r}|\mathbf{r}',\omega), \hat{\mathbf{n}}(\mathbf{r}')\rangle$ can be interpreted as the spatio-temporal transfer function of a dipole sound source whose main axis lies parallel to $\hat{\mathbf{n}}$. Thus, the K-H integral can be seen as made of two components displayed on the boundary of the source-free volume, one representing a layer of *secondary monopole sources* and a second layer of *secondary dipole sources*. A sound field synthesized with such a distribution would exhibit the desired properties, i.e. the reproduced field would match the desired one.

The reader may understand that having two superimposed layers of loudspeakers is quite impractical. In fact, usually the dipole layer is discarded. The first component of the K-H integral is what is called the acoustic SLP (2.6). By comparing (2.6) and (2.7) we can deduce that the function $D(\mathbf{r}',\omega)$ in 2.6 represent the gradient of the sound pressure in direction of the inward pointing surface normal on $\partial\Omega$, and in this work will be termed as distribution *density* of the potential, *driving function* or *volume velocity*.

With nowadays technology continuous distribution of loudspeakers are not feasible. Hence, in practical implementation discrete distribution are used. With L loudspeakers the above definition of the SLP 2.6 can be discretised as

$$P(\mathbf{r},\omega) = \sum_{l=1}^{L} G(\mathbf{r}|\mathbf{r}'_l,\omega)D(\mathbf{r}'_l,\omega). \quad (2.9)$$

Most of the SFC techniques rely on specific control points used to evaluate the reproduced field, i.e. the microphones used to measure the pressure values. We'll define a matrix representing the Green function between M control points and L control sources, the vector

representing the driving signal to each speaker and the pressure field vector describing the values of the pressure field in each point in space as

$$\mathbf{G}(\mathbf{r}|\mathbf{r}',\omega) = \begin{bmatrix} G(\mathbf{r}_1|\mathbf{r}'_1,\omega) & \cdots & G(\mathbf{r}_1|\mathbf{r}'_L,\omega) \\ \vdots & \ddots & \vdots \\ G(\mathbf{r}_M|\mathbf{r}'_1,\omega) & \cdots & G(\mathbf{r}_M|\mathbf{r}'_L,\omega) \end{bmatrix}, \tag{2.10}$$

$$\mathbf{d}(\mathbf{r}',\omega) = \begin{bmatrix} D(\mathbf{r}'_1,\omega) \\ \vdots \\ D(\mathbf{r}'_L,\omega) \end{bmatrix}, \tag{2.11}$$

$$\mathbf{p}(\mathbf{r},\omega) = \begin{bmatrix} P(\mathbf{r}_1,\omega) \\ \vdots \\ P(\mathbf{r}_M,\omega) \end{bmatrix}. \tag{2.12}$$

2.1.1. Pressure Matching

Pressure Matching, was first introduced by Nelson et. al [44] in the early nineties as a least square optimisation-based technique for sound field synthesis. In [47] the pressure matching method was first applied to the synthesis of personal sound zones. The present exposition will adhere to the last-mentioned problem description.

Even though, the mechanism of this procedure could be applied by considering multiple zones, in the following description we'll consider a case with a bright zone and a dark zone, being $2M$ the total number of control points of the desired zones to synthesize, precisely M_b and M_d are the points referred to the bright and dark zone, respectively. We define the desired sound field values at the control points as

$$P^{des}(\mathbf{r}_m,\omega) = \begin{cases} \sum_{l=1}^{L} G(\mathbf{r}_m|\mathbf{r}'_l,\omega)D(\mathbf{r}'_l,\omega), & m = 1,\ldots,M_b \\ 0, & m = M_b+1,\ldots,M_b+M_d \end{cases}. \tag{2.13}$$

The goal of this technique, as shown in Fig. 2.1, is to minimise the squared error between the values of the desired pressure field $P^{des}(\mathbf{r}_m,\omega)$ and the estimated pressure field $P^{est}(\mathbf{r}_m,\omega)$ at the control points, and can be written (by omitting the arguments $\mathbf{r}$ and ω) as

$$\mathcal{J}(\mathbf{d}) = \min_{\mathbf{d}\in\mathbb{C}^L} |\mathbf{G}\mathbf{d} - \mathbf{p}^{des}|^2 + \iota|\mathbf{d}|^2, \tag{2.14}$$

where ι is a regularization parameter used to constrain the power, and $\mathcal{J}(\mathbf{d})$ is the function to be minimised.

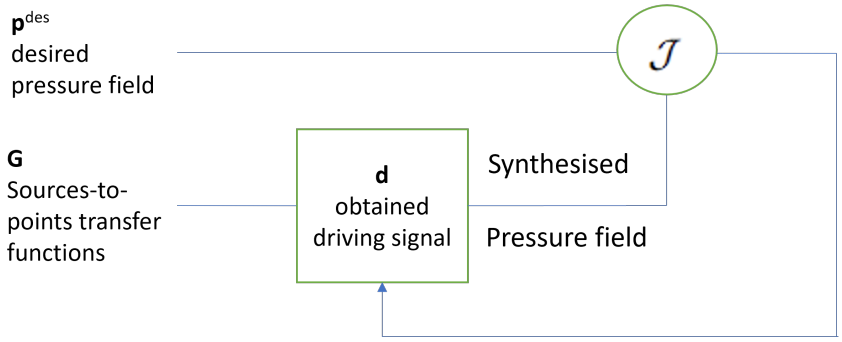


Figure 2.1: Scheme of the Pressure Matching algorithm

From this formulation, applying a Least Squares (LS) linear regression problem, we can obtain the closed-form solution

$$\hat{\mathbf{d}}_{pm} = (\mathbf{G}^H\mathbf{G} + \iota\mathbf{I}_L)^{-1}\mathbf{G}^H\mathbf{p}^{des}, \tag{2.15}$$

with the subscript pm denoting we refer to the optimal driving function given by the minimisation procedure, the apex H denoting the Hermitian matrix and $\mathbf{I}_L$ the $L \times L$ identity matrix.

2.1.2. Acoustic Contrast Control

The Acoustic Contrast Control (ACC) technique represents a sound zoning approach that aims to optimise the mean squared sound pressure within a specified zone, while maintaining a constant pressure level in the surrounding zones, as shown in Fig. 2.2. In the following description we will comply with its original formulation, which allows for the synthesis of only two zones. However is been recently demonstrated that is possible to use it for the synthesis of more than just one dark zone [1].

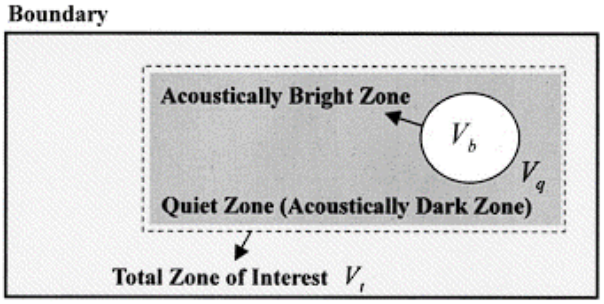


Figure 2.2: Schematic of the acoustic zones. The subscripts b and q refer to the bright and quite zones, respectively. Image from [17].

We will term *bright* the zone to which the mean squared sound pressure is optimised, while the surrounding zones will be termed as *dark* or *quite*. The Acoustic Contrast (AC) is the ratio between the acoustic potential energy in the bright zone and the acoustic potential energy in the dark zone. Thus, is necessary to express this ratio using driving vectors in the form of a cost function to be optimised. The averaged acoustic potential energy of a controlled zone can be written as

$$e(\omega) = \frac{\int_V p(\mathbf{r},\omega)^* p(\mathbf{r},\omega)\, dV}{V}, \tag{2.16}$$

where the apex * denotes the complex conjugate and V the volume of the control zone considered. This expression can be re-written as

$$e(\omega) = \mathbf{d}^H(\omega)(\frac{\int_V G(r|\mathbf{r}',\omega)^H G(r|\mathbf{r}',\omega)\, dV}{V})\mathbf{d}(\omega) = \mathbf{d}^H(\omega)\mathbf{R}(\omega)\mathbf{d}(\omega), \tag{2.17}$$

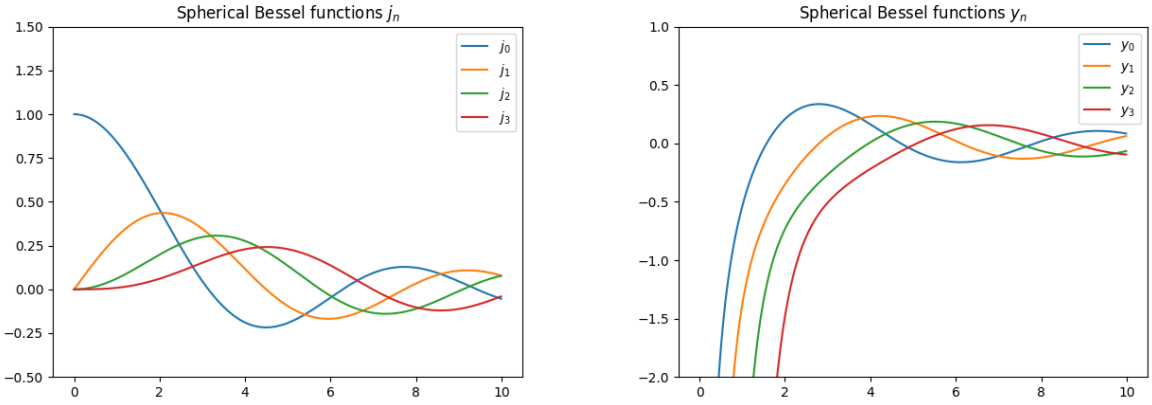
where $\mathbf{R}$ is the spatial correlation of the pressure field in the controlled zone produced by each control source. With this definitions we can derive an expression for a cost function to maximise the Acoustic Contrast. The AC can then defined as

$$\beta(\omega) = \frac{e_b(\omega)}{e_q(\omega)} = \frac{\mathbf{d}^H(\omega)\mathbf{R}_b(\omega)\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\mathbf{R}_q(\omega)\mathbf{d}(\omega)}, \tag{2.18}$$

and the optimisation problem can be formulated as

$$\max_{\mathbf{d}(\omega)\in\mathbb{C}^L} \beta(\omega). \tag{2.19}$$

Since the pressure fields produced by each control source, i.e. the sources used to synthesize the two zones, are linearly independent within the total zone of interest, the Hermitian



(a) (b)

Figure 2.3: Spherical Bessel Function of (a) First and (b) Second kind.

matrix R_q has full rank L and is invertible. Thanks to this characteristic it is possible to obtain a closed-form solution for the volume velocity vector that maximizes β as the eigenvector that corresponds to the largest eigenvalue of the matrix

$$\mathbf{R}_q^{-1}(\omega)\mathbf{R}_b(\omega) \tag{2.20}$$

It's noteworthy that the position of the control sources and of the control zones have to be determined.

2.1.3. Mode Matching

The mode-matching method aims to match the modes of the synthesized and desired sound fields at a certain control point. A mode in the three-dimensional (3D) context usually means a spherical wavefunction, by which a sound field can be expanded. The spherical wavefunctions, which are the products of spherical Bessel functions 2.3 and spherical harmonics 2.4, are solutions of the Helmholtz equation. The usage of the just represented model is a major difference between the mode-matching method and the pressure-matching method. However, it is necessary in the mode-matching method to truncate the modes, which strongly affects its reproduction accuracy, depending if the truncation order is excessively large or small.

We can represent the basis solutions to the Helmholtz equation (2.2) in spherical coordinates as

$$P(\mathbf{r},\omega) = R(r)\Theta(\theta)\Phi(\phi) = R(r)Y_o^m(\theta,\phi), \tag{2.21}$$

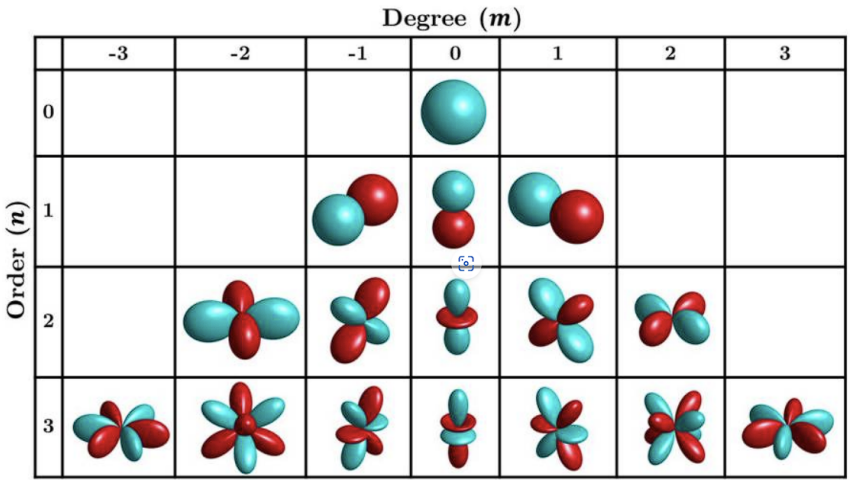


Figure 2.4: Spherical harmonics up to the third order. Image from [62]

with $\Theta(\theta)$ and $\Phi(\phi)$ representing respectively the co-elevation and azimuth dependencies components and with $R(r)$ being the radial component. $Y_o^m(\theta, \phi)$ incorporates the angular dependencies and is called *spherical harmonic* function

$$Y_o^m(\theta, \phi) = (-1)^m \sqrt{\frac{(2o+1)(o-|m|)!}{4\pi(o+|m|)!}} e^{jm\phi} P_o^{|m|}(\cos\theta), \tag{2.22}$$

where m denotes the degree of the spherical harmonic and o its order. Possible configurations of spherical harmonics are shown in Fig 2.4.

While radial dependency for interior field problems can be expressed in terms of Bessel functions

$$R(r) = R_1 j_o(\frac{\omega}{c} r) + R_2 y_o(\frac{\omega}{c} r), \tag{2.23}$$

where j_o and y_o refer to the Bessel functions of first and second kind, respectively.

A generic interior field can then be expressed using the *Inverse Spherical Harmonics* (ISH) expansion.

$$P(\mathbf{r}, \omega) = \sum_{o=0}^{\infty} \sum_{m=-o}^{o} (C_{o,m}(\omega) j_o(\frac{\omega}{c}\mathbf{r}) + B_{o,m}(\omega) y_o(\frac{\omega}{c} r)) Y_o^m(\theta, \phi). \tag{2.24}$$

We can characterize the sound field with the sets of coefficients $C_{o,m}, B_{o,m}$, that hereafter will be termed as *modes*.

All mode-matching methods try to synthesize a desired sound field by matching the modes of the reproduced field with the desired ones. It can be observed that with the presented characterization the number of modes to be matched should reach infinity. In practice, the

order of the spherical harmonics is truncated to an arbitrary degree M, and consequently of the maximum order O.

