University of Genova

PhD in Science and Technology for Electronic and Telecommunication Engineering
Department of Electrical, Electronic and Telecommunications Engineering and Naval Architecture (DITEN)
Curriculum: Electromagnetism, Electronics, Telecommunications

Advanced algorithms for audio and image processing

by

Danilo Greco

Thesis submitted for the degree of *Doctor of Philosophy* (33° cycle 2017-2020)

November 2020

Prof. Andrea Trucco	Supervisor
Prof. Mario Marchese	Head of the PhD program

Thesis Jury:

Prof. Farid Melgani, *University of Trento*	External examiner
Prof. Claudio Sacchi, *University of Trento*	External examiner

DITEN Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture
Polytechnic School, University of Genoa

Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department
(DITEN)

Ai miei genitori

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Danilo Greco
January 2021

Acknowledgements

It has been an incredible journey. Being able to finish the PhD as a working student at my age was not at all obvious at the beginning of this adventure. But you can reach any goal strongly desired, with determination and will in spite of everything, the difficulties and obstacles that life puts in front of us every day. I wish to thank prof. Mario Marchese for the opportunity granted and prof. Andrea Trucco for being my tutor in this path. I would like to thank Marco Crocco for his constant help and support, for being an exhaustible source of inspiration and knowledge but, above all, for being a real and good friend of mine. I want to thank prof. Giulio Sandini for my period in IIT at the RBCS department. In the same way I thank Francesco Rea for his availability with me and Alberto Inuggi to share with me as a teacher a beautiful course on magnetic resonance imaging (my first professional love) for DIBRIS PhD students. I would like to thank all the members of IIT's EDL group: Marco Crepaldi, Francesco Diotalevi, Giorgio Zini, Claudio Lorini, Andrea Merello, Mirko Di Salvo. It has been a privilege to know you and to be able to work with all of you. A sincere thank also to Luca Recla, Cinzia Pellini and Anastasia Bruzzone. I also wish to thank prof. Vittorio Murino for the opportunity he gave me to be part of the PAVIS department of IIT and the guys I met here and who supported me: Noman Haleem, Jacopo Cavazza, Matteo Bustreo, Carlos Beltran, Giancarlo Paoletti. I also thank all the Italian and foreign PhD students and post docs: I have learned a lot from you in human and professional terms. I wish to thank Alicia and her family for always being close to me in these difficult years, Dario and little Irina for teaching me how to stay together and to grow with the new generations, making me a better man and person. This thesis is dedicated to my parents for everything they have always done for me and for the possibility of being what I am today: to my beloved father Luigi who passed away in 2013, leaving an absence that will never be filled and to my mother Silvana whom I thank for being close to me and that I hope to have next to me for many years more despite the times we are experiencing are really difficult and full of uncertainties. Thank you all.

Danilo

Organization of the Thesis

This PhD thesis deals with advanced algorithms for audio and video processing. Chapter 2 deals with the topic of the reconstruction of ultrasound images for applications hypothesized on a software based device through image reconstruction algorithms processed in the frequency domain. An innovative beamforming algorithm based on seismic migration is presented, in which a transformation of the input RF data is carried out and the reconstruction algorithm can evaluate a masking of the *"k-space"* of the data, speeding up the reconstruction process and reducing the computational burden. Chapter 3 deals with the topic of audio beamforning algorithms for superdirective linear arrays. Two categories of algorithms (numerical and analytical) are faced and compared, highlighting advantages and disadvantages, leading to the selection for best performance of an algorithm which contains an evaluation of the errors of the microphones, extremely close to the experimental implementation. In the Chapter 4 the problem of the evaluation of the Room Impulse Response is explored through a challenging *"blind method"* using microphones in a number greater than two. Statistics on various types of signals (real and synthetic) are presented, highlighting improvements and limitations. Chapter 5 deals with the simultaneous audio and video processing methodologies, presenting the classification results through deep learning networks over a data set of acoustic images obtained with an innovative device called *Dual Cam*. In addition to the analysis of the presented data-set, the section is closed aiming the continuation of the work using a brand new data-set of musicians playing different musical instruments, with the goal of recognizing them in an audio/video scene where they play all together, after training of the data coming from instruments played individually. Chapter 6 closes the thesis, presenting a development activity of a new *Dual Cam* POC to build-up from it a spin-off, assuming to apply for an innovation project for hi-tech start-ups (such as a SME instrument H2020) for a 50K€ grant, following the idea of the technology transfer. Finally, a new version of the device (planar microphones array) simpler and easier to use than the current one with reduced dimensions and improved technical characteristics, is simulated, opening up new interesting possibilities of development not only technical and scientific but also in terms of business fallout.

Publication List

Chapter 2

- Danilo Greco and Marco Crocco, *"Ultrasound method and system for extracting signal components relating to spatial locations in a target region in the spatial and temporal frequency domain".* Patent, 2019; https://patentimages.storage.googleapis.com/b0/61/56/03d08623b8634b/EP3425423A1.pdf

Chapter 3

- Danilo Greco and Andrea Trucco, *"Superdirective Robust Algorithms' Comparison for Linear Arrays".* Accepted paper in Acoustics 2020, 2(3), 707-718; https://doi.org/10.3390/acoustics2030038

Chapter 4

- Danilo Greco, Jacopo Cavazza and Alessio Del Bue, *"Are Multiple Cross-Correlation Identities better than just Two? Improving the Estimate of Time Differences-of-Arrivals from Blind Audio Signals".* Accepted paper in 25th International Conference on Pattern Recognition (ICPR2020) - MiCo Milano Congress Center, ITALY 10 - 15 January 2021; https://arxiv.org/abs/2010.08428

Chapter 5

- Valentina Sanguineti, Pietro Morerio, Niccolo Pozzetti, Danilo Greco, Marco Cristani, Vittorio Murino, *"Leveraging Acoustic Images for Effective Self-Supervised Audio Representation Learning".* Accepted paper in 16th European Conference on Computer Vision ECCV 2020 23-28 august 2020; https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123670120.pdf

Table of contents

List of figures

List of tables

Chapter 1

Introduction and Abstract

1.1 Objectives of the thesis and motivations of the research

The objective of the thesis is the development of a set of innovative algorithms around the topic of *beamforming* in the field of acoustic imaging, audio and image processing, aimed at significantly improving the performance of devices that exploit these computational approaches. Therefore the context is the improvement of devices (ultrasound machines and video/audio devices) already on the market or the development of new ones which, through the proposed studies, can be introduced on new the markets with the launch of innovative high-tech start-ups. This is the motivation and the *leitmotiv* behind the doctoral work carried out. In fact, in the first part of the work an innovative image reconstruction algorithm in the field of ultrasound biomedical imaging is presented, which is connected to the development of such equipment that exploits the computing opportunities currently offered nowadays at low cost by GPUs (Moore's law). The proposed target is to obtain a new pipeline of the reconstruction of the image abandoning the architecture of such hardware-based devices towards a software-based technology that opens up technological scenarios that improve clinical and application potentials on low-cost and real-time imaging methods. Compared to the approaches already in use in the literature, the innovative algorithm based on seismic migration used in ultrasound imaging environment, proposes a masking of the RF acquired data in the $k_x - \omega$ space to significantly reduce the computational burden without the introduction of image artifacts or loss of signal to noise ratio, making this approach compatible with a real application coupled with plane wave imaging which is able to significantly increase the frame rate of the acquired images, thus opening new trends in technological development and diagnostic potential. In the first part of the thesis I faced the topic of the reconstruction of ultrasound images for applications hypothesized on a software based device through image reconstruction algorithms processed in the frequency domain. An innovative beamforming algorithm based on seismic migration is presented, in which a transformation of the RF data is carried out and the reconstruction algorithm can evaluate a masking of the *k-space* of the data, speeding up the reconstruction process and reducing the computational burden. The analysis and development of the algorithms responsible for carrying out the thesis has been approached from a feasibility point in an off-line context and on the Matlab platform, processing both synthetic simulated generated data and real RF data: the subsequent development of these algorithms within of the future ultrasound biomedical equipment will exploit an high-performance computing framework capable of processing customized kernel pipelines (henceforth called 'filters') on CPU/GPU. The type of filters implemented involved the topic of Plane Wave Imaging (PWI), an alternative method of acquiring the ultrasound image compared to the state of the art of the traditional standard *B-mode* which currently exploit sequential sequence of insonification of the sample under examination through focused beams transmitted by the probe channels. The PWI mode is interesting and opens up new scenarios compared to the usual signal acquisition and processing techniques, with the aim of making signal processing in general and image reconstruction in particular faster and more flexible, and increasing importantly the frame rate opens up and improves clinical applications. The innovative idea is to introduce in an offline seismic reconstruction algorithm for ultrasound imaging a further filter, named masking matrix. The masking matrices can be computed offline knowing the system parameters, since they do not depend from acquired data. Moreover, they can be pre-multiplied to propagation matrices, without affecting the overall computational load. Subsequently in the thesis, the topic of *beamforming* in audio processing on super-direct linear arrays of microphones is addressed. The aim is to make an in-depth analysis of two main families of data-independent approaches and algorithms present in the literature by comparing their performances and the trade-off between *directivity* and *frequency invariance*, which is not yet known at to the *state-of-the-art*. The goal is to validate the best algorithm that allows, from the perspective of an implementation, to experimentally verify performance, correlating it with the characteristics and error statistics. Frequency-invariant beam patterns are often required by systems using an array of sensors to process broadband signals. In some experimental conditions, the array spatial aperture is shorter than the involved wavelengths. In these conditions, *superdirective beamforming* is essential for an efficient system. I present a comparison between two methods that deal with a data-independent beamformer based on a filter-and-sum structure. Both methods (the first one numerical, the second one analytic) formulate a mathematical convex minimization problem, in which the variables to be optimized are the filters coefficients or frequency responses. In the described simulations, I have chosen a geometry and a set-up of parameters that allows us to make a fair comparison between the performances of the two different design methods analyzed. In particular, I addressed a small linear array for audio capture with different purposes (hearing aids, audio surveillance system, video-conference system, multimedia device, etc.). The research activity carried out has been used for the launch of a high-tech device through an innovative start-up in the field of glasses/audio devices (https://acoesis.com/en/). It has been proven that the proposed algorithm gives the possibility of obtaining higher performances than the state of the art of similar algorithms, additionally providing the possibility of connecting *directivity* or better *generalized directivity* to the statistics of phase errors and gain of sensors, extremely important in *superdirective* arrays in the case of real and industrial implementation. Therefore, the method proposed by the comparison is innovative because it quantitatively links the physical construction characteristics of the array to measurable and experimentally verifiable quantities, making the real implementation process controllable. The third topic faced is the reconstruction of the *Room Impluse Response* (RIR) using audio processing *blind methods*. Given an unknown audio source, the estimation of time differences-of-arrivals (TDOAs) can be efficiently and robustly solved using blind channel identification and exploiting the cross-correlation identity (CCI). Prior *blind* works have improved the estimate of TDOAs by means of different algorithmic solutions and optimization strategies, while always sticking to the case $N = 2$ microphones. But what if we can obtain a direct improvement in performance by just increasing N? In the fourth Chapter I tried to investigate this direction, showing that, despite the arguable simplicity, this is capable of (sharply) improving upon state-of-the-art blind channel identification methods based on CCI, without modifying the computational pipeline. Inspired by our results, we seek to warm up the community and the practitioners by paving the way (with two concrete, yet preliminary, examples) towards joint approaches in which advances in the optimization are combined with an increased number of microphones, in order to achieve further improvements. Sound source localisation applications can be tackled by inferring the time-difference-of-arrivals (TDOAs) between a sound-emitting source and a set of microphones. Among the referred applications, one can surely list room-aware sound reproduction, room geometry's estimation, speech enhancement. Despite a broad spectrum of prior works estimate TDOAs from a known audio source, even when the signal emitted from the acoustic source is unknown, TDOAs can be inferred by comparing the signals received at two (or more) spatially separated microphones, using the notion of cross-corrlation identity (CCI). This is the key theoretical tool, not only, to make the ordering of microphones irrelevant during the acquisition stage, but also to solve the problem as blind channel identification, robustly and reliably inferring TDOAs from an unknown audio source. However, when dealing with natural environments, such “mutual agreement” between microphones can be tampered by a variety of audio ambiguities such as ambient noise. Furthermore, each observed signal may contain multiple distorted or delayed replicas of the emitting source due to reflections or generic boundary effects related to the (closed) environment. Thus, robustly estimating TDOAs is surely a challenging problem and CCI-based approaches cast it as single-input/multi-output blind channel identification. Such methods promote robustness in the estimate from the methodological standpoint: using either energy-based regularization, sparsity or positivity constraints, while also pre-conditioning the solution space. Last but not least, the *Acoustic Imaging* is an imaging modality that exploits the propagation of acoustic waves in a medium to recover the spatial distribution and intensity of sound sources in a given region. Well known and widespread acoustic imaging applications are, for example, sonar and ultrasound. There are active and passive imaging devices: in the context of this thesis I consider a passive imaging system called *Dual Cam* that does not emit any sound but acquires it from the environment. In an acoustic image each pixel corresponds to the sound intensity of the source, the whose position is described by a particular pair of angles and, in the case in which the beamformer can, as in our case, work in near-field, from a distance on which the system is focused. In the last part of this work I propose the use of a new modality characterized by a richer information content, namely acoustic images, for the sake of audio-visual scene understanding. Each pixel in such images is characterized by a spectral signature, associated to a specific direction in space and obtained by processing the audio signals coming from an array of microphones. By coupling such array with a video camera, we obtain spatio-temporal alignment of acoustic images and video frames. This constitutes a powerful source of self-supervision, which can be exploited in the learning pipeline we are proposing, without resorting to expensive data annotations. However, since 2D planar arrays are cumbersome and not as widespread as ordinary microphones, we propose that the richer information content of acoustic images can be distilled, through a self-supervised learning scheme, into more powerful audio and visual feature representations. The learnt feature representations can then be employed for downstream tasks such as classification and cross-modal retrieval, without the need of a microphone array. To prove that, we introduce a novel multimodal dataset consisting in RGB videos, raw audio signals and acoustic images, aligned in space and synchronized in time. Experimental results demonstrate the validity of our hypothesis and the effectiveness of the proposed pipeline, also when tested for tasks and datasets different from those used for training. Chapter 6 closes the thesis, presenting a development activity of a new *Dual Cam* POC to build-up from it a spin-off, assuming to apply for an innovation project for hi-tech start- ups (such as a SME instrument H2020) for a 50K€ grant, following the idea of the technology transfer. A deep analysis of the reference market, technologies and commercial competitors, business model and the FTO of intellectual property is then conducted. Finally, following the latest technological trends (https://www.flir.eu/products/si124/) a new version of the device (planar audio array) with reduced dimensions and improved technical characteristics is simulated, simpler and easier to use than the current one, opening up new interesting possibilities of development not only technical and scientific but also in terms of business fallout.

Chapter 2

Advanced image reconstruction algorithms in medical ultrasound

2.1 Introduction and Motivation

Ultrasound establishes a huge innovation in the biomedical scenario of the clinical imaging with the possibility to disrupt the existing production and business of the medical devices [2; 3]. Thanks to its real-time imaging capabilities, its non-invasive properties and low cost compared to any other medical modality, ultrasound has fundamentally impacted clinical portions of the diagnostic imaging: radiology, obstetrics, vascular imaging, cardiology and so on. Ultrasound is creating as well new markets for emergency and intervention medicine. In the future the guidelines of screening (mammography market), analysis (with the normalization of elasto-sonography strategies in the prostate) and medical procedure (with HIFU - High Intensity Focused Ultrasound) could change. The developments in the field of ultrasound are consistently significant. Actually, throughout the entire existence of ultrasound, numerous advancements have been created since its foundation during the 1960s as a clinical imaging device, with the recurrence of around a couple for every decade [4]: we think for instance the key advancement which has delivered the method of ongoing imaging, through mechanical examining during the 1960s; multi-channel electronic control frameworks for transducer clusters created during the 1970s; the tool of flow and analysis/investigation created during the 1980s that prompted imaging of shading streams and quantitative Doppler modes (Pulse Wave Doppler - PWD); the noteworthy enhancements in picture nature of the 1990s with the presentation of constant arrangement of the imaging strategies. Albeit a considerable lot of these ideas were concentrated in research labs years going before the previously mentioned business dates, it is deliberately the development of another

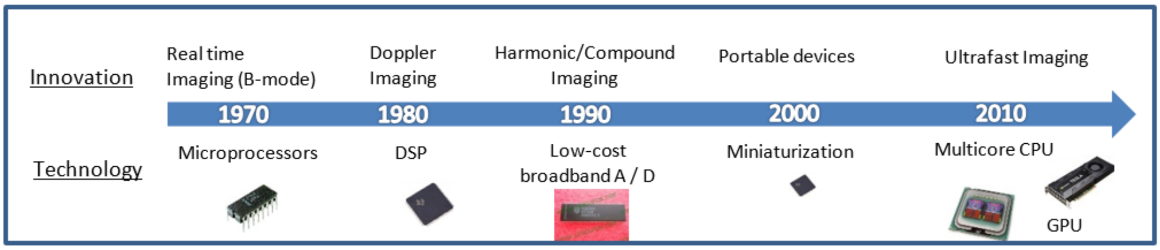


Figure 2.1 Some important steps in the innovation of ultrasound imaging with their corresponding enabling technologies.

innovation that triggers the presentation of advancements on accessible clinical routine stages in the commercial medical devices: for example, real-time imaging is made possible by the development of microprocessors, Doppler modes were made possible by the introduction of digital signal processing chips with sufficient dynamics to detect, at the same time, weak blood signals and strong tissue echoes; the introduction of low-cost analog-to-digital (A/D) converters has led to fully digital systems, significantly increasing the quality of the information provided; harmonic imaging was activated by large bandwidth transducers, enabling signal reception at twice the transmit frequency. In the first decade of the 21st century, technology has moved towards extensive miniaturization leading to the introduction of high-performance portable devices. Portable devices have created new markets for ultrasound, for example the emergency market, again underlining the destructive potential of the modality. Today, portable devices are the primary sources of market growth in the industry and miniaturization can be considered as a global trend of the ultrasound industry: available technologies and innovations are progressively integrated into portable systems. The following figure summarizes the evolution of ultrasound over the past decades (Figure 2.1). Nowadays a new technological adventure is underway with the advent of enormous parallel computing capabilities. This is due to the tremendous demand for performance processing and visualization needed in the video game industry. In addition to the multi-core architecture CPUs, the new graphics processing units (GPUs) allow parallel processing on thousands of channels simultaneously. This technology is available for the ultrasound industry and is the activator of software-based architecture systems. In 2009, SuperSonic Imagine introduced the first complete ultrasound system with a software approach (Aixplorer®): instead of increasing the integrated hardware processing channels, all processing is done by the software unit (CPU and GPU). The concept of processing channels disappears and the system is able to calculate in parallel many channels required by the acquisition. This architecture opens up a new way to perform ultrasound imaging by allowing you to review standard ultrasound modes using ultra-fast capabilities and improving projection performance of current ultrasound devices.