The choice of the order is usually driven by two reasons, depending by the context: when it comes to simulations, the order that gives the best result is chosen empirically; while in practical situations the order is upper bounded by the maximum number of loudspeakers available, since there is a direct correlation between the maximum order used and the number of loudspeakers used. Thus, the expression 2.24 can be approximated

$$P(\mathbf{r},\omega) \approx \sum_{o=0}^{O} \sum_{m=-o}^{o} (C_{o,m}(\omega) j_o(\frac{\omega}{c} r)) Y_o^m(\theta,\phi). \tag{2.25}$$

Recently a Weighted Mode-Matching (WMM) technique [35] has been proposed, where the order that leads to the minimum reproduction error is obtained by means of an optimization procedure.

In the following description we'll use $\sum_{o,m}$ to express $\sum_{o=0}^{O} \sum_{m=-o}^{o}$.

With the goal of having

$$C_{o,m}(\mathbf{r}_c,\omega)^{des} = C_{o,m}(\mathbf{r}_c,\omega)^{syn}, \tag{2.26}$$

with the superscripts *syn* and *des* used to refer to the synthesized and desired sound fields respectively, we can obtain a result by solving

$$\min_{\mathbf{d}\in\mathbb{C}^L} \sum_{o,m} |C_{o,m}(\mathbf{r}_c,\omega)^{des} - C_{o,m}(\mathbf{r}_c,\omega)^{syn}|^2 + \iota \mathbf{d}^H \mathbf{d}. \tag{2.27}$$

Here $\mathbf{r}_c$ represents the expansions center, that in our case is the origin of our system. The solution of 2.27 can be expressed as

$$\hat{\mathbf{d}}_{mm} = (\mathbf{A}_{mm} + \lambda \mathbf{I}_L)^{-1} \mathbf{b}_{mm}, \tag{2.28}$$

with $\mathbf{A}_{mm} \in \mathbb{C}^{LxL}$ and $\mathbf{b}_{mm} \in \mathbb{C}^L$ given by

$$(\mathbf{A}_{mm})_{l_1,l_2} = \sum_{o,m} C^{\mathbf{G}}_{l_1,o,m}(\mathbf{r}_c) * C^{\mathbf{G}}_{l_2,o,m}(\mathbf{r}_c), \tag{2.29}$$

$$(\mathbf{b}_{mm})_l = \sum_{o,m} C^{\mathbf{G}}_{l,o,m}(\mathbf{r}_c) * C^{des}_{o,m}(\mathbf{r}_c). \tag{2.30}$$

Here, $(\cdot)_{l_1,l_2}$ denotes the $(l_1,l_2)^{th}$ element of the matrix, $(\cdot)_l$ denotes the l^{th} element of the

vector, and the apex $\mathbf{G}$ denotes that the coefficients are referred to the spherical expansion of the Green's function. For the synthesis of multiple zones Ω^q, with $q = \{0, \ldots, Q\}$, the problem 2.27 is reformulated as

$$\min_{\mathbf{d}\in\mathbb{C}^L} \sum_{q=0}^{Q} \gamma_q \sum_{o,m} |C_{o,m}(\mathbf{r}_c, \omega)^{des,q} - C_{o,m}(\mathbf{r}_c, \omega)^{syn}|_q^2 + \lambda \mathbf{d}^H \mathbf{d}, \tag{2.31}$$

where γ_q is a constant parameter used to weight all the square differences.

The optimum solution, is then obtained as

$$\hat{\mathbf{d}}_{mm,Q} = (\sum_{q=0}^{Q} \gamma_q \mathbf{A}_{mm}^q + \lambda \mathbf{I}_L)^{-1} \sum_{q=0}^{Q} \gamma_q \mathbf{b}_{mm}^q. \tag{2.32}$$

2.1.4. Amplitude Matching

Amplitude Matching aims to minimize the error between amplitude distributions of synthesized and desired sound fields. In some applications, it is necessary to synthesize a sound field of the desired amplitude inside the target region, whereas the phase of the desired sound field is arbitrary [11]. Thus, more than actually reproduce a specific sound field, the result obtained is to have an equal power distribution over the target zone by optimising the desired amplitude and discarding the phase. In Amplitude Matching's optimisation procedure is present a modulus operator. This leads to a non-linear optimisation problem, which requires non-linear optimization algorithms.

In order to solve the Amplitude Matching problem, it is usually applied the Minimisation-Maximisation [34].
As many others non-linear optimisation algorithms the basis of the method is to construct a surrogate linear function of the non-linear objective function (at each iteration). So reformulating Pressure Matching minimisation problem (2.14) by adding the amplitude constraint it becomes

$$\min_{\mathbf{d}\in\mathbb{C}^L} \mathcal{J}(\mathbf{d}) = \||\mathbf{G}\mathbf{d}| - |\mathbf{p}^{des}|\|^2 + \iota\|\mathbf{d}\|^2, \tag{2.33}$$

where $|\cdot|$ is the element-wise absolute value. The monotonic decrease in the objective function can be guaranteed by alternately updating the variable of the surrogate function and the variable to be optimized. We first define the variable of the surrogate function as

$$\mathbf{v}_k = |\mathbf{p}^{des}| \exp\left(j \arg(\mathbf{G}\mathbf{d}_k)\right), \forall \mathbf{d}_k \in \mathbb{C}^L, \tag{2.34}$$

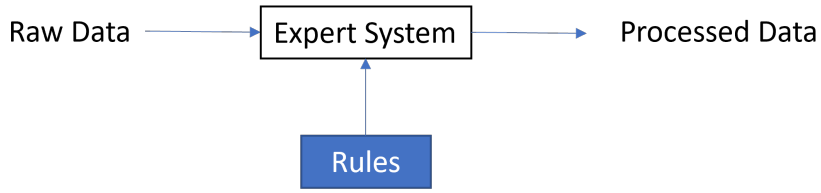


Figure 2.5: Traditional *expert system* used to process some data

where $\arg(\cdot)$ is the argument of a complex variable and k is index of the iteration. Then we can define the surrogate function as

$$\mathcal{J}^{+}(\mathbf{d}|\mathbf{v}_k) = \|\mathbf{G}\mathbf{d} - \mathbf{v}_k\|^2 + \iota\|\mathbf{d}\|^2 \geq \mathcal{J}(\mathbf{d}). \tag{2.35}$$

Thus, by alternately updating $\mathbf{v}_k$ and $\mathbf{d}_k$ we obtain a solution for the driving signal $\mathbf{d}$ as

$$\mathbf{v}_k = |\mathbf{p}^{des}| \exp\left(j \arg(\mathbf{G}\mathbf{d}_k)\right), \tag{2.36}$$

$$\mathbf{d}_{k+1} = \min \mathbf{d} \in \mathbb{C}^L \mathcal{J}^{+}(\mathbf{d}|\mathbf{v}_k) = (\mathbf{G}^H\mathbf{G} + \iota\mathbf{I}_L)^{-1}\mathbf{G}^H\mathbf{v}_k. \tag{2.37}$$

The monotonic non-increase of the objective function can be verified as

$$\mathcal{J}(\mathbf{d}_{k+1}) \leq \mathcal{J}^{+}(\mathbf{d}_{k+1}|\mathbf{v}_k) \leq \mathcal{J}^{+}(\mathbf{d}_k|\mathbf{v}_k) = \mathcal{J}(\mathbf{d}_k) \tag{2.38}$$

This algorithm is iterated until a stopping condition is met, e.g. a threshold for the variation of $\mathcal{J}(\mathbf{d}_k)$ or $\mathbf{d}_k$.

2.2. Deep Learning

Deep Learning is a sub-category of Machine Learning (ML), which is a branch of the Artificial Intelligence field. With ML we use the available data to automatically learn from the data itself the set of rules and algorithms needed to perform a determined task. By comparing Figs. 2.5 and 2.6 we can see an example on how ML could be applied for a generic task. The procedure by which the model learns the function that maps input to output is called *training*. Basically, as described in Fig. 2.7 the ML model is trained using a dataset, called *training set*, and later tested on a different dataset that must not contain the same examples of the training set, called *test set*.

Usually, machine learning's models are not fed with raw data, but with data that has

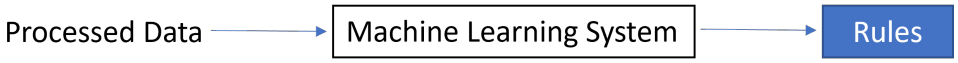


Figure 2.6: ML system used to learn a set of rules that can be applied to process data

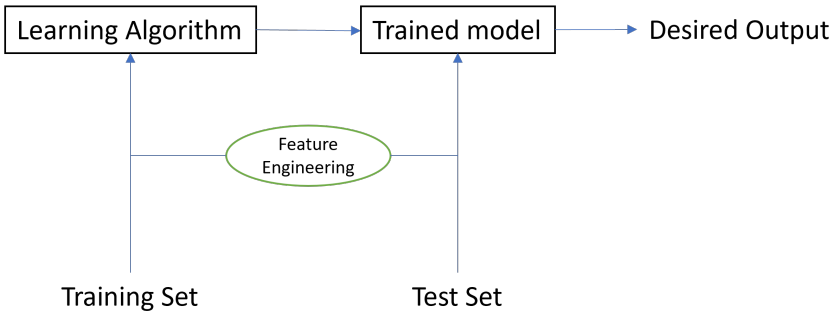


Figure 2.7: Schematic of ML system usage

passed through preprocessing operations. Through this process some aspects of the input are extracted using well-known, established algorithms and procedures. The outputs of the preprocessing steps are called *features* and the choice of the right ones to use is fundamental to obtain a model capable to perform the chosen task. This pre-processing step is commonly called *feature engineering* (FE).

2.2.1. Features

A Neural Network (NN) is a class of machine learning models inspired by the structure and function of biological neurons. It is composed of layers of artificial neurons that process input data and produce corresponding output data. We call a *layer* a set of artificial neurons that process input data and produce corresponding output data: as shown in Fig. 2.8, these sets of neurons are organized into a sequential or hierarchical structure, with the input of one layer serving as the output of the preceding layer.

One of the main advantages of Neural Networks, if compared to standard Machine Learning methods, is that they don't need any feature extraction procedure. NN models can be fed with raw data, and the features are automatically extracted as part of the learning procedure; in this case we talk about *feature learning* (FL). Basically every layer extracts features from the previous layer. Thus, first layers will extract low-level features, and as we go deep in the network, higher-level features are derived. The last layer can be used for many purposes, for example classification and regression.
Hence, classification and regression are computed based on the features which correspond to the higher level of abstraction.

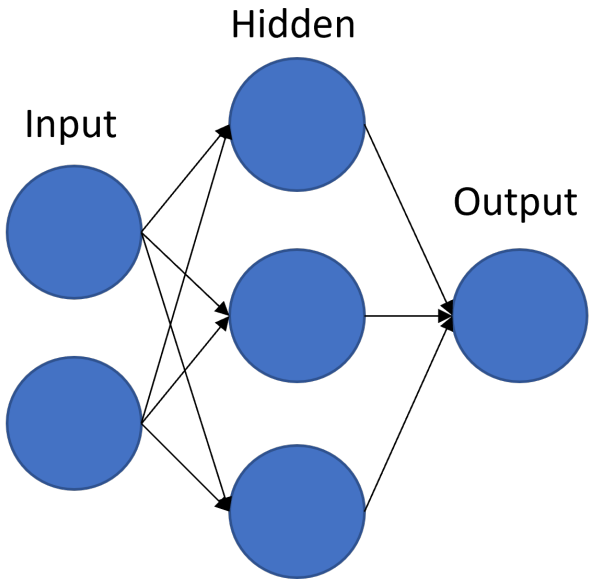


Figure 2.8: Simple NN Architecture composed by 3 layers of 2, 3 and 1 neurons, respectively

An issue that is frequently encountered is that is not easy to understand what the features extracted by a deep learning model represent: usually for engineered features, there is a strong background theory related to the field of research; in deep learning models, features are obtained through probability and statistics, applied to a non-linear optimisation procedure.

2.2.2. Learning

To describe the learning procedure is necessary to explain how is structured the "oldest" family of Neural Networks, called Multi-Layer Perceptron (MLP) or Feed-Forward Dense Neural Networks (FFDNN). A Feed-Forward Neural Network is composed by a series of layers, where each layer itself composed by a set of elemental operators. What these elemental operators - called Neurons or Perceptron - do is to take an input signal z, apply a weight y, add a bias term v and then apply an activation function ϵ to produce an output signal. This description can be synthesized in Fig. 2.9 and the following function

$$f(z|\eta) := \epsilon(y^T z + v), \tag{2.39}$$

where $\eta = (y, v, \epsilon)$ are the parameters updated at each iteration. Weights represent the strength of the connections between the input and the output, while the biases control the threshold for the following activation. Activations are a crucial elements when it comes to

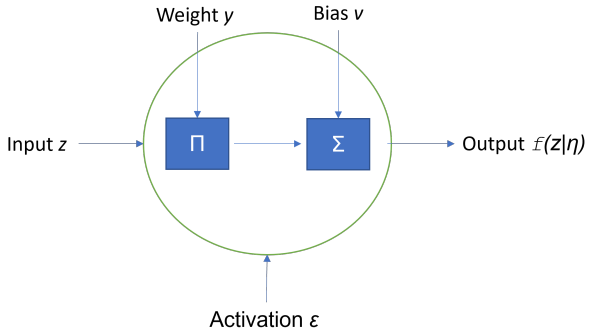


Figure 2.9: Description of operations performed by a single Perceptron

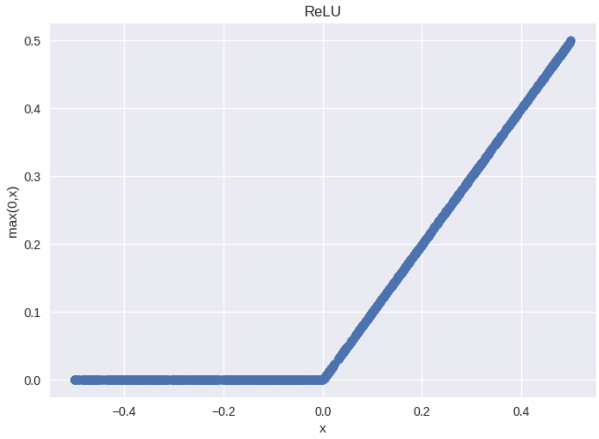


Figure 2.10: Rectified Linear Unit

the learning procedure: these are non-linear transfer functions that determine the output and make the neural network a non-linear model; without them a neural network would simply be a linear combination with biases. One of the most used activations is the Rectified Linear Unit (ReLU) [42] shown in Fig. 2.10, that saturates to zero all negative results

$$\epsilon_{ReLU}(z) = \max(0, z). \tag{2.40}$$

As described above, the intermediate layers can be interpreted as feature extractors. By jointly optimising the intermediate and output layers, the model finds a feature extractor which processes the data so that the output layer performs well.

Parameters are estimated by an iterative method. The method used to optimize the parameters is the Gradient Descent (GD), i.e. through the gradient the model learns how it changes the function as we change its weights. We'll refer to the scaling factor as *learning rate* $lr \in (0, 1]$, that is the size of the step taken at each iteration for the exploration of the space of solutions.

The learning algorithm that NNs use to compute the gradient weights is called *back-propagation* [27]. The name of the algorithm is due to the fact that the error at the output layer is propagated to the previous layers.

In DL, is very rare and improbable to have a convex function to optimize, so we also have local minima and not only a global one. To avoid to get stuck in local minima, Stochastic Gradient Descent is used (SGD), that is a stochastic approximation of GD where the whole dataset is replaced with randomly selected samples of the data, called batch: the gradient direction is estimated from these randomly selected samples. The size of this minibatch must ensure that we get enough stochasticity to avoid local minima. It behaves in an erratic way, and this randomness helps to avoid getting trapped before reaching the function's minimum, and so to find a better (or hopely the best) solution.

We refer to the term *training* to the process of learning the parameters, i.e. the optimisation is done during training time, in which we try to fit the parameters.

2.2.3. Loss

In Deep Learning, the optimisation process involves learning parameters through a continuous and differentiable *loss* function, which compares the predicted output with the actual output, or Ground Truth (GT). Back-propagation of the gradient is used to update the parameters η iteratively until the desired result is reached.

Also the choice of the loss function is fundamental, for enabling successful model training. Specifically, selecting an appropriate loss function is necessary for the model to be trainable and converge to an optimal solution. One of the simplest loss functions that can be thought is the Mean-Squared Error (MSE), i.e. the mean of the squared differences between the obtained and desired result

$$MSE = \frac{\sum_{n=1}^{N}(p_{des}(n) - p_{out}(n))^2}{N}, \tag{2.41}$$

where N is the length of the vector containing the examples evaluated and the subscripts *des* and *out* indicate if we're referring to the desired or obtained output, respectively.

2.2.4. Convolutional Neural Networks

Convolutional networks also known as convolutional neural networks, or CNNs, are a specialized kind of neural network for processing data that has a known grid-like topology. An example is image data, which can be thought of as a 2-D grid of pixels. The name "convolutional neural network" indicates that the network employs a mathematical op-

eration called convolution which is a specialised kind of linear operation. Convolutional networks are simply neural networks that use convolution or correlation in place of general matrix multiplication in at least one of their layers.

First, differently from MLPs where every input of a layer interacts with every output of the previous layer, in convolutional networks typically we have sparse interactions. This is accomplished by using *kernels* - the CNN counterpart of MLP's neurons, also called *filters* - smaller than the input. In this way, in image processing we can detect meaningful features such as edges with kernels that occupy only a small number of pixels.

Secondly, the sharing of parameters between filters lead CNNs to be equivariant to translations, i.e. if we move an object in the input, its representation will move the same amount in the output.
This is useful for when we know that some function of a small number of neighboring pixels is useful when applied to multiple input locations. For example, when processing images it could help to detect edges in the first layer of a convolutional network. The same edges appear more or less everywhere in the image, so it is practical to share parameters across the entire image.

Like in standards MLPs we have biases and activation functions, but the main difference - as described above - is that the activation is not applied to an inner product but to a correlation, i.e. Eq. (2.42) becomes

$$f_{CNN}(z|\eta) := \epsilon(z * y^T + v), \tag{2.42}$$

with $*$ representing the convolution operation.

The good thing about CNNs is that layers at different depths tend to specialize to different patterns, rather than abstracting the previous layer as in FFDNNs.

Convolutional Neural Networks are the most common tool used in image processing and computer vision. We can apply them to any kind of data that is in matrix form, e.g. time-frequency representation of audio signals and pressure field representations. We can imagine treating our acoustic fields as if they are images from which we try to learn edges and corners, e.g. directions and radiation properties of pressure waves.

2.2.5. Conclusive Remarks

In the first section of this chapter, we described the main mathematical and physical concepts related to acoustics, needed in order to develop the technique proposed in this

thesis, particularly the concepts used for SFS and SFC. We've made a detailed description of the most used techniques, namely Pressure Matching, Acoustic Contrast Control, Mode-Matching and Amplitude Matching. All the above-mentioned methods are at the state-of-the-art for what concerns the multi-zone sound field synthesis problem. They all have in common that they try to calculate the optimal driving functions with by means of an optimisation problem: for most of the approaches the problem is linear and thus they present a closed-form solution; the only exception is for the Amplitude Matching algorithm, for which an iterative procedure is necessary since the problem to solve presents a non-linearity.

The second part of this chapter gave a brief description of the basic concepts behind the functioning of Deep Neural Networks. It is shown the optimisation mechanism of Neural Networks and why CNNs achieve greater results w.r.t. standard MLPs. NNs are based on an optimisation procedure, this make them perfectly fit the purpose of finding an optimal driving function to synthesize a sound field, in particular with multi-zones.

It's noticeable how each of the SFC methods optimises differently due to the fact that they aim at reaching different objectives. Deep Learning could easily be applied to each of these optimisation problems. In this work we focus on the reproduction of the desired pressure field through the minimisation of the reproduction error, since our goal is to correctly reproduce a pressure field.

3 | Proposed Method

In this chapter we will describe the proposed method in detail, from the mathematical theory used, to the design of the Neural Network. After a first formulation of our problem, we carry a detailed description of the proposed model. The last part of this chapter is focused on the explaination of our training procedure.

The objective of our system is to to correctly reproduce a desired sound field in a determined area, while attenuating the pressure field in a different a different zone, inside the same environment. Our problem follows the rationale of the Pressure Matching method presented in 2.1.1, i.e. to minimise the error between our estimation and the ground truth at the control points inside the two regions. The main difference with the aforementioned approach is that the optimisation is not performed through Least Squares, but through a Deep Neural Network 2.2.4.

Through the adoption of a neural network model is possible to find a solution that analyse the non-linearities of the system. This last characteristic is not possible to be investigated in PM since LS is an algorithm that can be applied only to linear problems. Our network is designed to extrapolate from the values at the control points of a desired pressure field the driving signals that allow us to replicate the sound field in a determined region of our environment. Due to the different characteristic of the sound zones to be synthesized, during the training we treat bright and dark zones differently.

3.1. Problem Formulation

Let us consider - as shown in Fig 1.6 - L loudspeakers deployed in positions $\mathbf{r'}_l$, $l = 1, \cdots, L$ and M control points $\mathbf{r}_m$, $m = 1, \cdots, M$ used to measure the pressure in the q^{th} area, being $q = 1, \cdots, Q$ and Q the number of examined regions inside the considered environment $\mathcal{Q}$. The goal of a SFS technique is to obtain the optimal driving function $\mathbf{d}(\mathbf{r'}, \omega)$ that allows us to best approximate the desired acoustic field. By expressing each term of the discrete Single Layer Potential (2.9) in terms of vectors as defined in (2.12), (2.11) and

(2.10), we can reformulate the equation as

$$\mathbf{p}(\mathbf{r},\omega) = \sum_{l=1}^{L} \mathbf{G}(\mathbf{r}|\mathbf{r'}_l,\omega)\mathbf{d}(\mathbf{r'}_l,\omega). \tag{3.1}$$

It's noteworthy that the set of points computed of the pressure field directly depends on the set of points considered in the transfer function loudspeaker-to-point $G(\mathbf{r}|\mathbf{r'}_l,\omega)$. That means, that even though for training procedure we only compute the pressure field at the control points, we could use the same driving function to compute the entire acoustic scene inside our environment $\mathcal{Q}$: the only difference needed to obtain such result would be to use a transfer function that considers every point inside $\mathcal{Q}$.

In the proposed method, we apply a Deep-Learning based Pressure Matching (DLPM) approach as proposed in [19] and modify it in order to perform the synthesis of multiple acoustic scenes, precisely for the optimal formulation of the driving function. Our procedure follows the Pressure Matching technique for multi-zone synthesis as described in 2.1.1, but the main difference is in the optimisation algorithm: instead of using the Least Square regression, we used a DNN.
Thus, by reformulating PM's minimisation problem (2.14), we can express our optimisation problem as

$$\min_{\mathbf{d}_{dlpm}\in\mathbb{C}^L} \sum_{m=1}^{M} |\mathbf{G}(\mathbf{r_m})\mathbf{d}_{dlpm} - \mathbf{p}^{des}(\mathbf{r}_m)|^2, \tag{3.2}$$

where $\mathbf{d}_{dlpm}$ is the output of our optimisation procedure, i.e. the DNN training.

In classic DL methods, a NN is fed with some input data, and the output of the system is compared with a predefined ground truth, by means of a loss function that often needs to be minimised.

For our problem we don't have a ground truth set of driving functions, but a set of ground truth sound fields is easily obtainable as described in Sec. 2.1. Hence, instead of directly use driving functions for the comparison, we apply the loss to two sound fields. For this purpose, we use the output of our system, i.e. the estimated driving function $\mathbf{d}^{est}(\mathbf{r'},\omega)$, to obtain our estimated pressure field $\mathbf{p}^{est}(\mathbf{r},\omega)$ through the vectorised discrete SLP (3.1), and we apply the loss function to compare our estimated acoustic field with the desired one $\mathbf{p}^{des}(\mathbf{r},\omega)$ at the control points.

The input of the neural network model consists of the pressure field of the desired bright zone $\mathbf{p}_b^{des}(\mathbf{r},\omega)$. We omit the dark zone because it's an area where all values are equal, hence it would not add any discriminative information from the learning purpose. Follow-

ing [36], we don't feed the proposed model with complex matrixes representing pressure fields, but as input we used a vector containing a concatenation of real and imaginary parts, i.e.

$$\tilde{\mathbf{p}}_b^{des}(\mathbf{r},\omega) = \begin{bmatrix} \Re(\mathbf{p}_b^{des}(\mathbf{r},\omega) \\ \Im(\mathbf{p}_b^{des}(\mathbf{r},\omega) \end{bmatrix}, \tag{3.3}$$

with $\Re(\cdot)$ and $\Im(\cdot)$ representing the real and imaginary part of a complex number, respectively.
We will use $\mathcal{U}$ to refer to a series of nested functions that represent our Neural Network, defined as

$$\mathcal{U}(\cdot) = \bigcirc_{i=1}^{I} f_i = f_I \circ \cdots \circ f_1. \tag{3.4}$$

Thus, the solution of our system can be expressed by

$$\mathbf{d}_{dlpm} = \mathcal{U}(\tilde{\mathbf{p}}_b^{des}), \tag{3.5}$$

where $\mathbf{d}_{dlpm}$ is a vector containing the driving signals that minimise the error between the desired and estimated pressure fields.

3.2. Deep Learning for Multi-zone Sound Field Synthesis

In this section we will present the model for multi-zone synthesis. A first part will be dedicated to the depiction of the network architecture. The second part of this section will describe the training procedure used for our purpose. A 2D sound field in the frequency domain can be represented by a complex matrix, whose entries represent the pressure values at specific points in the space, conveniently taken from a grid. Hence, given the success of convulotional neural networks in computer vision it is straightforward to use a pressure fields as input to a CNN.

Our model it's composed by an encoder-decoder part, with a physics inference for the calculation of the loss during training. The physic inference consists in the application of some of the acoustics principles described in 1.1, i.e. in our loss we estimate our pressure field by convolving the estimated driving functions with the loudspeaker-to-CP transfer functions.

The principle of our global system - i.e. to replicate at the output its input - makes the use of an encoder-decoder structure a suitable choice for our objective.

3.2.1. Neural Network Architecture

In the following description we present the details of the neural network architecture used to perform multi-zone sound field synthesis. As already proposed in [36], we use only real values as inputs, by concatenating the real and imaginary part. We adopt a Neural Network that follows the basis of an encoder-decoder structure. The encoder takes as input the pressure field and outputs with the bottle-neck layer high-level features. The decoder takes as input the high-level features from the encoder and learns the optimal driving function.

In the above and following descriptions, we use the term *decoder* to refer to the part after the bottle-neck layer, and *encoder* for the previous part. The structure is shown in tables 3.1 3.2, and Figs 3.1 3.2.

Layer Type	Output	Parameters	Filters/Units	Kernel	Stride	Activation
Input	[(450, 64, 1)]	0				
Conv2D	[(225, 32, 32)]	230720	32	3×3	2×2	PReLU
BN	[(225, 32, 32)]	128				
Conv2D	[(225, 32, 32)]	239648	32	3×3	1×1	pReLU
Conv2D	[(113, 16, 64)]	134208	64	3×3	2×2	PReLU
BN	[(113, 16, 64)]	256				
Conv2D	[(113, 16, 64)]	152604	64	3×3	1×1	PReLU
Conv2D	[(57, 8, 128)]	134208	128	33×3	2×2	PReLU
BN	[(57, 8, 128)]	512				
Conv2D	[(57, 8, 128)]	205952	128	3×3	1×1	PReLU
Conv2D	[(29, 4, 256)]	324864	256	3×3	2×2	PReLU
BN	[(29, 4, 256)]	1024				
Conv2D	[(29, 4, 256)]	619776	256	3×3	1×1	PReLU
Conv2D	[(15, 2, 512)]	1195520	512	3×3	2×2	PReLU
BN	[(15, 2, 512)]	2048				
Conv2D	[(15, 2, 512)]	2375168	512	3×3	1×1	PReLU
Flatten	[(15360)]	0				
Dense	[(8)]	122888	8			
Reshape	[(4, 2, 1)]	0				

Table 3.1: Encoder Architecture and Parameters

Layer Type	Output	Parameters	Filters	Kernel	Stride	Activation
Conv2DT	[(8, 4, 512)]	21504	512	3×3	2×2	PReLU
Conv2DT	[(8, 4, 512)]	2376192	512	3×3	1×1	PReLU
Conv2DT	[(16, 8, 256)]	1212672	256	3×3	2×2	PReLU
Conv2DT	[(16, 8, 256)]	622848	256	3×3	1×1	PReLU
Conv2DT	[(32, 16, 128)]	360576	128	3×3	2×2	PReLU
Conv2DT	[(32, 16, 256)]	213120	128	3×3	1×1	PReLU
Conv2DT	[(64, 32, 64)]	204864	64	3×3	2×2	PReLU
Conv2DT	[(64, 32, 64)]	168000	64	3×3	1×1	PReLU
Conv2DT	[(128, 64, 32)]	280608	32	3×3	2×2	PReLU
Conv2DT	[(128, 64, 32)]	271392	32	3×3	1×1	PReLU
Conv2DT	[(128, 64, 1)]	289	1	3×3	1×1	

Table 3.2: Decoder Architecture and Parameters

The encoder is composed of 10 convolutional layers having (i) 32, (ii) 32, (iii) 64, (iv) 64, (v) 128, (vi) 128, (vii) 256, $(viii)$ 256, (ix) 512, (x) 512 filters, respectively. The first layer takes as input the vector $\hat{\mathbf{p}}_b^{des}(\mathbf{r}, \omega) \in \mathbb{R}^{2M \times K}$, being K the number of frequencies used for training; while the output of the last layer is then flattened to a monodimensional vector. The bottle-neck layer, that contains the highest-level features is a dense layer composed of (2L/32)(K/32) neurons.
The decoder is composed of 10 de-convolutional layers that mirror the encoder. Thus the number of filters increases, being respectively (xi) 512, (xii) 512, $(xiii)$ 256, (xiv) 256, (xv) 128, (xvi) 128, $(xvii)$ 64, $(xviii)$ 64, (xix) 32, (xx) 32 filters, respectively. The input of the decoder is the output of the bottle-neck, reshaped as a $(2L/32)(K/32) \times 1$ tensor; the output is a tensor of shape $(2L)(K) \times 1$.