2.2 Ultrafast ultrasound imaging: state of the art

Ultrasound imaging is typically performed by sequential insonification of the medium using focused beams. Each beam allows the reconstruction of an image line. A typical 2D image is made up of a few dozen lines (from 64 to 512): the overall sequence is illustrated in Figure 2.2. The transmission rate of the imaging mode is determined by the time it takes to transmit a beam, receive and process the backscattered echoes from the media, and repeat that for all lines of the image. For a traditional 2D image then, the time to build an image is (two ways):

$$T_{image} = \frac{N_{lines} * 2 * Z_{max}}{c} \tag{2.1}$$

Where Z_{max} is the depth of the image, c is the ultrasonic wave speed assumed constant (1540 m/s) and N_{lines} the number of lines in the image. The maximum frame rate achievable with this technique is:

$$FR_{max} = \frac{1}{T_{image}} \tag{2.2}$$

For example, an image 5 cm deep and 256 lines wide would have $FR_{max} = 60Hz$ for ultrasound system architectures designed to process one line of image at a time. The limitations of the conventional approach appear where they are needed higher FR in clinical applications, typically in ultrasound cardiography for heart movement analysis, as well as in 3D/4D images where the number of lines becomes significant (about a few thousand) . To overcome these functional limitations, parallelization schemes have been considered: in the academic field this has been reported to start from the late 1970s [5; 6; 7; 8]. Most of the current systems on the market have the multiline functionality: for each transmission beam, several reconstruction lines are calculated (typically from 2 to 16). Multiline processing can be used both to increase the frame rate (for example for echocardiography) and to increase the number of lines calculated per image (for the 3D image). With or without multiline capability, current ultrasound systems are built on a serialized architecture and images are sequentially reconstructed from several equivalent transmissions. Ultrafast imaging breaks this paradigm: an ultra-fast imaging system is able to calculate as many lines in parallel as required and is therefore able to calculate a complete image from a single transmission regardless of the size and characteristics of the image. In such a system, the FR of the image is no longer

Table 2.1 Example of typical FRs in various clinical applications for conventional and ultra-fast architectures.

Application	**Depth of imaging**	**Conventional Architecture**	**Ultrafast**
Abdominal imaging	20 cm	20 Hz	3800 Hz
Cardiac Imaging	15 cm	150 Hz	5000 Hz
Breast imaging	5 cm	60 Hz	15000 Hz

limited by the number of reconstructed lines but by the flight time of a single pulse to propagate in the medium and return to the transducer. Table 2.1 provides the comparison of typical FR_{max} for different clinical ultrasound applications using conventional and ultra-fast architectures. New applications of ultrafast systems have therefore been reported in the literature. Fink has shown for the first time that transient shear waves, never seen before on an ultrasound scanner, can be picked up [9]. Jensen used an ultra-fast device to implement synthetic imaging techniques by deriving flow motion vector estimation [10]. The ultrafast device prototypes reported in these works allowed the storage of acquisition streams in a digital memory stack and then transfer to a PC. The processing was then performed off-line starting from the stored data. Although the concept of ultrafast imaging has been explored in academia during the last decade, it has only recently entered the commercial realm due to major technological barriers that needed to be overcome. For example, to obtain an ultra-fast image, the image calculation must be performed on a completely parallel platform, typically a platform based on software and not on hardware. However, there are two technologically demanding aspects to building a software-based platform:

- the data transfer rate from the acquisition module to the processing unit. Since raw RF signals (i.e. native and non-beam-formed) are transferred directly to the PC, the transmission rate required to perform real-time imaging is enormous: several Giga Byte / sec.
- the processing unit must be powerful enough to ensure real-time imaging. For example, conventional grayscale scanning requires 1 to 2 Giga-flops per second.

Powerful new processing units (GPUs) have achieved a satisfying level of performance. As a result, GPUs are increasingly used in the medical sector to speed up processing algorithms [11; 12; 13]. The ultra-fast architecture takes advantage of this processing power by combining it with fast numeric links (PCI Express technology) capable of transferring huge amounts of data to these drives. This combination enables the beam-forming process - the most challenging step in an ultrasound system - to be shifted from hardware to software, enabling complete parallelization of the ultrasound image calculation (Figure 2.3).

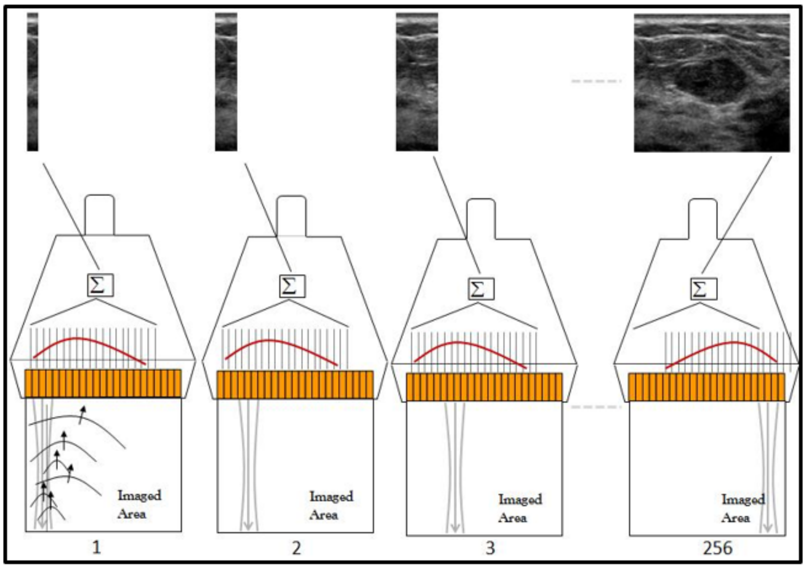


Figure 2.2 Conventional acquisition process of ultrasound images on a linear probe.

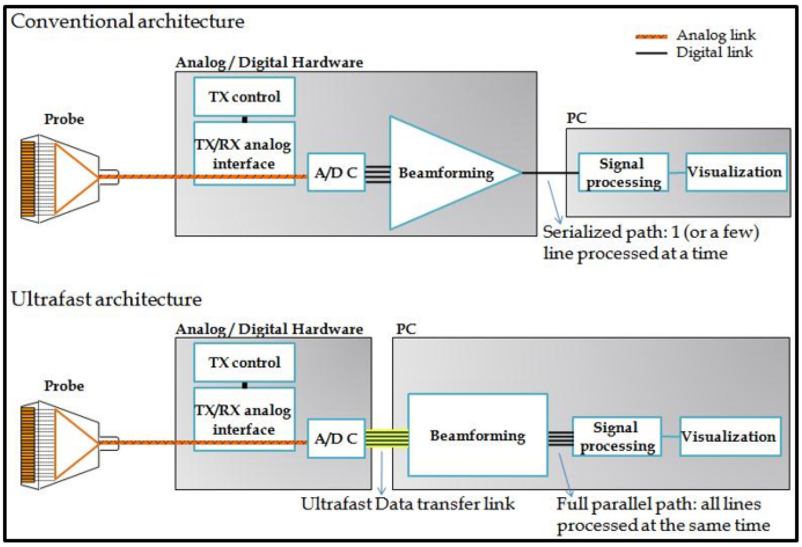


Figure 2.3 Through the beamforming performed by the software, it is possible to perform a perfect parallelization of the image formation. Each insonification can therefore lead to a complete picture.

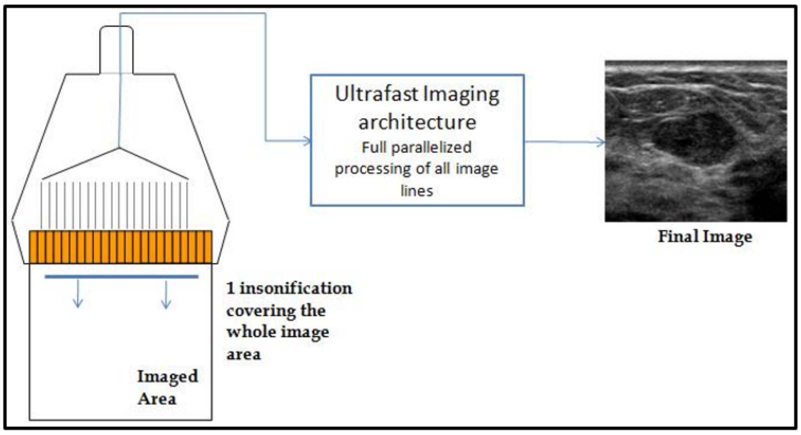

Ultrafast Imaging architecture
Full parallelized processing of all image lines
Final Image
1 insonification covering the whole image area
Imaged Area

Figure 2.4 A plane wave is sent by a linear transducer and insonifies the entire region of interest in a single shot. The image is calculated by processing this single insonification.

There are many ways to exploit an ultrafast imaging architecture [14; 15]. The SuperSonic Imagine approach, for example, is based on the use of insonification with unfocused plane waves (Plane Wave Imaging PWI). A plane wave is generated by applying delays on the transmission elements of the ultrasound probe as shown in Figure 2.4: the wave generated will insonify the entire area of interest. The backscattered echoes are then recorded and processed by the scanner to calculate an image of the insonified area. Plane wave imaging allows to calculate a complete ultrasound image for transmission at the expense of image quality. As the transmission focus step is removed, image contrast and resolution are consequently reduced. To overcome this limitation, several inclined plane waves at different steering angles are sent in the middle [16] and are consistently summed to calculate a complete image. Using this method, the transmission focusing step is performed retrospectively from this sum (Figure 2.5). The acquisition sequence with plane wave compounding has several advantages:

- Firstly, retrospective transmission focusing can be performed dynamically for each pixel of the image increasing the homogeneity of the final image with respect to physical isolation.

- Secondly, the number of steps required to obtain an image of equivalent quality to a focused mode (in terms of contrast and resolution) is approximately 5 to 10 times lower [16]. Consequently, the frame rates of the ultrasound images can be increased by the same factor by using strategies of using coherent plane waves on an ultrafast ultrasound system. The maximum achievable frequencies increase from 30 Hz to more than 300 Hz.