All convolutional layers have a kernel size of 3×3. Both in the encoder and decoder, all layers have Parametric ReLU (PReLU)[28] as activation function, odd layers have a stride of 2×2, have a stride of 1×1. PReLU 3.3, can be defined as

$$\epsilon_{PReLU}(z) = \max(0, z) + \alpha \min(0, z), \tag{3.6}$$

and is a variant of the ReLU activation function that allows the slope of the negative part of the function to be learned during training; this makes the activation function more

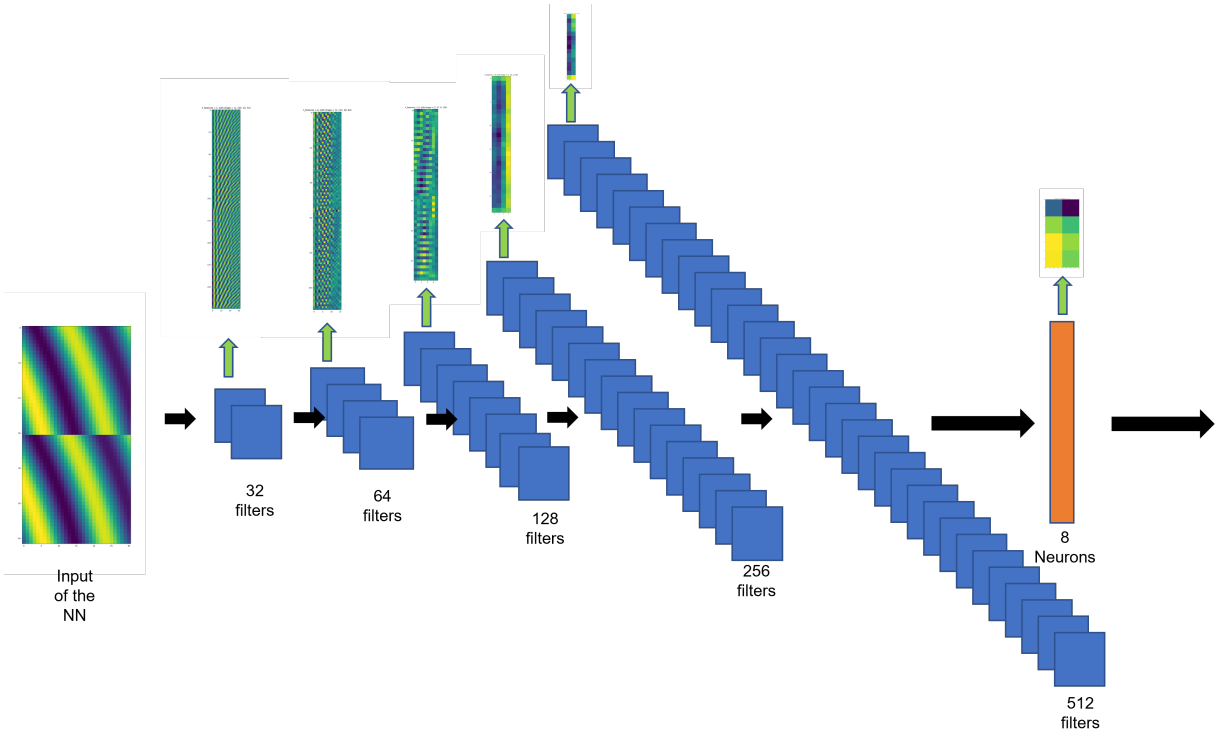


Figure 3.1: Schematic representation of the Encoder. For simplicity we represent only the layers with stride 2×2, the reshape layer and their outputs. The Encoder takes as input the concatenation of the real and imaginary part and outputs high-level features.

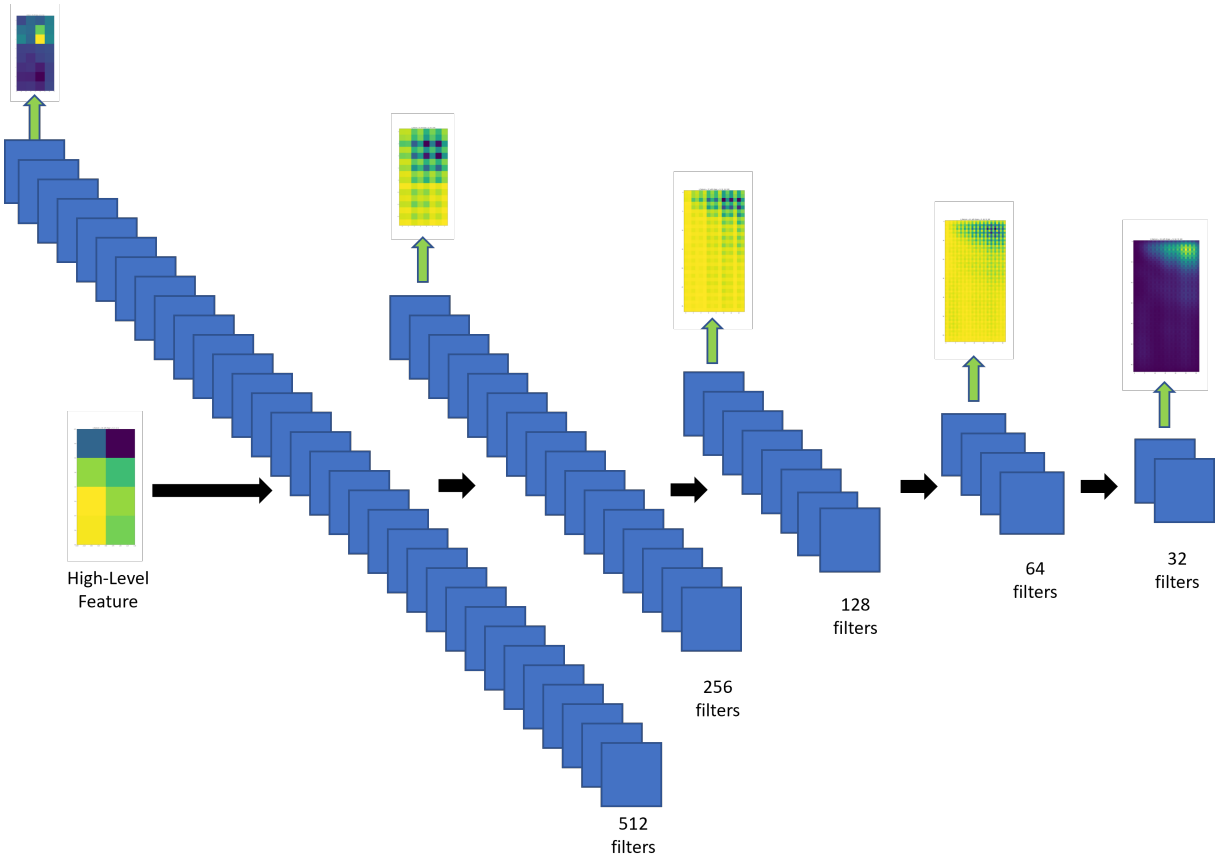


Figure 3.2: Schematic representation of the Decoder. For simplicity we represent only the layers with stride 2×2 layer and their outputs. The decoder takes as input high-level features, and outputs the concatenation of the real and imaginary part of the driving function.

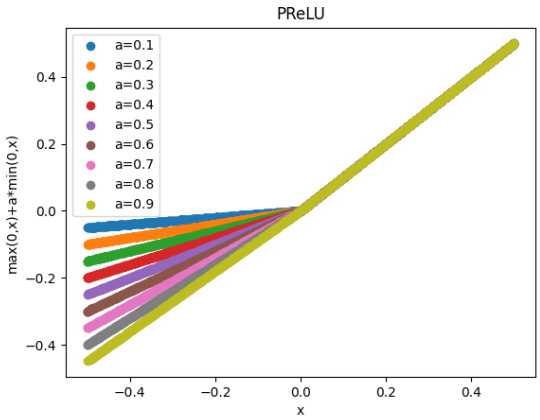


Figure 3.3: Parametric Rectified Linear Unit with various values of α

flexible and can improve the model's ability to capture complex nonlinear relationships in the data.

In the encoder's odd layers, kernels are regularised using the *L2* regularisation [45]. Every convolutional layer's input is zero-padded evenly in the left/right and up/down parts. With this configuration, in layers with $stride = (1 \times 1)$, the output has the same size as the input. At the output of the bottle-neck layer is applied a *L1L2* regularisation [23].

3.2.2. Procedure

In this section we present a formal description of how the proposed model can be trained. Terming as $\mathcal{S}$, the set of virtual sources placed outside the listening environment $\mathcal{Q}$, and using the subscript cp to underline that we consider only the M control points, we use transfer functions to compute the pressure field $\mathbf{p}^{des}_{b,cp}(\mathbf{r}_{cp})$ emitted by each virtual source $\mathbf{r}_s \in \mathcal{S}$, i.e.

$$\mathbf{p}^{des}_{b,cp}(\mathbf{r}_{cp}) = \mathbf{G}(\mathbf{r}_{cp}|\mathbf{r'}_s),\ cp = 1, \dots M, \text{ and } s \in \mathcal{S}. \tag{3.7}$$

For simplicity, in the above definition, we omit the frequency argument ω. This omission is used to represent that the computation is for every ω_k frequencies, with $k = 1, \dots, K$. This notation will be maintained during the following description. Since we are considering only control points, (3.5) can hence be reformulated as

$$\mathbf{d}_{dlpm} = \mathcal{U}(\tilde{\mathbf{p}}^{des}_{b,cp}). \tag{3.8}$$

As shown in section 3.2.1, the output of our DNN is a real tensor $\mathbf{d}_{dlpm} \in \mathbb{R}^{2L\times K\times 1}$. With this shape is possible to characterise the first dimension of the tensor as being the concatenation of real and imaginary part. This characterisation allow us to reorganise our model in a complex formulation as

$$\mathbf{d}^{C}_{dlpm,l} = \mathbf{d}_{dlpm,l} + j\mathbf{d}_{dlpm,L+l},\ l = 1,\dots,L \tag{3.9}$$

being $\mathbf{d}^{C}_{dlpm}$ the complex reformulation of our estimated driving signal, and j the imaginary unit.

With this complex formulation of our driving signal, we can now use the vectorised equation for discrete Single Layer Potential (3.1), to compute our estimated pressure field at the control points as

$$\mathbf{p}^{est}_{b,cp}(\mathbf{r}_{cp}) = \sum_{l=1}^{L} \mathbf{G}(\mathbf{r}_{b,cp}|\mathbf{r}'_l)\mathbf{d}^{C}_{dlpm}(\mathbf{r}'_l), \tag{3.10}$$

$$\mathbf{p}^{est}_{d,cp}(\mathbf{r}_{cp}) = \sum_{l=1}^{L} \mathbf{G}(\mathbf{r}_{d,cp}|\mathbf{r}'_l)\mathbf{d}^{C}_{dlpm}(\mathbf{r}'_l). \tag{3.11}$$

In the last equations we can see hoe once we have a driving function, the pressure zone of the pressure field to be synthesised depends only on the transfer function $\mathbf{G}$.

The derivation of our loss function is based the Mean Absolute Error, i.e.

$$MAE = \frac{(|\mathbf{p}^{des}_{s} - \mathbf{p}^{est}_{s}|)}{\sum_s s}, s \in \mathcal{S}, \tag{3.12}$$

where $\mathbf{p}_s$ is used to represent a generic pressure field produced by a virtual source s.

Using $\mathcal{L}(\cdot,\cdot)$ to refer to a loss function, we describe our loss as

$$\begin{aligned}\mathcal{L}_{MAE}(\mathbf{p}^{des}_{cp}, \mathbf{p}^{est}_{cp}) = (\lambda_{abs}(||\mathbf{p}^{des}_{b,cp}| - |\mathbf{p}^{est}_{b,cp}||) + (|\angle\mathbf{p}^{des}_{b,cp} - \angle\mathbf{p}^{est}_{b,cp}|))+ \\ +\lambda_d(\lambda_{abs}(||\mathbf{p}^{des}_{d,cp}| - |\mathbf{p}^{est}_{d,cp}||)\end{aligned}, \tag{3.13}$$

where the absence of the r_s is used to represent the whole batch over which the loss is computed, and λ_{abs} and λ_{dark} are two weights empirically estimated. Note that since our goal is to correctly reproduce the bright zone and only to attenuate the dark zone, we completely discarded the phase of the dark zone.
A schematic representation of the training procedure is shown in Fig. 3.4.

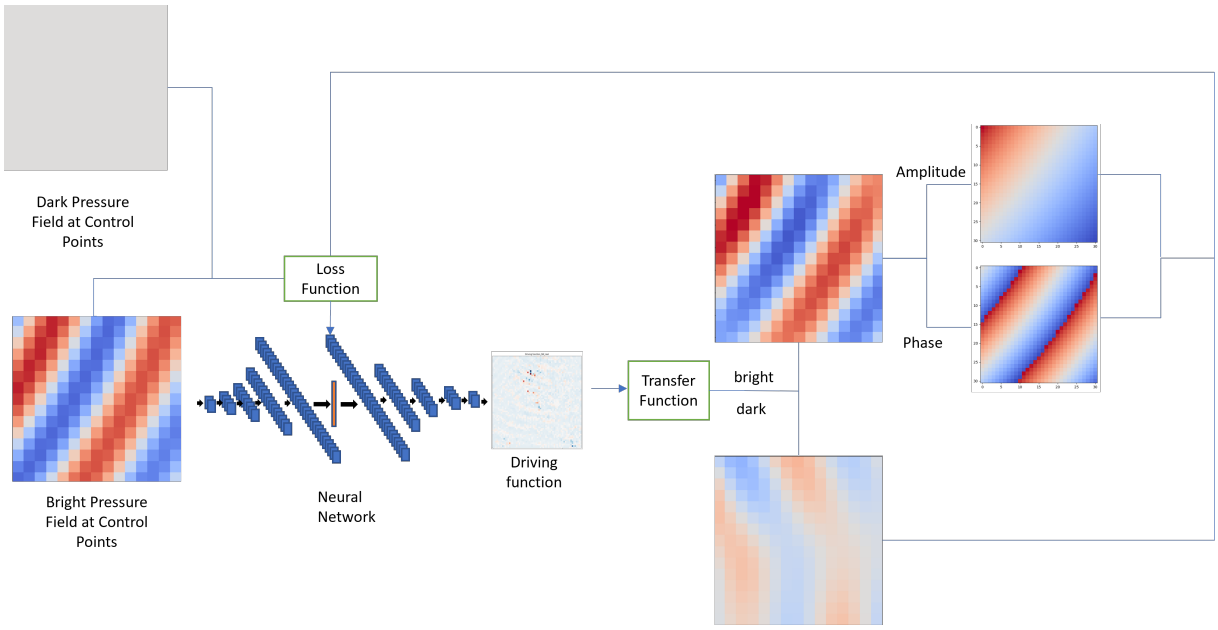


Figure 3.4: Schematic representation of the training procedure

3.3. Conclusive Remarks

In this chapter, we presented the problem formulation and the procedure used to find the optimal driving function to minimise error between the estimated and desired pressure fields. We described the problem of multi-zone sound field control through an array of loudspeakers. The goal of our system was to find the optimal driving vector containing the weights to be applied to each loudspeaker to reproduce our desired pressure field. Following the Pressure Matching approach, the reproduced field was estimated using a set of control points inside our regions of interest. We proposed a deep network, where the convolutional part was inspired by the structure of autoencoders, and the output was computed through a multi-layer perceptron.

Our model takes as input the concatenation of the real and imaginary parts of a pressure field, and outputs the concatenation of real and imaginary parts of a driving function. The output, is then recomposed as a complex vector and convolved to loudspeaker's Green Functions to obtain our estimated field.

Finally for the comparison of the estimated and desired pressure fields we used the mean absolute error as a loss function: we compared separately and weighted the sum of the MAEs for the bright and dark zones; furthermore, for the bright zone we compared separately the MAEs of the amplitude and the phase and weighted their sum, while for the dark zone we only used the MAE of the amplitudes.

4 | Results

In this chapter we present simulation results aiming to demonstrate the effectiveness of the proposed technique, namely MZ-DLPM, when compared with state-of-the-art methods such as Pressure Matching (PM), Amplitude Matching (AM), and Acoustic Contrast (AC).
We conduct numerical experiments in a 3D sound field to evaluate our proposed method and perform the evaluation for the pressure values in the horizontal plane where the loudspeakers are placed.

In the first section we present the metrics that will be used in order to evaluate the models. We mainly use metrics that are frequently used in the field of PSZ reproduction. In order to compare the wave fronts we also applied a metric coming from the field of computer vision.

In the second section we present the setup used to obtain our results. We firstly describe the physical layout, consisting on the considered environment and the distribution of control points and secondary sources. We then move to the virtual layout, describing how the virtual sources are placed and modeled.

The last part of the chapter is dedicated to the presentation of the obtained results and corresponding discussions and interpretations.

In the following, we'll term as *bright* the zone with high acoustic power, and *dark* or *quite* the zone with low acoustic power.

4.1. Evaluation Metrics

In this section we describe the evaluation metrics we will use to asses the performance of our system and the rationale behind each choice. We will use the subscripts *des* and *est* to refer to the pressure values taken by the desired and estimated pressure fields. The expression *ground truth* is also used to refer to the desired pressure field, and the terms *synthesised* and *predicted* will be used as alternative to "estimated".

4.1.1. Mean Squared Error

We measure the reproduction error between the desired and estimated field by applying the MSE as defined in 2.41 to their complex values. To have insights on which are the components of the pressure field that are better reproduced, we measure the MSE also for the amplitude and phase values of the two acoustic fields, separately. The three metrics are termed as MSE, MSE_{abs} and MSE_{angle}, and are defined as

$$MSE = \frac{\sum_{n=1}^{N}(\mathbf{p}_{des}(\mathbf{r}_n,\omega) - \mathbf{p}_{est}(\mathbf{r}_n,\omega))^2}{N}, \tag{4.1}$$

$$MSE_{abs} = \frac{\sum_{n=1}^{N}(|\mathbf{p}_{des}(\mathbf{r}_n,\omega)| - |\mathbf{p}_{est}(\mathbf{r}_n,\omega)|)^2}{N}, \tag{4.2}$$

$$MSE_{angle} = \frac{\sum_{n=1}^{N}(\angle\mathbf{p}_{des}(\mathbf{r}_n,\omega) - \angle\mathbf{p}_{est}(\mathbf{r}_n,\omega))^2}{N}, \tag{4.3}$$

where $|\cdot|$ represent the absolute value operator, $\angle\cdot$ the phase operator and N is the number of evaluation points.

4.1.2. Image Similarity

As described in 3.2.2 the MSE mainly provides a mean error over all locations between the two acoustic fields. Due to the square operation, a high MSE value may result from a poor perfomance locally, while performing well in the remaining spatial locations. Hence, following [19], we also compute the Structural Similarity Index Measure (SSIM) [57], which is usually applied in image processing problems. SSIM is used to quantify how much two images are similar, being 1 the value obtained in the case of two identical images. Considering two matrices $\mathbf{P}_{des}(\omega), \mathbf{P}_{est}(\omega) \in^{N\times N}$ is defined as

$$SSIM(\mathbf{P}_{des}, \mathbf{P}_{est}) = \frac{(2\mu_{\mathbf{P}_{des}}\mu_{\mathbf{P}_{est}} + c_1)(2\sigma_{\mathbf{P}_{des},\mathbf{P}_{est}} + c_2)}{(\mu^2_{\mathbf{P}_{des}} + \mu^2_{\mathbf{P}_{est}} + c_1)(\sigma^2_{\mathbf{P}_{des}} + \sigma^2_{\mathbf{P}_{est}} + c_2)}, \tag{4.4}$$

where μ is the mean of the corresponding matrix entries, σ^2 the estimate of the variance of the entries, $\sigma_{\mathbf{p}_{des},\mathbf{p}_{est}}$ is the covariance estimate between the entries $\mathbf{p}_{des}$ and $\mathbf{p}_{est}$; c_1 and c_2 are constants meant to stabilise the division for small denominators.

4.1.3. Acoustic Contrast

The Acoustic Contrast (AC), as described in 2.1.2 is the ratio between the acoustic potential energy between two considered zones. It can be applied in order to measure how the noise or the energy is distributed inside an environment, once a desired pressure field is synthesised in a determined region. For our purpose, we use the AC to compare the energy difference between the bright and dark zones. By using the definitions of acoustic potential energy (2.16), (2.17) and considering that we have a discrete representation of the pressure fields, we redefine AC as

$$AC = \frac{e_b(n)}{e_d(n)} = \frac{\sum_{n=1}^{N} \mathbf{p}_b^*(n)\mathbf{p}_b(n)}{\sum_{n=1}^{N} \mathbf{p}_d^*(n)\mathbf{p}_d(n)}, \tag{4.5}$$

where the apex * denotes the complex conjugate and the subscripts b and d refer to the bright and dark zones respectively.

4.2. Setup and Dataset Generation

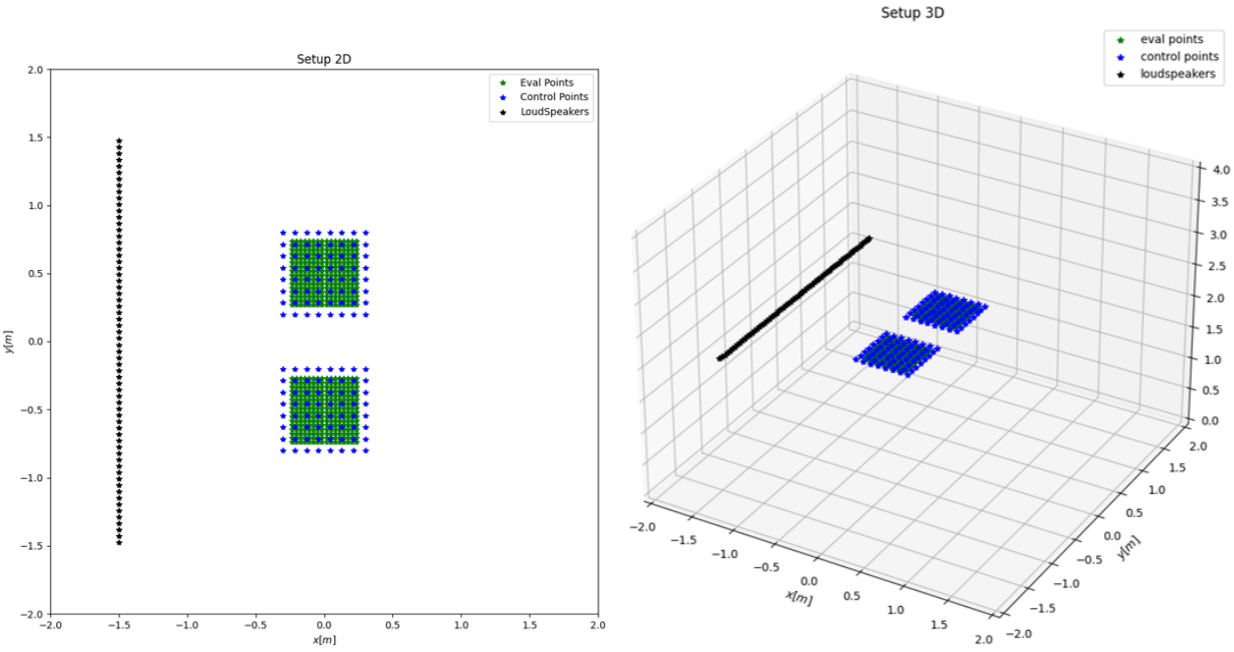
In this section we describe how we set up the system used in order to simulate the environment. In the first part we focus on the considered environment, by showing the room we use and how we position secondary sources and control points. We also show how we compute two different sets of points: a sparse set of control points for the training of the model and a dense set of evaluation points for the evaluation of our system.

The second part of this section is dedicated to description of the generation of our datasets. We show the signals used to represent our virtual sources and how we set the parameters of our model for the training procedure.

4.2.1. Reproduction System Layout

As shown in Fig 4.1 our environment is a free-field cubic room of dimensions $[-2m, 2m] \times [-2m, 2m] \times [0m, 4m]$, with the position $\mathbf{r}_0 = (0m, 0m, 2m)$ being its centre and origin. Two square target regions for the generation of the two zones with high-and-low acoustic potential energy are placed.
The bright evaluation zone is centered at $(0.0m, 0.5m, 2m)$ m and has a side of 0.5 m, while the dark evaluation zone is centered at $(0.0m, -0.5m, 2m)$ and has the same side of the bright zone.

We can define the two regions $\mathcal{A}_b$ and $\mathcal{A}_d$ with two different sets of points each. We refer



(a) (b)

Figure 4.1: Experimental setting in (a) 2D plane (b) 3D environment.

to *evaluation points* to define the points used for the evaluation of our system, i.e. the ones over which the metrics are calculated. This point distribution is dense, i.e. with a spacing of $\delta_{eval} \approx 0.02\ m$. While we refer to *control points* to define the points used for training our system, precisely the ones over which the loss is minimised. Control points are more sparse since they have a spacing of $\delta_{cp} \approx 0.05\ m$. Furthermore the zone covered by the control points is enlarged w.r.t. to the evaluation zone. In facts the control zones have the same center of evaluation zones, but their side is of 0.6 m.

To sum up the above description, the evaluation zone is composed by 512 evenly distributed points in an area of 0.25 m^2, while the control zone is composed by 128 evenly distributed points in an area of 0.36 m^2. We'll use the subscripts *eval* and *cp* to refer to evaluation and control points, respectively.

To reproduce the desired sound field we use a Uniform Linear Array (ULA) of $L = 64$ secondary sources, linearly distributed in the range $-1.5m \times [-1.5m, 1.5m] \times 2m$. With this configuration, the spacing between secondary sources is $\delta_l \approx 0.05m$. Hence, since - as mentioned in Ch. 1 - closed-cabinet loudspeakers behave similarly to point sources, we can model our secondary sources using the Green's Function (2.5).

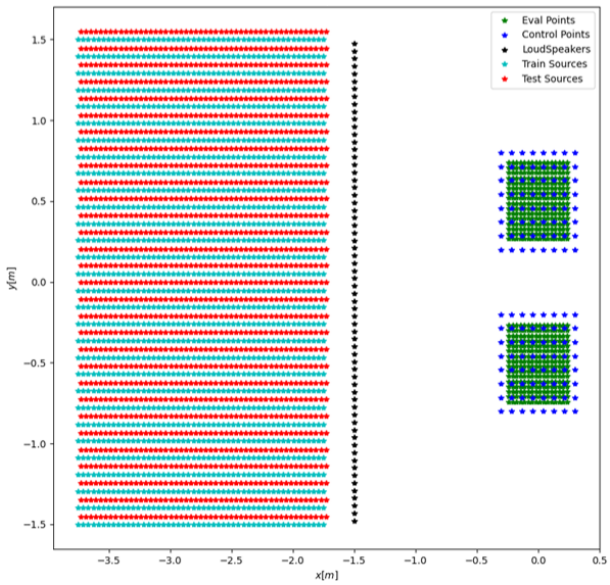


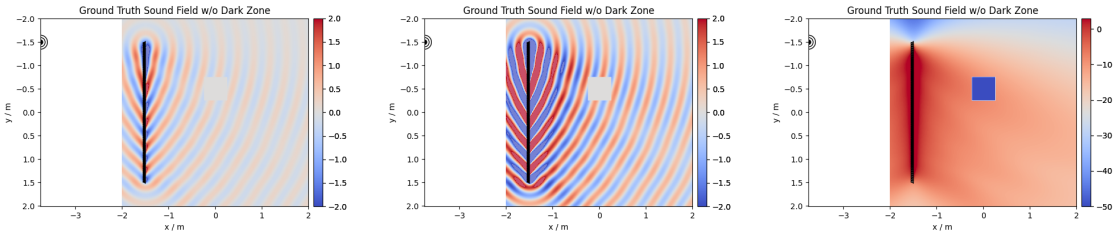
Figure 4.2: Virtual sources distribution for the generation of train set and test set

4.2.2. Training and Test Sets

In order to train the network we consider a dataset of $\mathcal{S}$ virtual sources, with a cardinality of $\#\mathcal{S} = 1500$. We then separate our dataset in two datasets $\mathcal{S}_{train}$ and $\mathcal{S}_{val}$ for training and validation, respectively. These two datasets have a cardinality of $\#\mathcal{S}_{train} = 1200$ and $\#\mathcal{S}_{val} = 300$. The $\mathcal{S}_{train}$ and $\mathcal{S}_{val}$ sets are generated by randomly sampling from the whole dataset. The sources of $\mathcal{S}$ are placed in a rectangular area covering the range $[-3.75m, -1.75m] \times [-1.5m, 1.5m] \times 2m$, with a spacing of 0.04 m along the x axis and a spacing of 0.1 m along the y axis. A last dataset $\mathcal{S}_{test}$ of cardinality $\#\mathcal{S}_{test} = 1500$ is created by shifting the $\mathcal{S}$ by $0.02m$ on the x axis and by 0.05 on the y axis. The latter is the dataset that be we use to test our system and compare the results with other methods. A representation of the above description is in Fig. 4.2.