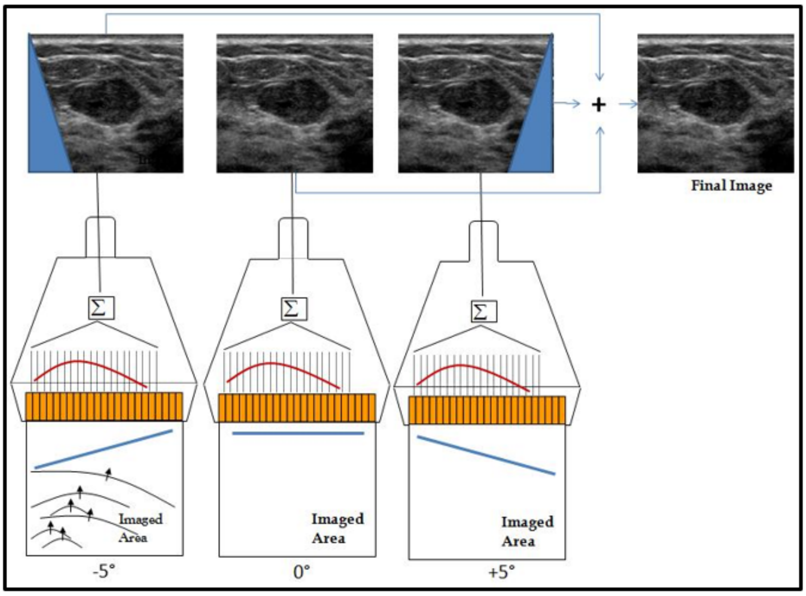


Figure 2.5 Ultrasound image obtained using the coherent compounding of field of view insonifications by plane waves at different steering angles.

Data sets of plane waves at different transmission angles were then acquired according to preliminary physical/mathematical evaluations for the best compromise between resolution, signal/noise ratio and contrast/noise ratio and reworked according to different coherent compounding algorithms to reach the best reconstructed result.

2.3 Plane wave imaging: data acquisition and image reconstruction

The analysis and development of the algorithms responsible for carrying out the thesis were approached from a feasibility point in an off-line context and on the *Matlab* platform, processing both synthetic simulated generated data and real RF data: the subsequent development of these algorithms within of the future ultrasound biomedical equipment will exploit a proprietary high-performance computing framework capable of processing customized kernel pipelines (henceforth called ’filters’) on CPU/GPU. The type of filters implemented has involved the theme of Plane Wave Imaging (PWI) described above in a general context, that is an alternative method of acquiring the ultrasound image compared to the state of the art of the traditional standard B-mode that currently exploit insonification sequence of the sample under examination through focused beams transmitted by the probe channels (Figure 2.2). The PWI mode is interesting and opens up new scenarios compared to the usual signal acquisition and processing techniques, with the aim of making signal processing in general and image reconstruction in particular faster and more flexible, and increasing importantly, the frame rate opens up and improves clinical applications. For the validation of the implemented beamforming algorithms in reception (see below) synthetic input signals generated digitally in advanced simulation environments were used (**Field II** https://field-ii.dk/, **K-Wave** http://www.k-wave.org/) on point scatterers, and real acquisitions of data acquired with ultrasound on a static phantom with PWI mode: these algorithms, as anticipated above, open the horizon to the ultrasound machine with a "software" approach on a computational platform compared to the state of the art of "hardware based" processing. To acquire the PWI mode, ad hoc focusing files had to be implemented to allow the acquisition in this mode as it is not standard, subsequently transferring them to the ultrasound machine by testing all the conditions of experimental and application interest on all the geometry modes of the probes used (linear, convex, phased array).

2.4 Beamforming algorithms

The PWI gives the possibility to address the problem of image reconstruction (called beamforming) with various approaches: reconstruction algorithms in the time domain and reconstruction algorithms in the frequency domain. Let's see them in detail.

2.4.1 Time Domain Algorithm: Delay And Sum (DAS)

The technique of transmission and acquisition of plane waves requires specific algorithms to migrate the temporal signal acquired by the elements of the probe to the depth of the field of view of the acquired image. First of all, the algorithm currently performed via hardware defined DAS (Delay And Sum) of refocusing in reception of the signal has been implemented in Matlab which is based on the identification within the RF data matrix of the contributions belonging to the image pixels through the Euclidean geometry of the probe (Figure 2.6). The inverse problem consists in finding the energy at the position of the scatterator points using the data acquired at the elements of the probe in the different time samples. The migration problem can be seen as the transposition of the acquired temporal signal to the depth coordinate. Consider a simplified case with a medium composed of a single scatterer element. Using the plane wave approach, the diffusion point is stimulated by an approximation of the plane wave generated by the emission of the same signal from each probe element with a constant delay curve. The wavefront generated approaches the plane wave well, ignoring the boundary regions that are usually not of interest during an ultrasound

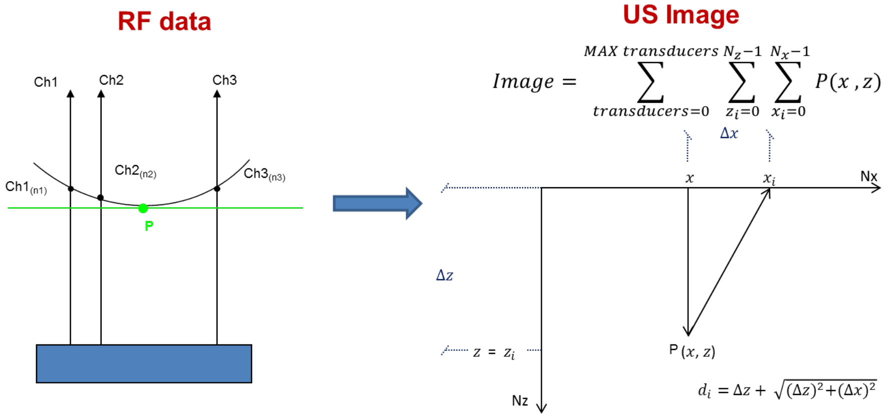

RF data
Ch1
Ch2
Ch3
Ch1(n1)
Ch2(n2)
Ch3(n3)
P
US Image
$Image = \sum_{transducers=0}^{MAX\ transducers} \sum_{z_i=0}^{N_z-1} \sum_{x_i=0}^{N_x-1} P(x,z)$
Δx
x
x_i
Nx
Δz
$z = z_i$
P (x, z)
Nz
$d_i = \Delta z + \sqrt{(\Delta z)^2+(\Delta x)^2}$

Figure 2.6 Beamforming obtained with algorithm in DAS times following excitation with plane wave.

survey. Using a linear delay curve, the generated wavefront approaches a plane wave at a specific angle that allows for insonification from a different point of view. Using the delay and sum approach the basic idea is to calculate for the considered depth point the time of arrival of the signal on each probe element, delay the time signal of each element of the corresponding time delay and add all the contributions to obtain the corresponding energy scattered from the point. The idea is that for the right point the delays produce signals in phase for each element causing a constructive summation. If we try to reconstruct the energy deriving from a point that does not characterize a scatterator, the signals are not in phase causing a destructive summation and consequently a lower energy. Considering the discrete case, the goal is to find the correct index for each column of the RF data array and sum the contribution on the channels. The process requires the assumption of a constant or known sound wave velocity. Each signal point of the ultrasound matrix is the result of the sum of the contributions of the signals of each individual component of the array on the arc at that point. Repeating the operation for all points in the matrix creates the image. The scattering point stimulated by the plane wave generates a spherical response according to the Huygens principle (Figure 2.7). The spherical wave front propagates to the elements of the probe that are affected by the signal with a delay that depends on the distance from the element and the point of dispersion. The energy of the acquired signal is attenuated by a factor that depends on the inverse of the distance and on the frequency attenuation component with respect to the depth equal to 0.5 dB * Mhz / cm. The basic principle of the DAS algorithm presents some problems that must be highlighted:

- The delay derived from the distance of the probe element to the scatterer element is not limited to being a multiple of the time phase of the signal and it is generally necessary

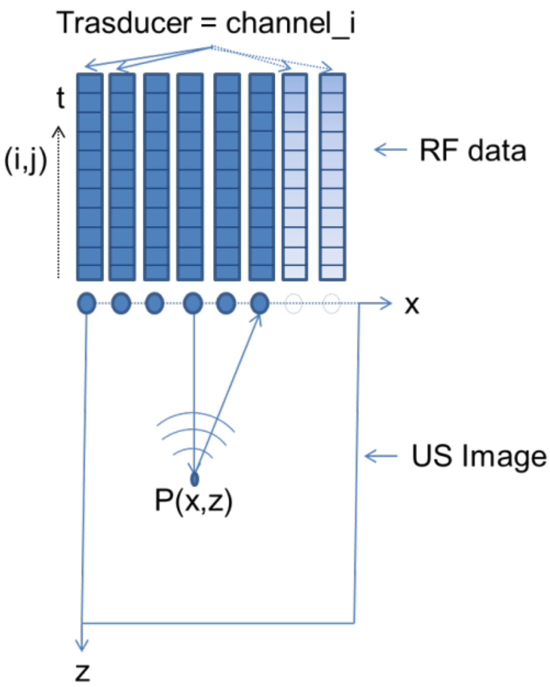


Figure 2.7 Generation of spherical waves by a point scatterer excited by a plane wave.

to make an approximation to select the closest sample. If the approximation is not checked it can cause an incorrect reconstruction of the image. The required time step is less than the acquisition accuracy (which complies with the Nyquist principle) by something nearly x4 and consequently higher sampling and interpolation is required before the delay and sum algorithm is applied. This oversampling interpolation causes a tradeoff between focus and performance that must be considered. A possible alternative to oversampling in the time domain and a potential theoretical solution could be to apply the delay in the frequency using the "shift theorem" (or delay) given by the properties of the Fourier transform:

$$F\left\{f\left(t-t_0\right)\right\}(\omega)=e^{-j2\pi\omega t_0}F(\omega) \tag{2.3}$$

Considering the Fourier transform of the signal, it is possible to apply the delay using a multiplication of the phase factor with infinite precision. The resulting focus is much better than the temporal focus (Figure 2.9), but requires the calculation of the direct and inverse Fourier transformation and the resulting algorithm, however, has a performance that is not competitive in computational terms for a real-time method.