The signals emitted by the virtual sources are sinusoids, with $K = 64$ frequency values linearly spaced between 23.4375 Hz and 1500 Hz.
Since we consider a free-field environment, we can model also the transfer functions as spatio-temporal impulses with the Green's Function defined in (2.5), as done with the secondary sources.

We train our model for 5000 epochs and apply early stopping with a patience of 100 epochs, tracking the value of the loss of $\mathcal{S}_{val}$. Approximately the overall training lasts for ≈ 500 epochs. We adopt the Adaptive Moment (Adam) optimiser [31], that is a stochastic gradient descent method that adapts its learning rate during the training phase. Finally, we initialise the learning rate $lr = 0.001$ and set the parameters for the loss (3.13) $\lambda_{abs} = 25$



(a) (b) (c)

Figure 4.3: Pressure field emitted by virtual point source at position $\mathbf{r}_{\hat{\mathbf{s}}}$ (a) without normalisation, (b) normalised w.r.t. the amplitude at position $\mathbf{r}_0$ and (c) expressed in dB.

and $\lambda_{dark} \approx 1$.

4.3. Discussion

In this section we qualitatively and numerically compare the proposed technique with some of the state-of-the-art approaches described in Ch. 2, namely Pressure Matching, Acoustic Contrast Control and Amplitude Matching. For the qualitative comparison we show and discuss the resulting pressure fields obtained by each method. For the numerical comparison we use the metrics described in 4.1 - i.e. MSE 4.1.1, SSIM 4.1.2 and AC 4.1.3 - to show the behaviour of each technique as a function of frequency and position.

Before presenting the results related to the whole dataset, we show an example of the generated sound fields for a single virtual source located outside our considered reproduction zone at position $\mathbf{r}_{\hat{\mathbf{s}}} = [-3.75m, 1.5m, 2m]$ emitting a spherical wave at frequency $f_k = 961$ Hz, as shown in Fig. 4.3.

In Fig. 4.4, we can see the characteristics of each of the considered techniques as real pressure fields, while in Fig. 4.5 are shown the amplitude and phase distribution of the same acoustic scenes. PM tends to accurately confine the two evaluation zones and focuses the reproduction only on those areas. Amplitude matching has a similar behavior to PM; despite its capability to reproduce a bright zone with a higher acoustic potential energy, the waves reproduced tend to be more similar to plane waves. ACC is capable of achieving a great acoustic contrast, by generating a bright area with a very high acoustic potential energy. However, the pressure waves reproduced are completely different w.r.t. the desired ones. Finally, the proposed method is able achieve an acoustic contrast similar to the ACC while maintaining the directionality of the desired pressure field. It's interesting to notice

how, despite being out of the scope of our work, the proposed technique is able to maintain a point-source-like phase distribution along the whole environment.

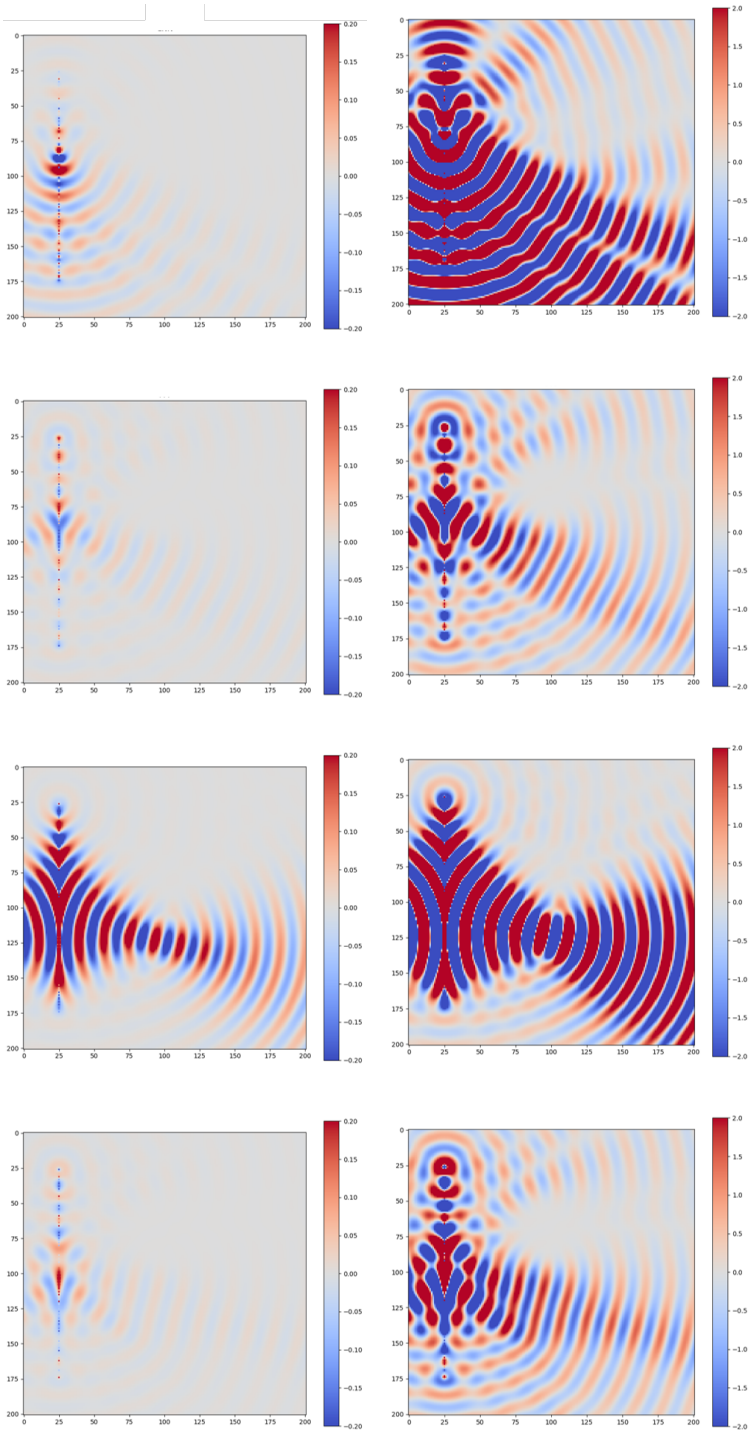
For the numerical evaluation, firstly we use the MSE, MSE_{abs} and MSE_{angle} averaged over all test positions as function of frequency f_k for all the evaluation points $\mathcal{A}$ 4.6 comprising both evaluation points of the bright zone $\mathcal{A}_b$ and the evaluation points of the dark zone $\mathcal{A}_d$ 4.7.

It is clear how the proposed method is capable of being more accurate. As expected, between the other methods, the one that best performs in terms reproduction error is the PM when we consider the complex pressure field.
If we further analyse the results, we can see that our method outperforms the other techniques both in terms of error of the amplitude and phase. Also in this case is expected for the Amplitude Matching to perform better when considering only the amplitude w.r.t. considering the whole values. Acoustic Contrast Control is by far the one that creates a major distortion in the reproduced sound field.
It's noteworthy how by separately optimising the amplitude and the phase, our approach is able to achieve great results in terms of phase accuracy.

In the same way we show the MSE, MSE_{abs} and MSE_{angle} averaged over all frequencies as a function of the distance to the line that connects the centres of $\mathcal{A}_b$ $(0m, 0.5m, 2m)$ and $\mathcal{A}_d$ $(0m, -0.5m, 2m)$. Also this simulation is performed for the evaluation points of the bright and dark zones separately 4.13 and together 4.12.

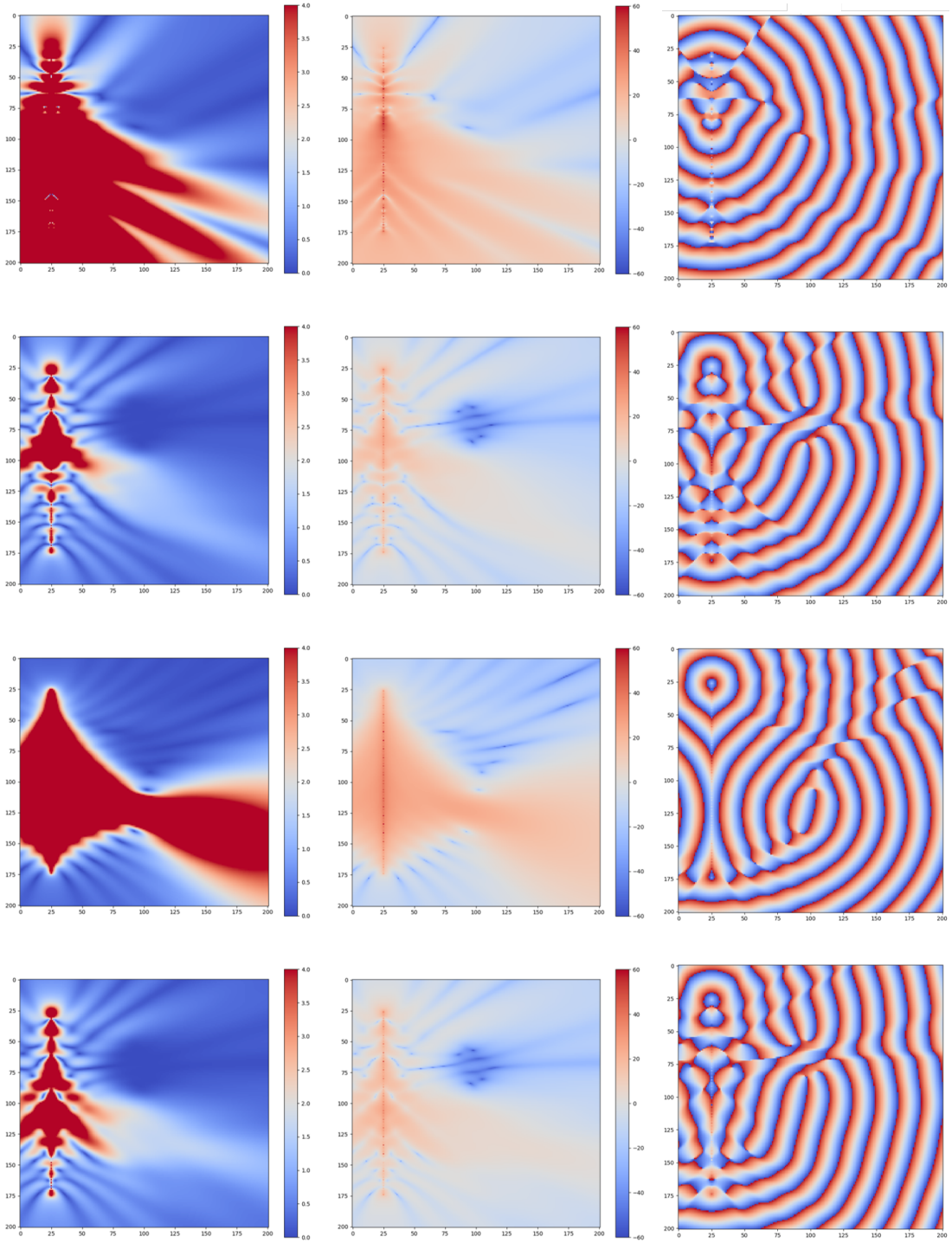
Also in this situation our approach has a reproduction error clearly lower w.r.t. the other approaches. It is interesting to note how all techniques have the same performance trend. There could be two complementary reasons to explain this trend.
Firstly the tendency is more pronounced in the dark zone, which could mean that is harder in general to attenuate a zone when the pressure values in the bright zone are higher, i.e. when it's necessary a greater contrast: farther virtual sources have more space to decrease their amplitude and could be that the values of the pressure field are already low when the pressure wave arrives to the bright zone.
However this trend is present also in the bright zone. This could suggest, that for multi-zone systems is easier to reproduce plane waves, w.r.t. spherical waves: even though we used points sources as virtual sources, the waves coming farther w.r.t. the origin will tend to flatten.
Also with this representation is straightforward how the phase benefits by the separated optimisation.

The SSIM results are computed only for the bright zone. The average over all test positions



(a) (b)

Figure 4.4: In top-bottom order, acoustic fields of MZ-DLPM, PM, ACC and AM (a) without and (b) with normalisation w.r.t. the amplitude at position $\mathbf{r}_0$



(a) (b) (c)

Figure 4.5: In top-bottom order, acoustic fields of MZ-DLPM, PM, ACC and AM of (a) amplitude distribution of the acoustic fields, (b) amplitude distribution of the acoustic fields expressed in dB, (c) phase distribution of the acoustic fields.