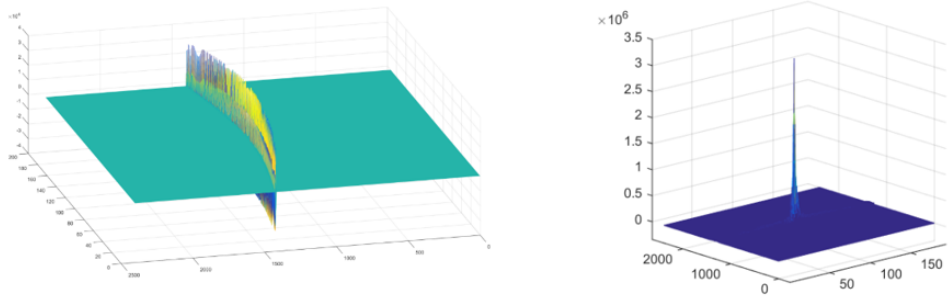


Figure 2.8 Simulated RF data of the delay curve generated by a point scatterer excited by a plane wave (left) and result of the Matlab beamforming algorithm of the DAS type obtained on this data in linear scale (right).

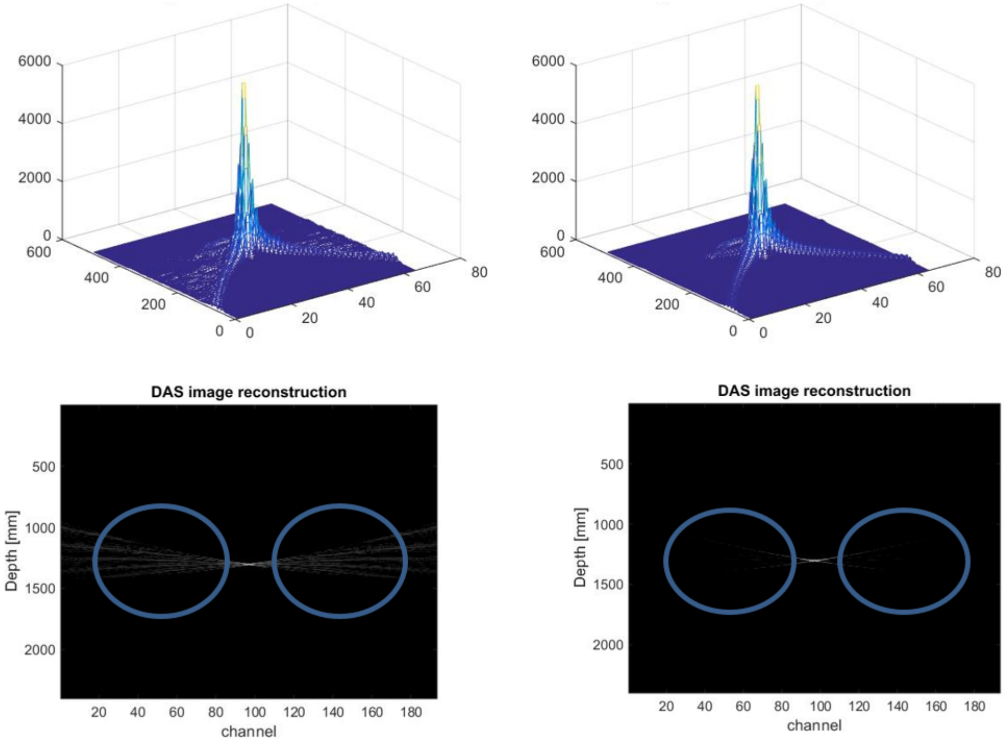


Figure 2.9 Comparison between the result of the DAS in the time domain without preliminary oversampling (left) and the DAS obtained through the application of the shift theorem for calculating the delays always obtained from a point scatterer excited by a plane wave (right). The focusing in frequency is more precise (see the indication of the areas inside the blue curves on the image to highlight the difference).

- Considering the ideal case of a probe with infinite length and infinite density of the elements, all the scattering points of the image will be perfectly reconstructed, but in a real application it is necessary to deal with the finite length of the probe that produces artifacts on the reconstructed image. These artifacts are more relevant considering the points at greater depth where the plane wave assumption and the non-infinite length of the probe produce a higher effect (the elements of the probe cover a smaller angular portion from the point of view of the scatterer). To make the calculation more efficient and minimize artifacts, a dynamic focal aperture was implemented simultaneously with the beamforming algorithm in order to intelligently weigh the contributions of the various image pixels. It can be shown that the delay and sum algorithm creates a reconstructed image that can be approximated as the convolution of the medium with the Fourier transform of the weighted probe profile with an extrapolation to the depth axis depending on the maximum depth of analysis (which corresponds to the maximum acquisition time). Using the signals acquired by the elements of the probe without applying gains, the result is to apply the inverse Fourier transform of a Heaviside function which is a *sinc* function to the medium under examination. The *sinc* function has a good resolution related to the size of the main lobe but has side lobes of lesser intensity. The consequence is that a single scattering point produces a number of side lobes on the reconstructed ultrasound image. To avoid this behavior, a profile gain is applied that approximates a center of a Gaussian curve on the spatial coordinates of the point under analysis. The Fourier transform of a Gaussian curve is another Gaussian curve and consequently the reconstructed image of a medium composed of a single scattering point will have a larger main lobe than the case obtained using a Heaviside function (less spatial resolution) but without side lobes. This additional necessary image quality optimization algorithm is called *apodization*.

- Working with the complex signal at the end of the reconstruction process to sum the contributions is necessary to perform an envelope of the signals. The envelope can then be calculated by applying the Hilbert transform and calculating the absolute value of the resulting complex signal. The problem is that the input signal is not a Dirac but a finite signal and after the delay and sum process the resulting signal is "distorted" by the operations performed. By calculating the Hilbert transform of the distorted signal downstream of the beamforming to extract the envelope, the final result is affected by artifacts. The solution adopted consists in applying the Hilbert filter directly to the input values of the RF data as the first stage through the software beamforming algorithm. Subsequently, all the delay and sum operations are performed on the I and

Q components (phase and quadrature) and only the absolute value is calculated at the end of the process to extract the envelope (pixel beamforming). The consequence is that the operations must be performed on both components I and Q (and therefore duplicated) but there is a gain in terms of quality of the final resulting image. A considerable advantage of the delay and sum algorithm seen on a linear probe is that it can be extended without significant problems to the case of convex and phased array geometry probes. The only difference compared to the linear case presented above is the calculation of the delays of the signal arriving at each probe element which must be defined specifically for the geometry of the probe taken in that case.

2.4.2 Algorithms in the frequency domain

The software reconstruction approach of the RF data to arrive at the final image allows a versatile cascade application of 'filters' that make the calculation more efficient and minimize artifacts (such as the filter with the dynamic focal aperture used in a to intelligently weigh the contributions of the various pixels of the image). The versatility of the field of view excitation modalities described above allows to explore alternative processing techniques of the RF data by treating it in alternative domains to that of time, through algorithms that evaluate the spectrum of the RF signal in the frequency domain and operate on it. reconstructing the final image. In particular, the defined approaches of "beamforming in frequency" and/or "beamforming in the domain of "k-spaces" (conceptual MRI derivation [17]) which is the mapping in a frequency context of the RF data simulated or acquired on the machine. The translation "tout-court" of the DAS algorithm in frequency in fact, as mentioned above, while giving correct results from an iconographic point of view, it is not feasible from the point of view of the computational burden resulting in a number of accounts not compatible with the frame rate required by a method real-time such as ultrasound imaging The further necessary bibliographic research has led to the identification of various algorithms known to date, highlighting the strengths/weaknesses of each as well as different areas of application. Three different viable paths have been identified corresponding to as many beamforming algorithms implemented in Matlab environment taken from the three different scientific articles attached:

1. Reconstruction with interpolation based on a truncated and inverse FFT algorithm on k-space.
2. Migration in F-K space commonly called Stolt's migration.
3. The seismic migration.

Let's see the details of these roads.

1. Frequency algorithm based on truncated and inverse k-space FFT

This algorithm was the subject of M. Jaeger's PhD thesis in Physics at the University of Bern (Switzerland) [18], [19], [20]. It has been used essentially to date in the context of research projects relating in particular to ultrasound imaging related to the photo-acoustics research environment. We start from the well-known second degree equation of partial differential waves applied to the scalar field acoustic pressure $p(x,z,t)$ with propagation speed $c = 1540$ m/s but common to any electromagnetic vector coming from the Maxwell relations of physics classical:

$$\frac{\partial^2 p(x,z,t)}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 p(x,z,t)}{\partial t^2} - \frac{\partial^2 p(x,z,t)}{\partial x^2} \tag{2.4}$$

The initial boundary condition is given by:

$$\mathrm{p_0}(\mathbf{x}) = \mathrm{p}(\mathbf{x}, \mathrm{t} = 0) = \iiint \mathrm{d^3k\tilde{p}}(\mathbf{k})\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\mathbf{x})} \tag{2.5}$$

where **k** is the wavenumber. Rewriting the solution of the equation in terms of the Fourier integral we have:

$$\mathrm{p}(\mathbf{x}, \mathrm{t}) = \iiint \mathrm{d^3k\tilde{p}}(\mathbf{k})\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\mathbf{x})}\mathrm{e}^{-\mathrm{i}\omega \mathrm{t}} \quad \text{where } \omega = \mathrm{c}\|\mathbf{k}\| \tag{2.6}$$

Given the finite nature of the opening of the probe, it follows that only some frequencies are mapped, so the problem arises of interpolating the missing data. The description of the algorithm steps is as follows (Figure 2.10) Given the dispersion relationship $\omega = \mathrm{c}\|\mathbf{k}\|$, before the inverse Fourier transform a regularization operation is carried out that allows to have the components on a Cartesian grid (Figure 2.11) The steps of the algorithm are as follows:

1. Step: transform through FFT $\mathrm{s(x,t)}$ in $\tilde{\mathrm{s}}\,(\mathrm{k'_x k'_t})$
2. Step: for each $\mathrm{k_x} = \mathrm{k'_x}$ calculate $\mathrm{M^{-1}}_{\mathrm{k'_x}}\,(\mathrm{k'_t}, \mathrm{k_z})$
3. Step: for each $\mathrm{k_x}$, calculate $\widetilde{\mathrm{p_0}}\,(\mathrm{k_x}, \mathrm{k_z}) = \sum_{\mathrm{k_z}} \mathrm{M^{-1}}'_{\mathrm{k'_x}}\,(\mathrm{k'_t}, \mathrm{k_z}) \cdot \widetilde{\mathrm{s}}\,(\mathrm{k_x}, \mathrm{k'_z})$
4. Step: Inverse FFT $\widetilde{\mathrm{p_0}}\,(\mathrm{k_x}, \mathrm{k_z})$ in $\mathrm{p_0(x,z)}$

If the matrix M is non-invertible, then use the pseudoinverse matrix $\mathrm{M^{inv}}$

$$\mathrm{M_{k'_x}}\,(\mathrm{k'_t}, \mathrm{k_z}) = \tfrac{\mathrm{e}^{-\mathrm{i}(\mathrm{k_t} - \mathrm{k'_t})\mathrm{T}} - 1}{\mathrm{i}(\mathrm{k_t} - \mathrm{k'_t})\mathrm{T}} \qquad \Rightarrow \qquad \mathrm{M^{inv}_{k'_x}}\,(\mathrm{k'_t}, \mathrm{k_z}) = \tfrac{\mathrm{k_z}}{\mathrm{k'_t}} * \mathrm{conj}\left(\tfrac{\mathrm{e}^{\mathrm{i}(\mathrm{kt} - \mathrm{k'_t})\mathrm{T}} - 1}{\mathrm{i}(\mathrm{k_t} - \mathrm{k'_t})\mathrm{T}}\right)$$

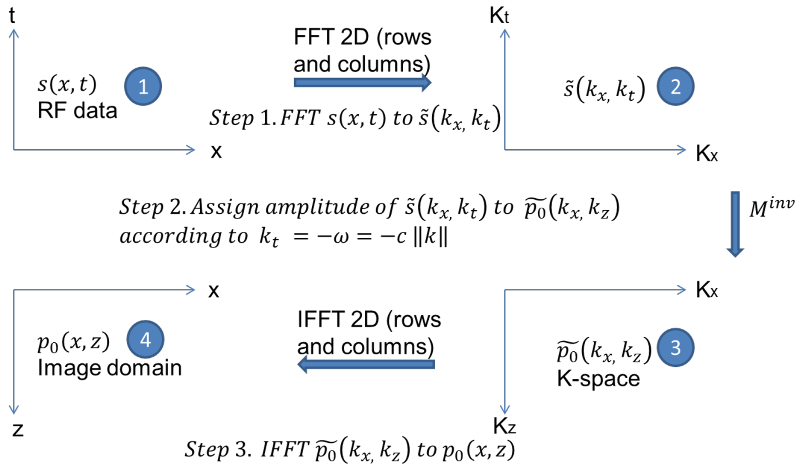

t
$s(x,t)$
RF data
1
x
FFT 2D (rows and columns)
Step 1. FFT $s(x,t)$ to $\tilde{s}(k_x, k_t)$
Kt
$\tilde{s}(k_x, k_t)$
2
Kx
Step 2. Assign amplitude of $\tilde{s}(k_x, k_t)$ to $\widetilde{p_0}(k_x, k_z)$ according to $k_t = -\omega = -c\,\|k\|$
M^{inv}
x
$p_0(x,z)$
4
Image domain
z
IFFT 2D (rows and columns)
Kx
$\widetilde{p_0}(k_x, k_z)$
3
K-space
Kz
Step 3. IFFT $\widetilde{p_0}(k_x, k_z)$ to $p_0(x,z)$

Figure 2.10 Reconstruction steps algorithm based on truncated and inverse FFT on k-space.

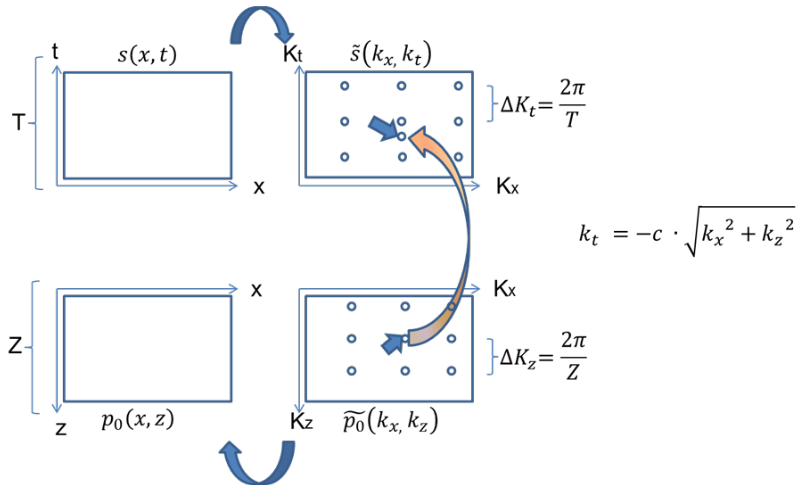

t
$s(x,t)$
T
x
Kt
$\tilde{s}(k_x, k_t)$
$\Delta K_t = \frac{2\pi}{T}$
Kx
$k_t = -c \cdot \sqrt{k_x^2 + k_z^2}$
x
Z
z
$p_0(x,z)$
Kx
$\Delta K_z = \frac{2\pi}{Z}$
Kz
$\widetilde{p_0}(k_x, k_z)$

Figure 2.11 Reconstruction steps algorithm based on truncated and inverse FFT on k-space (cont'd).

For computational efficiency only N elements of M^{inv} (N <= 20)

In summary, the conclusions with respect to the algorithm implemented and tested on synthetic and simulated signals and on real data acquired are as follows:

- Good candidate as a fast reconstruction algorithm.
- Effective solution for simple geometries and linear probes.
- Under feasibility, the possibility of applying the algorithm in more complicated geometries (phased array, convex), but the implementation solution is anything but clear and simple already at a theoretical level.

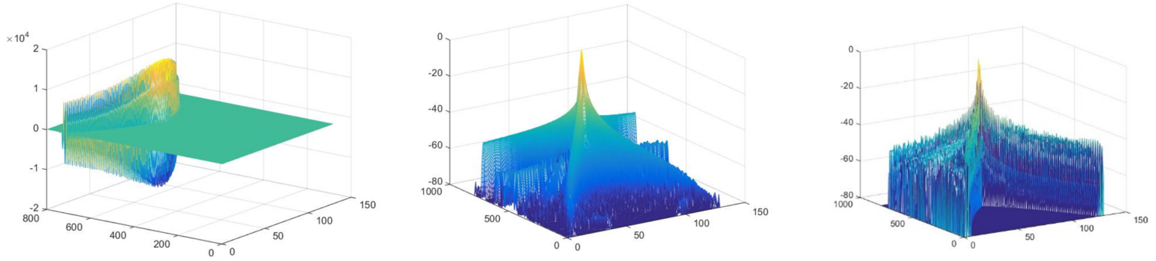


Figure 2.12 Simulated RF data of a plane wave generated by a single point scatterer (left) and reconstruction with an algorithm based on FFT truncated and inverse on k-space (center, logarithmic scale) and its comparison with respect to the reconstruction with the DAS algorithm in times without dynamic opening or "apodization" (right, logarithmic scale).

2. Frequency algorithm based on Stolt's F-K migration

Let $\psi(x,z,t)$ "a scalar wave field" under the hypothesis of the ERM (Exploding Reflector Model) that satisfies the known two-dimensional linear wave equation [21]:

$$\frac{\partial^2 \psi(\mathrm{x,z,t})}{\partial \mathrm{z}^2} = \frac{1}{\mathrm{c}^2}\frac{\partial^2 \psi(\mathrm{x,z,t})}{\partial \mathrm{t}^2} - \frac{\partial^2 \psi(\mathrm{x,z,t})}{\partial \mathrm{x}^2} \tag{2.7}$$

As in the seismic migration algorithm, which we will see later, we want to determine the ERM wave field at the moment of the explosion, i.e. $\psi(x,z,t=0)$ nowing the wave field on the probe surface $\psi(x,z=0,t)$. Let $\phi\ (\mathrm{k_x,z,f})$ the Fourier transform of $\psi(x,z,t)$ so that

$$\psi(\mathrm{x,z,t}) = \iint_{-\infty}^{+\infty} \phi\ (\mathrm{k_x,z,f})\, \mathrm{e}^{2\pi \mathrm{i}(\mathrm{k_x}-\mathrm{ft})} \mathrm{dk_x df} \tag{2.8}$$

Applying the Fourier transform produces the following Helmholtz equation:

$$\frac{\partial^2 \phi}{\partial z^2} + 4\pi^2 \widehat{\mathrm{k_z}}^2 \phi = 0 \tag{2.9}$$

with wavenumber

$$\widehat{\mathrm{k_z}}^2 = \frac{\mathrm{f}^2}{\mathrm{c}^2} - \mathrm{k_x^2} \tag{2.10}$$

The unique boundary condition is $\phi\ (\mathrm{k_x,0,f})$ which is the Fourier transform of $\psi(x,z=0,t)$ As we will do with the seismic migration we assume that $\psi(x,z,t)$ contains only the waves moving upwards. The ERM wave field is therefore authorized to propagate in the -z direction, as would happen with primary reflections. The wave equation can then be solved and the "migrated" wave field is obtained:

$$\psi(\mathrm{x},\mathrm{z},0)=\iint_{-\infty}^{+\infty}\phi\left(\mathrm{k_x},0,\mathrm{f}\right)\mathrm{e}^{2\pi\mathrm{i}\left(\mathrm{k_x x}-\widehat{\mathrm{k_z}}\mathrm{z}\right)}\mathrm{dk_x df} \tag{2.11}$$

To take full advantage of the Fourier transforms, Stolt proposed changing the variable $\mathrm{k_z}$ by introducing

$$\mathrm{f}\left(\mathrm{k_z}\right)=\widehat{\mathrm{c}}\,\mathrm{sign}\left(\widehat{\mathrm{k_z}}\right)\sqrt{\mathrm{k_x^2}+\widehat{\mathrm{k_z}}^2} \tag{2.12}$$

This expression describes the spectral re-mapping of the Stolt f-k migration for the PWI. Using the change of variables the Stolt migration is finally

$$\psi(\mathrm{x},\mathrm{z},0)=\iint_{-\infty}^{+\infty}\frac{\widehat{\mathrm{c}}\widehat{\mathrm{k_z}}}{\sqrt{\mathrm{k_x^2}+\widehat{\mathrm{k_z}}^2}}\phi\left(\mathrm{k_x},0,\mathrm{f}\left(\mathrm{k_z}\right)\right)\mathrm{e}^{2\pi\mathrm{i}\left(\mathrm{k_x x}-\widehat{\mathrm{k_z}}\mathrm{z}\right)}\mathrm{dk_x d}\widehat{\mathrm{k_z}} \tag{2.13}$$

The migrated solution is basically the inverse Fourier transform of

$$\frac{\widehat{\mathrm{c}}\widehat{\mathrm{k_z}}}{\sqrt{\mathrm{k_x^2}+\widehat{\mathrm{k_z}}^2}}\phi\left(\mathrm{k_x},0,\mathrm{f}\left(\mathrm{k_z}\right)\right) \tag{2.14}$$

In summary, the conclusions with respect to the Stolt algorithm implemented and tested in the Matlab environment on synthetic and simulated signals and on real data acquired are as follows:

- Good candidate as a fast reconstruction algorithm.
- Poor focus along the direction of the probe length.
- Impossibility of applying the algorithm in geometries more complicated than linear (phased array, convex).
- Performances worse than the truncated and inverse FFT algorithm on the k-space analyzed previously (see Figure 2.13).