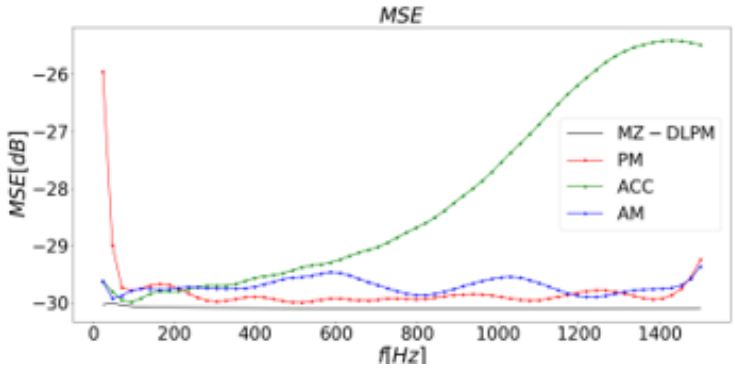
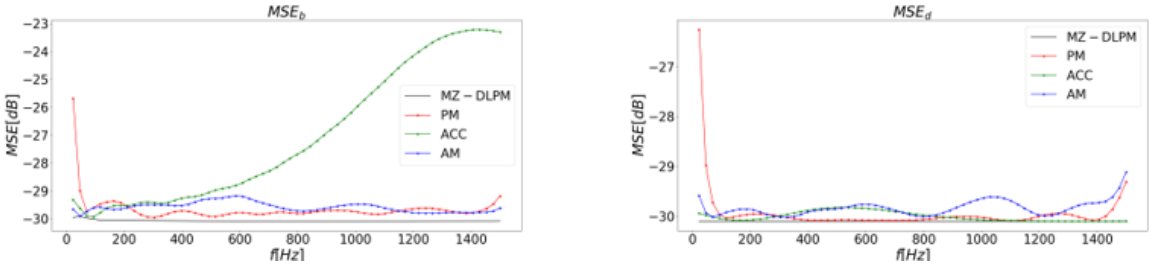


Figure 4.6: MSE as a function of frequency in $\mathcal{A}$



(a) (b)

Figure 4.7: MSE as a function of frequency in (a) $\mathcal{A}_b$ and (b) $\mathcal{A}_d$

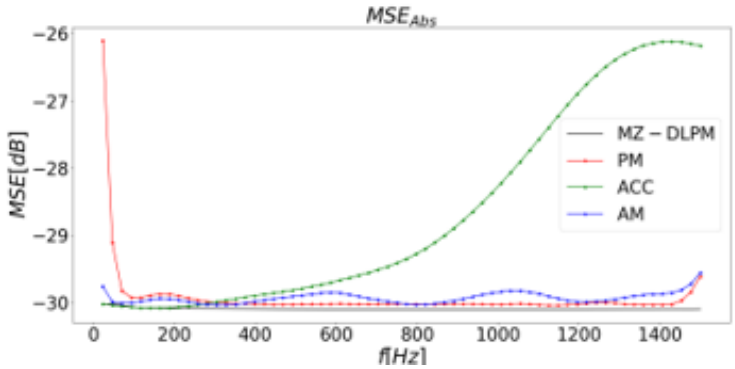
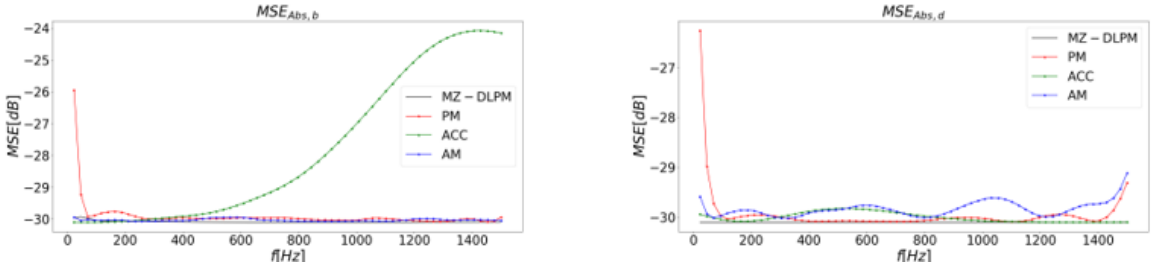


Figure 4.8: MSE of the absolute values as a function of frequency in $\mathcal{A}$



(a) (b)

Figure 4.9: MSE of the absolute values as a function of frequency in (a) $\mathcal{A}_b$ and (b) $\mathcal{A}_d$

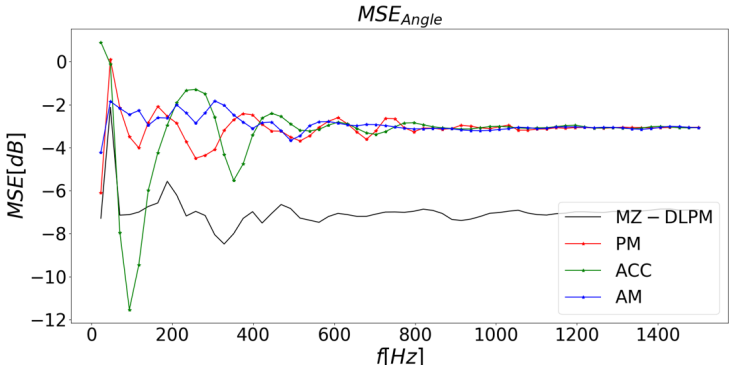
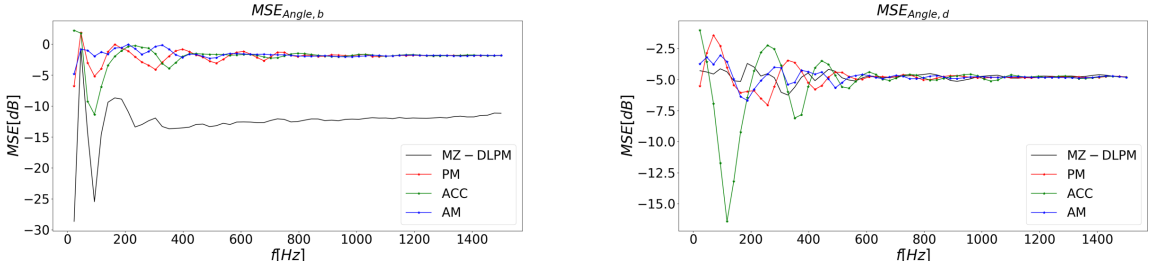


Figure 4.10: MSE of the phase as a function of frequency in $\mathcal{A}$



(a) (b)

Figure 4.11: MSE of the phase as a function of frequency in (a) $\mathcal{A}_b$ and (b) $\mathcal{A}_d$

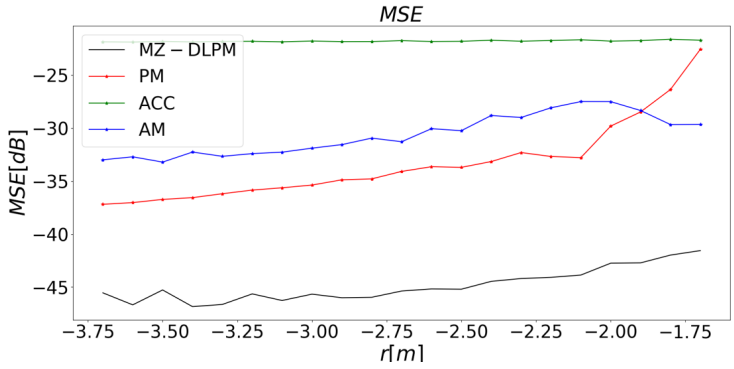
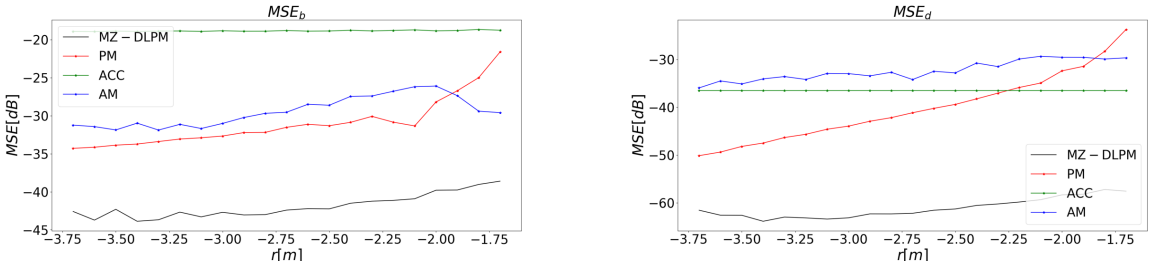


Figure 4.12: MSE as a function of the position in the x axis in $\mathcal{A}$



(a) (b)

Figure 4.13: MSE as a function of the position in the x axis in (a) $\mathcal{A}_b$ and (b) $\mathcal{A}_d$

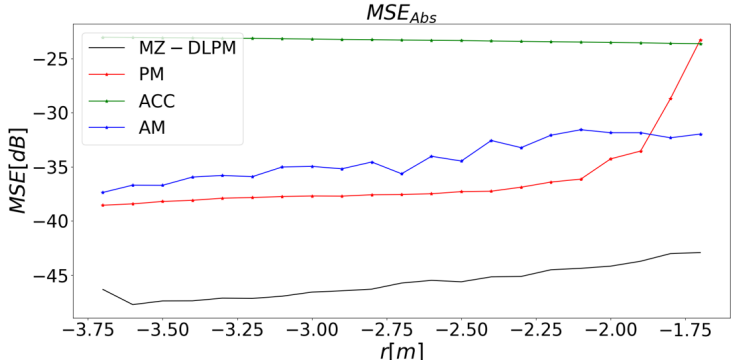
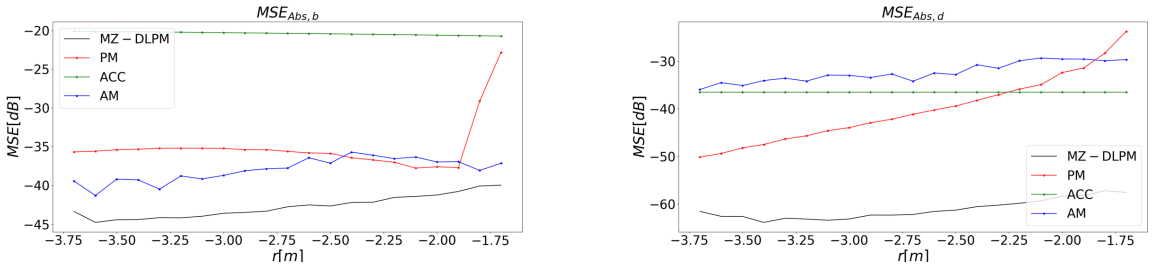


Figure 4.14: MSE of the absolute values as a function of the position in the x axis in $\mathcal{A}$



(a) (b)

Figure 4.15: MSE of the absolute values as a function of the position in the x axis in (a) $\mathcal{A}_b$ and (b) $\mathcal{A}_d$

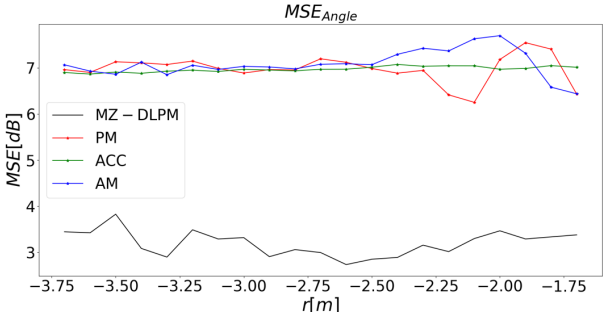
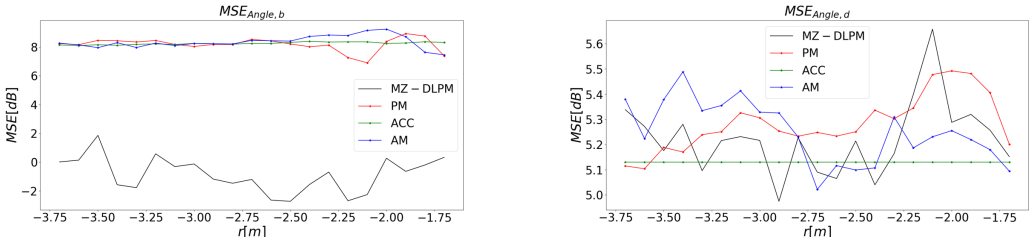
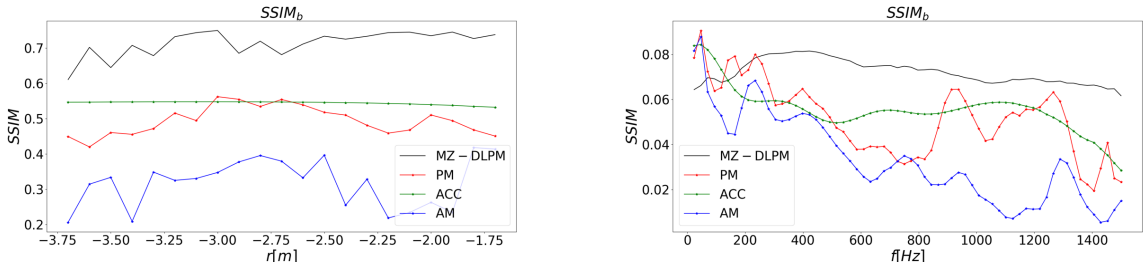


Figure 4.16: MSE of the phase as a function of the position in the x axis in $\mathcal{A}$



(a) (b)

Figure 4.17: MSE of the phase as a function of frequency in (a) $\mathcal{A}_b$ and (b) $\mathcal{A}_d$



(a) (b)

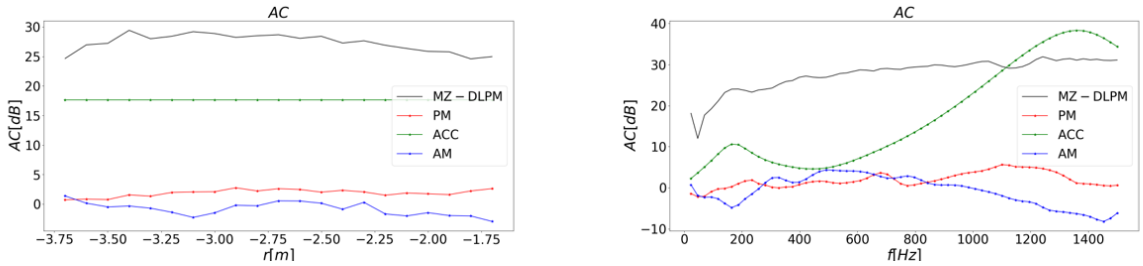
Figure 4.18: SSIM as a function of (a) the position in the x axis and (b) frequency in $\mathcal{A}_b$

and over all frequencies are shown in Fig. 4.18.

In this case there is a strong trend, that shows how performance degrade as we increase the frequency. Since we are below aliasing condition and we are considering only the bright zone, it is possible that this tendency is correlated with the trend of the reproduction error of the phase in the bright zone 4.11a. By considering a human empirical evaluation of two pressure fields coming from the same position, what could be perceived in a case like ours - where the amplitude error is small - is the difference in the phase. Since SSIM aims to evaluate the similarity between two images, this could be the reason for this trend. For what concerns its dependency from the position also here there is a small tendency of reaching a better performance as we come closer to the origin.
At first sight this behaviour seems in contrast w.r.t. to the one of the MSE 4.13, however SSIM is less bounded by the single values of the pressure fields and more correlated to their statistical distribution. This means that the similarity between the distribution of the pressure field values increases as we go closer to the origin.
Also this metric shows how our methods tends to better represent the desired pressure field in the bright zone.

Finally Fig. 4.19 show the Acoustic Contrast computation as a function of frequency and position, respectively.

The proposed technique achieves a great result in terms of acoustic contrast, being able to surpass ACC for low frequencies, and achieving by far a higher contrast w.r.t. AM and PM. However as we approach higher frequencies the ACC outperforms our method.