3. Frequency algorithm based on seismic migration

We define again with $p(x,z,t)$ the acoustic pressure field acquired at the depth of the probe as a function of the continuous temporal and spatial coordinates. The pressure field received on

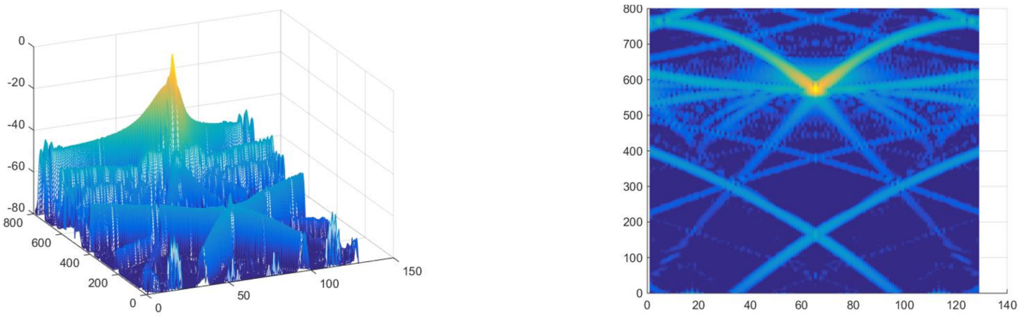


Figure 2.13 Reconstruction result with algorithm based on Stolt's algorithm (logarithmic scale) on a single point scatterator. It can also be seen from the "top" view of the graph (right) that the reconstruction is worse than the above-mentioned algorithms of the DAS in times and of Jaeger in frequency.

the probe surface can be seen as the boundary condition at the probe depth of the Helmholtz equation (wave equation in two dimensions in a homogeneous medium):

$$\frac{\partial^2 p(x,z,t)}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 p(x,z,t)}{\partial t^2} - \frac{\partial^2 p(x,z,t)}{\partial x^2} \tag{2.15}$$

The solution of the equation for the wave equation can be calculated starting from the boundary condition on the probe surface (z = 0), assuming that no signal arrives from the region above the probe and imposing a condition of Neumann homogeneous to the added operator of the equation for the upper limit of the region (requiring that all signals reaching the maximum depth do not return to the region of interest and there are no scattering elements after the region of interest to be considered) . The principle of this algorithm is that note the wave in the receiving position $p(x,0,t)$ (on the transducer), even the time-propagated wave can be known, given a pre-calculated matrix of "steps of delay ". The knowledge of the signals of the transducer upon reception allows us to know the pressure distribution in $p(x,0,t)$ and its Fourier transformation in space and time (2D) in $\mathrm{P}(\mathrm{k_x},0,\omega)$. Given a depth z_0 the 2D Inverse Fourier Transform of the product $\mathrm{e}^{\mathrm{jk_z z_0}} * \mathrm{P}(\mathrm{k_x},0,\omega)$ returns the pressure wave propagated back in time (how the signal looked at the position z_0). Applying the Fourier transforms to the acquired signal for the x coordinate and the time we obtain the following relations:

$$\mathrm{P}(\mathrm{k_x},\mathrm{z},\omega) = \mathrm{P}(\mathrm{k_x},0,\omega)\,\mathrm{e}^{\mathrm{jk_x z}}$$
$$\mathrm{p}(\mathrm{x},\mathrm{z},\mathrm{t}) = \textstyle\sum_{\mathrm{k_x}}\sum_{\omega}\mathrm{P}(\mathrm{k_x},\mathrm{z},\omega)\,\mathrm{e}^{\mathrm{j}(\mathrm{k_x x}+\omega \mathrm{t}))}$$
$$\mathrm{p}(\mathrm{x},\mathrm{z},\mathrm{t}) = \textstyle\sum_{\mathrm{k_x}}\mathrm{e}^{(\mathrm{jk_x x})}\sum_{\omega}\mathrm{e}^{\mathrm{jk_z z}}\mathrm{P}(\mathrm{k_x},0,\omega)\,\mathrm{e}^{(\mathrm{j}\omega \mathrm{t})}$$

Where the element k along z is the frequency of the signal along the z coordinate. The acquired signal is decomposed with respect to the frequency component along the x coordinate and the time coordinate. All the possible acquired signals can be described as the superposition of a number of plane waves with different orientation and frequency along the direction of propagation. By calculating the Fourier transform along the dimension of the probe x, the projection of the omega time frequency is calculated. Moving along z the signal is calculated as the superposition of the different frequency components along the time coordinate and x using the back propagation approach (the corresponding component is derived for the fixed depth for each temporal and x-spatial frequency). Then from the Fourier transform to the probe the pulsations along x and t are calculated and from them the pulsation along z is calculated using the Pythagorean theorem:

$$\frac{\omega^2}{c^2} = k_x^2 + k_z^2 \Rightarrow k_z = \frac{\omega}{c}\sqrt{1 - \frac{ck_x}{\omega}} \tag{2.16}$$

From the pulsation of the wave along z we can derive the wave propagation of any point along the depth line for a given time. By adding the contribution for each pulsation along x and t, the output of the 3D matrix is calculated. $p(x,z,t)$ represents the field that can be calculated at each point at any time generated by the medium of the region of interest. The next step is to derive the dispersion profile of the medium from the full pressure range in each sample of space and time. Imagine a fictional scenario in which a set of scatterator points emit a signal at a given synchronized time. Considering the function, the complete image of the medium is reconstructed. In a real case, each point scatters the signal at a different time depending on the position of the point with respect to the initial wave front and the speed of the wave. Considering a scattering point in the position z_0, where the scattering intensity of the point can be reconstructed considering the pressure field $p\left(x_0, z_0, \frac{z_0}{c}\right)$ and repeating the operation for each point, the signal is derived and then the Hilbert filter is applied and the absolute value reconstructs the complete image. In the real implementation of the algorithm, as in the delay and sum algorithm, the Hilbert filter is applied at the beginning of the process and all operations are applied to the I and Q components. At the end of the migration only the absolute value is calculated. The seismic migration algorithm can be summarized as follows:

1. Calculate the I and Q components from the RF signal using the Hilbert filter.
2. Calculate the FFT of I and Q after the Hilbert filter.
   START ITERATION ON THE DEPTH.

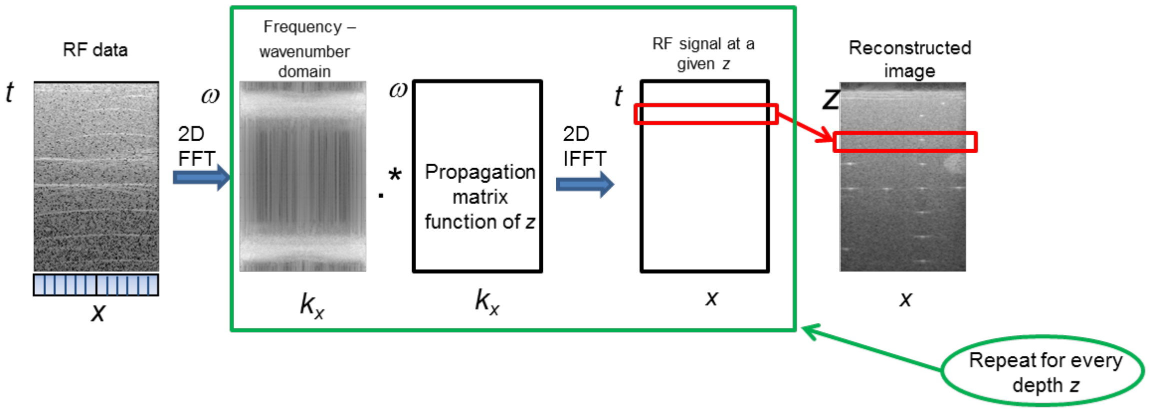


Figure 2.14 Reconstruction steps algorithm based on seismic migration.

3. Apply the propagation matrix $\mathrm{e}^{\mathrm{j}\sqrt{\frac{\omega^2}{\mathrm{c}^2-\mathrm{k}_\mathrm{x}^2}}*\mathrm{dZ}} = \mathrm{e}^{\mathrm{jk}_\mathrm{z}*\mathrm{dZ}}$ to *migrate* from ω and k_x to ω and k_z

4. Calculate the inverse Fourier transform and select the correct time for the current depth. Only the correct row of the output matrix is needed, so after calculating all the FFT for the time coordinate you can calculate the FFT along the x coordinate only for the correct row.

5. Apply the absolute value to get the row of the reconstructed matrix.
   END ITERATION ON THE DEPTH.

In summary, the conclusions with respect to the migration algorithm implemented and tested on synthetic and simulated signals and on real data acquired are as follows:

- Good candidate as a reconstruction algorithm not extremely fast compared to the previous ones but robust.

- Simple solution for simple geometries and linear probes.