(a) (b)

Figure 4.19: AC as a function of (a) the position in the x axis and protect(b) frequency in $\mathcal{A}$

4.4. Conclusive Remarks

In this chapter, we presented the configuration used perform our simulations. We used an anechoic shoebox room ($4 \times 4 \times 4$) to represent a free field. Inside this environment we placed 256 control points in the two regions $\mathcal{A}_b$, $\mathcal{A}_d$ under study. To reproduce the estimated sound fields we used ULA of secondary sources.
We then described the procedure used for the generation of our datasets. We used two different datasets for training and testing composed by 1500 virtual sources each emitting sinusoidal signals at 64 different frequencies.
In the last part of the chapter we show how our method performs compared with other state-of-the-art systems. For the comparison we used two classic metrics used for the evaluation of multi-zone problems, namely MSE and AC, and also used a metric used in computer vision that for our scope takes into consideration the statistic distribution of the pressure values, i.e. SSIM.
From the comparison is evident, how our method outperforms all the other approaches in all the tested metrics: is capable of maintaining a small reproduction error, while also achieving a high acoustic contrast and a distribution similar w.r.t. the ground truth.

It's noteworthy how approaches described in the literature tend always to choose between the trade-off accuracy-vs-contrast. When it comes to accuracy, PM and and WMM are the two methods that till this day achieved the best results. While for the acoustic contrast, the ACC is considered as an upper-bound to refer to, since it's an algorithm that aims to only maximise it. Hence, the fact that our method has in most cases simultaneously both a greater accuracy w.r.t. PM and a greater AC w.r.t. ACC is a great achievement.

5 | Conclusions and Future developments

In this thesis we have presented a technique for multi-zone sound field synthesis through a deep neural network. Our method aims to synthesise a desired pressure field over two target regions using multiple secondary sources, precisely a Uniform Linear Array (ULA) of loudspeakers. We characterised two target regions with high and low acoustic potential energy, termed as *bright* and *dark*, respectively. Our goal was to find the optimal driving function to apply to the loudspeakers, through a Deep Learning-based optimisation. Specifically, for our purpose we retrieved the desired driving signals by feeding the ground truth sound field at a series of control points into an Encoder-Decoder-structured Convolutional Neural Network. Since we didn't have a set of ground truth driving signals, we first convolved the output of our model with the acoustic transfer function between the loudspeaker positions and the control points, obtaining our estimated pressure field. We then computed the loss between the ground truth and estimated sound field at control points by weighting separately the bright and dark zone, and by weighting separately the amplitude and phase of the bright zone.
Since the dark region was modeled as matrix of near-zero values, we omitted its points from the input of the CNN, as it wouldn't give any discriminative information for learning purposes. For the same reason, in the loss function, we only considered its amplitude, since adding its phase information would over-complicate the training procedure without necessarily improving the performance.

We compared the proposed technique with other state-of-the-art methods for multi-zone sound field synthesis, namely Pressure Matching (PM), Acoustic Contrast Control (ACC), and Amplitude Matching (AM). Results demonstrate the effectiveness of the proposed method and the ability to overcome the trade-off between accuracy of the reproduction in the bright zone and AC between the acoustic potential energy of the two target regions. Precisely our method was able to achieve an acoustic contrast at the same level of the ACC method, that is usually by far the best-performing technique for such metric, and simultaneously maintain a reproduction error lower then PM and AM.

Future works could include noise and reverberation into the environment. A further study could aim to analyse the behaviour of the proposed system as we change the number and/or position of control points. Also how a reduction of the number of loudspeakers impacts on the performance could be of interest. Finally, the proposed method could be tested for the synthesis of more then two zones.

Bibliography

[1] T. Abe, S. Koyama, N. Ueno, and H. Saruwatari. Amplitude matching for multi-zone sound field control. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:656–669, 2023. doi: 10.1109/TASLP.2022.3231715.

[2] J. Ahrens. The single-layer potential approach applied to sound field synthesis including cases of non-enclosing distributions of secondary sources. 01 2010. doi: 10.14279/depositonce-2663.

[3] J. Ahrens. *Analytic Methods of Sound Field Synthesis*. T-Labs Series in Telecommunication Services. Springer Berlin Heidelberg, 2012. ISBN 9783642257421. URL https://books.google.it/books?id=z8WWkMcjGp8C.

[4] J. Ahrens and S. Spors. An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions. *Acta Acustica United With Acustica*, 94:988–999, 2008.

[5] T. Ajdler, L. Sbaiz, and M. Vetterli. The plenacoustic function and its sampling. *Trans. Sig. Proc.*, 54(10):3790–3804, oct 2006. ISSN 1053-587X. doi: 10.1109/TSP.2006.879280. URL https://doi.org/10.1109/TSP.2006.879280.

[6] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, 2017. doi: 10.1109/PlatCon.2017.7883728.

[7] J. Benesty, M. M. Sondhi, and Y. A. Huang. *Springer Handbook of Speech Processing*. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 3540491252.

[8] J. Benesty, M. Christensen, and J. Jensen. *Signal enhancement with variable span linear filters*, pages 1–172. Springer Topics in Signal Processing. 2016. doi: 10.1007/978-981-287-739-0_1.

[9] A. Berkhout, D. Vries, and P. VOGEL. Acoustic control by wave field synthesis. *J.Acoust.Soc.Am.*, 93:2764–2778, 05 1993. doi: 10.1121/1.405852.

[10] T. Betlehem and T. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. *The Journal of the Acoustical Society of America*, 117:2100–11, 05 2005. doi: 10.1121/1.1863032.

[11] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala. Personal sound zones: Delivering interface-free audio to multiple listeners. *IEEE Signal Processing Magazine*, 32:81–91, 2015.

[12] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro. Rendering of directional sources through loudspeaker arrays based on plane wave decomposition. In *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, pages 013–018, 2013. doi: 10.1109/MMSP.2013.6659256.

[13] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro. Model-based acoustic rendering based on plane wave decomposition. *Applied Acoustics*, 104:127–134, 03 2016. doi: 10.1016/j.apacoust.2015.10.010.

[14] 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, 2019.

[15] Y. Cai, M. Wu, and J. Yang. Sound reproduction in personal audio systems using the least-squares approach with acoustic contrast control constraint. *The Journal of the Acoustical Society of America*, 135 2:734–41, 2014.

[16] C. Chen, R. Gao, P. Calamia, and K. Grauman. Visual acoustic matching, 2022.

[17] J.-W. Choi and Y.-H. Kim. Generation of an acoustically bright zone with an illuminated region using multiple sources. *The Journal of the Acoustical Society of America*, 111:1695–700, 05 2002. doi: 10.1121/1.1456926.

[18] G. Ciaburro and G. Iannace. Acoustic characterization of rooms using reverberation time estimation based on supervised learning algorithm. *Applied Sciences*, 11:1661, 02 2021. doi: 10.3390/app11041661.

[19] L. Comanducci, F. Antonacci, and A. Sarti. A deep learning-based pressure matching approach to soundfield synthesis. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5, 2022. doi: 10.1109/IWAENC53105.2022.9914712.

[20] L. Comanducci, F. Antonacci, and A. Sarti. Synthesis of soundfields through irregular loudspeaker arrays based on convolutional neural networks, 2022.

[21] E. Corteel. On the use of irregularly spaced loudspeaker arrays for wave field synthesis, potential impact on spatial aliasing frequency. 2006.

[22] j. daniel. spatial sound encoding including near field effect: introducing distance coding filters and a viable, new ambisonic format. *journal of the audio engineering society*, may 2003.

[23] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2009.01.002. URL `https://www.sciencedirect.com/science/article/pii/S0885064X0900003X`.

[24] P. Fellgett. Ambisonic reproduction of directionality in surround-sound systems. *Nature*, 252:534–538, 1974.

[25] S. Gao, J. Lin, X. Wu, and T. Qu. Sparse dnn model for frequency expanding of higher order ambisonics encoding process. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1124–1135, 2022.

[26] M. J. Gerzon. Periphony: With-height sound reproduction. *Journal of The Audio Engineering Society*, 21:2–10, 1973.

[27] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[29] W. Jin and W. Kleijn. Multizone soundfield reproduction in reverberant rooms using compressed sensing techniques. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4728–4732, 2014.

[30] 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.

[31] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[32] O. F. Kirkeby, P. A. Nelson, F. Orduña-Bustamante, and H. Hamada. Local sound field reproduction using digital signal processing. *Journal of the Acoustical Society of America*, 100:1584–1593, 1996.

[33] S. Koyama, G. Chardon, and L. Daudet. Optimizing source and sensor placement

for sound field control: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:696–714, 2020.

[34] S. Koyama, T. Amakasu, N. Ueno, and H. Saruwatari. Amplitude matching: Majorization–minimization algorithm for sound field control only with amplitude constraint. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 411–415, 2021. doi: 10.1109/ICASSP39728.2021.9414855.

[35] S. Koyama, K. Kimura, and N. Ueno. Sound field reproduction with weighted mode matching and infinite-dimensional harmonic analysis: An experimental evaluation, 2021.

[36] M. S. Kristoffersen, M. B. Møller, P. Martínez-Nuevo, and J. Østergaard. Deep sound field reconstruction in real rooms: Introducing the isobel sound field dataset, 2021.

[37] T. Lee, J. Nielsen, J. R. Jensen, and M. Christensen. A unified approach to generating sound zones using variable span linear filters. 04 2018. doi: 10.1109/ICASSP.2018.8462477.

[38] Q. Lei, H. Chen, J. Hou, L. Chen, and L. Dai. Deep neural network based regression approach for acoustic echo cancellation. *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*, 2019.

[39] N. Liu, H. Chen, K. Songgong, and Y. Li. Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays. *The Journal of the Acoustical Society of America*, 149:1069–1084, 02 2021. doi: 10.1121/10.0003445.

[40] Y. Long, Y. Li, S. Wei, Q. Zhang, and C. Yang. Large-scale semi-supervised training in deep learning acoustic model for asr. *IEEE Access*, PP:1–1, 09 2019. doi: 10.1109/ACCESS.2019.2940961.

[41] L. Lu, K.-L. Yin, R. C. de Lamare, Z. Zheng, Y. Yu, X. Yang, and B. Chen. A survey on active noise control techniques – part i: Linear systems, 2021.

[42] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013.

[43] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang. Self-supervised generation of spatial audio for 360 video, 2018.

[44] P. A. Nelson. Active control of acoustic fields and the reproduction of sound. *Journal of Sound and Vibration*, 177:447–477, 1994.

[45] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance.

In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 78, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015435. URL https://doi.org/10.1145/1015330.1015435.

[46] M. Ochmann, M. Vorländer, and J. Fels, editors. *Proceedings of the 23rd International Congress on Acoustics : integrating 4th EAA Euroregio 2019 : 9-13 September 2019 in Aachen, Germany*, Berlin, Germany, Sep 2019. 23. International Congress on Acoustics, Aachen (Germany), 9 Sep 2019 - 13 Sep 2019, Deutsche Gesellschaft für Akustik. ISBN 978-3-939296-15-7. URL https://publications.rwth-aachen.de/record/767416.

[47] m. poletti. an investigation of 2-d multizone surround sound systems. *journal of the audio engineering society*, october 2008.

[48] M. A. Poletti. Three-dimensional surround sound systems based on spherical harmonics. *Journal of The Audio Engineering Society*, 53:1004–1025, 2005.

[49] v. pulkki. virtual sound source positioning using vector base amplitude panning. *journal of the audio engineering society*, 45(6):456–466, june 1997.

[50] G. Routray, S. Basu, P. Baldev, and R. M. Hegde. Deep-sound field analysis for upscaling ambisonic signals. In *EAA Spatial Audio Signal Processing Symposium*, pages 1–6, Paris, France, Sept. 2019. doi: 10.25836/sasp.2019.14. URL https://hal.science/hal-02275176.

[51] H. Sallandt, P. Krah, and M. Lemke. Supervised learning for multi zone sound field reproduction under harsh environmental conditions, 2021.

[52] P. N. Samarasinghe, M. Poletti, S. M. A. Salehin, T. D. Abhayapala, and F. M. Fazi. 3d soundfield reproduction using higher order loudspeakers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 306–310, 2013. doi: 10.1109/ICASSP.2013.6637658.

[53] L. Shi, G. Ping, X. Shen, and M. G. Christensen. Generation of personal sound fields in reverberant environments using interframe correlation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1066–1070, 2022. doi: 10.1109/ICASSP43922.2022.9747574.

[54] S. Spors, R. Rabenstein, and J. Ahrens. The theory of wave field synthesis revisited. 1, 01 2008.

[55] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter. Spatial sound

with loudspeakers and its perception: A review of the current state. *Proceedings of the IEEE*, 101:1920–1938, 2013.

[56] N. Ueno, S. Koyama, and H. Saruwatari. Three-dimensional sound field reproduction based on weighted mode-matching method. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1852–1867, 2019. doi: 10.1109/TASLP.2019.2934834.

[57] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[58] D. Ward and T. Abhayapala. Reproduction of a plane-wave sound field using an array of loudspeakers. *IEEE Transactions on Speech and Audio Processing*, 9(6):697–707, 2001. doi: 10.1109/89.943347.

[59] E. G. Williams and J. A. Mann. Fourier acoustics: Sound radiation and nearfield acoustical holography. 1999.

[60] Y. J. Wu and T. D. Abhayapala. Theory and design of soundfield reproduction using continuous loudspeaker concept. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:107–116, 2009.

[61] Y. J. Wu and T. D. Abhayapala. Spatial multizone soundfield reproduction: Theory and design. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:1711–1720, 2011.

[62] N. Xiang and C. Landschoot. Bayesian inference for acoustic direction of arrival analysis using spherical harmonics. *Entropy*, 21, 2019.

[63] L. Zhang, X. Wang, R. Hu, D. Li, and W. Tu. Estimation of spherical harmonic coefficients in sound field recording using feed-forward neural networks. *Multimedia Tools and Applications*, 80:1–16, 02 2021. doi: 10.1007/s11042-020-09979-z.

[64] W. Zhang, T. D. Abhayapala, T. Betlehem, and F. M. Fazi. Analysis and control of multi-zone sound field reproduction using modal-domain approach. *The Journal of the Acoustical Society of America*, 140 3:2134, 2016.

List of Figures

List of Tables

Acknowledgements

Firstly, I want to express my deepest gratitude to my Advisor Prof. Fabio Antonacci for giving me the possibility to develop this project and my Co-Advisor Luca Comanducci for his invaluable patience and feedback. I also could not have undertaken this journey without all the professors that I've encountered through my Master of Science, who provided the knowledge and expertise necessary not only to develop this work but also to face the world after University. Additionally, this endeavour would not have been possible without the generous support from the ISPL, which let me use its technological resources.

I cannot avoid mentioning my family, especially my parents. Their belief in me have kept my spirits and motivation high during this process, from the first day of my Bachelor's till this moment.

I'm also grateful to my colleagues. This journey wouldn't have been the same without your company and friendship. In particular, I would like to thank Giovanni, for being almost like a second teacher to me at the beginning of this path, and Davide for reminding me every single day that I'm also an artist and not only an engineer.

Lastly, I cannot forget to thank my oldest friends, which are my second family. Thank you all for your constant support and for reminding me to take breaks and have fun when I've been stressed out. You're too many to thank individually, but this would have been a much more difficult accomplishment without you.