- It is possible to apply the algorithm in more complicated geometries (phased array, convex).

Given the greater flexibility of the seismic migration algorithm on all types of transducers, more tests and simulations were carried out on this candidate, always in comparison to the standard algorithm DAS in the time domain.

Experimental results obtained on different experimental conditions

First of all, in the case of the convex probe (the case of the linear probe has been described above), the equation of the waves in polar coordinates must be rewritten, which in this reference system becomes:

$$\left(\partial_{rr} + \frac{1}{r}\partial_r + \frac{1}{r^2}\partial_{\vartheta\vartheta} - \frac{1}{c^2}\partial_{tt}\right) P(r,\vartheta,t) = 0 \tag{2.17}$$

The theory of the approach to be followed comes from the study of two papers [22; 23]. In the convex case, the wave equation in polar coordinates must be rewritten, which becomes in this reference system the equation above. Similarly to the approach followed for the linear probe, we assume that it can be expressed through a sum of Fourier components formalized by the following equation:

$$P(r,\vartheta,t) = \sum_{\omega,k_\vartheta} \hat{P}\left(r,k_\vartheta,\omega\right) e^{i(k_\vartheta \vartheta + \omega t)} \tag{2.18}$$

By substituting the solution 2.18 within equation 2.17, similarly to what was done for the linear, we find the equation for the Fourier components of the acoustic wave in spherical coordinates which appears to be:

$$\left(\partial_{rr} + \frac{1}{r}\partial_r + \frac{k_\vartheta}{r^2} + \frac{\omega^2}{c^2}\right) \hat{P}\left(r,k_\vartheta,\omega\right) = 0 \tag{2.19}$$

This equation has solution in the form of a linear combination of two Bessel functions. By placing $z = \frac{\omega}{c} r$; $p = k_\vartheta$ we have:

$$e^{i\omega t} H_p^{(1,2)}(z) \underset{z\to\infty}{\rightarrow} z^{-1/2} e^{i\omega t \pm i[z-(2p+1)\pi/4]} \left(1 + O\left(z^{-1}\right)\right) \tag{2.20}$$

The appearance of the two solutions is given by the Hankel functions.The relationship between two two-ray solutions r_1 and $r_2 = r_1 + \Delta r$ is given by:

$$\hat{P}\left(r_2,k_\theta,\omega\right) = \frac{H_{k_\theta}^{(1)}\left(\frac{\omega}{c}r_2\right)}{H_{k_\theta}^{(1)}\left(\frac{\omega}{c}r_1\right)} \cdot \hat{P}\left(r_1,k_\theta,\omega\right) \tag{2.21}$$

If $1/r$ varies slowly in the interval $[r_1, r_1 + \Delta r]$ then $\Delta r << r_1$. Then if $\Delta r/r << 1$ we can solve the Bessel equation in an approximate way in that $1/r$ is considered as a constant in the range $[r_1, r_1 + \Delta r]$: the approximate solution of the wave propagating forward (similarly

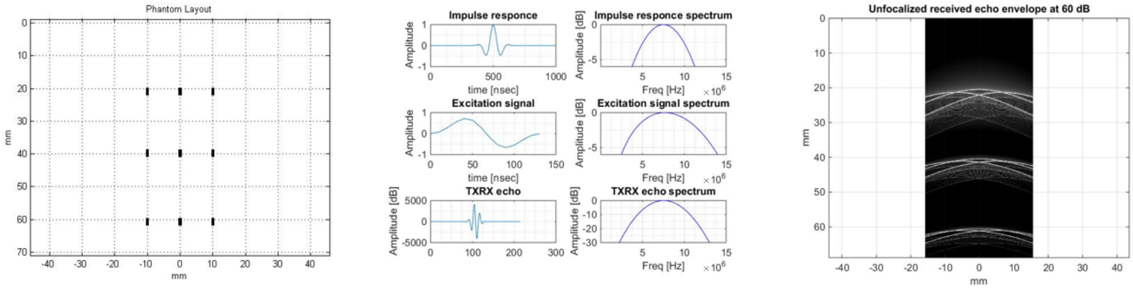


Figure 2.15 Position of the scatterers, characteristics of the Tx/Rx pulse, simulated cylindrical wave.

to the linear case, only the one with the + sign is considered) is given by:

$$e^{\left(-\frac{1}{2r_1}\pm ik_r\right)}\Delta r \qquad \text{where} \qquad k_r = \operatorname{sgn}(\omega)\sqrt{\left(\frac{\omega}{c}\right)^2 - \frac{1+4k_\vartheta^2}{4r_1^2}} \tag{2.22}$$

So in conclusion the solution that binds Equation 2.21 to the solutions of the Fourier series components on two radii at distances $r, r+\Delta r$ is given by the following equation:

$$\hat{P}\left(r+\Delta r, k_\theta, \omega\right) = e^{\left(-\frac{1}{2r}+ik_r\right)\Delta}\cdot\hat{P}\left(r, k_\theta, \omega\right) \tag{2.23}$$

Compared to the case of the linear probe, it is immediately evident that while in the first case the retro-propagation matrix is constant and can be calculated at the beginning of the algorithm only once based on the geometry of the probe, in the convex case the matrix must be recalculated to each iteration step thus increasing the computational burden; furthermore, the reconstructed image is expressed in coordinates r, θ for which a conversion into Cartesian coordinates is required at the end of the algorithm (scan converter). As a simulation test, nine point scatterers were set as in the first graph on the left: the transmission characteristics are the one shown in the central figure and the last graph represents the simulated cylindrical plane wave on the right (RF data as input to the algorithm of reconstruction of seismic migration according to convex geometry). At this point, the synthetic data input to the seismic migration algorithm was given and the promising results are reported in terms of numerical graphs and images in gray scale. The feasibility activity of developing the seismic migration algorithm also for the convex probe can therefore be said to have concluded with a positive outcome: apart from the increased and expected computational burden justified by the mathematical considerations expressed above, the reconstructed images are aligned with expectations and also free of artifacts, proving the accuracy of the implemented algorithm. It remains to convert the final image into Cartesian coordinates (scan-converter routine).

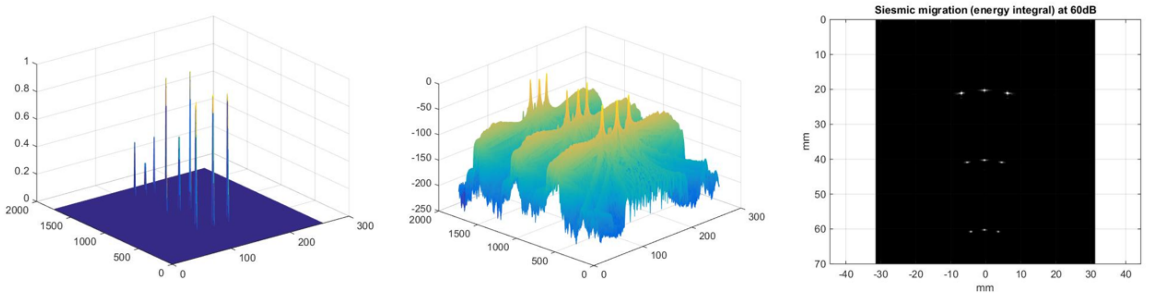


Figure 2.16 Position and amplitudes of the scatterers reconstructed by the seismic migration algorithm on a linear scale (left), in the graph in db (center) and finally the image reconstructed in shades of gray (right)

The logical steps of the seismic migration algorithm on the convex probe are therefore the following:

1. Double two-dimensional FFT (along the times and along the channels).
2. Hilbert transform which is equivalent to zeroing half matrix in the direction of the frequencies obtained at the previous point (IQ beamforming).
3. Calculation of the back-propagation matrix (unlike the linear, it must be recalculated at each iteration).
4. Iteration on all lines of the product matrix between the initial matrix and the retro-propagation factor.
5. Absolute value of each reconstructed row to be stored in the image matrix.
6. Scan conversion from polar coordinates to Cartesian's coordinates
7. Mapping of the numeric matrix in *dB* (decibel) and gray levels.

With a scan conversion routine always validated on the synthetic data (9 scatterers) the following result is obtained. Subsequently, real RF data was acquired from plane wave transmission on a static phantom (one shot data). The result processed on Matlab is put after the scan conversion where the goodness of the reconstruction obtained only with a single plane wave is highlighted. Given the validation of the seismic migration algorithm on synthetic and real data and on various geometries of the probes, comparisons were made between the different beamforming algorithms, comparing the offline reconstructions with the results obtained from a standard ultrasound system with beamforming hw-based. The comparisons between the results of the algorithms on a datum acquired with a plane wave on

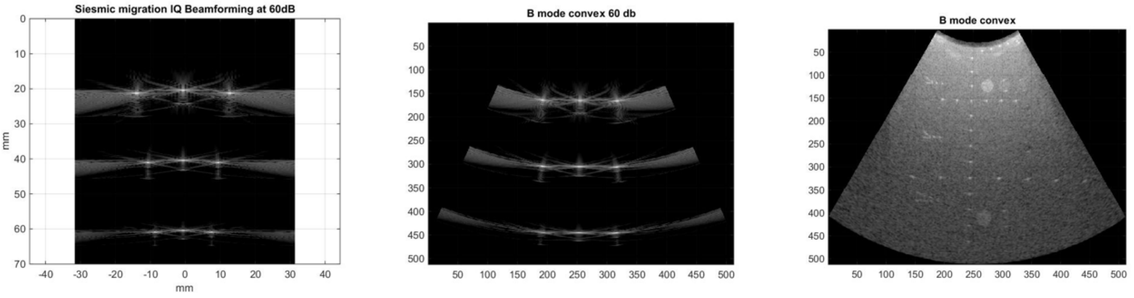


Figure 2.17 Position and amplitudes of the scatterers reconstructed by the seismic migration algorithm on a convex probe before the scan conversion routine (left) and after the application of the routine (center). Once the validation of the algorithm has been obtained on a synthetic data, a real RF data has been acquired on a convex probe and has been successfully reconstructed (last image reconstructed in shades of gray on the right)

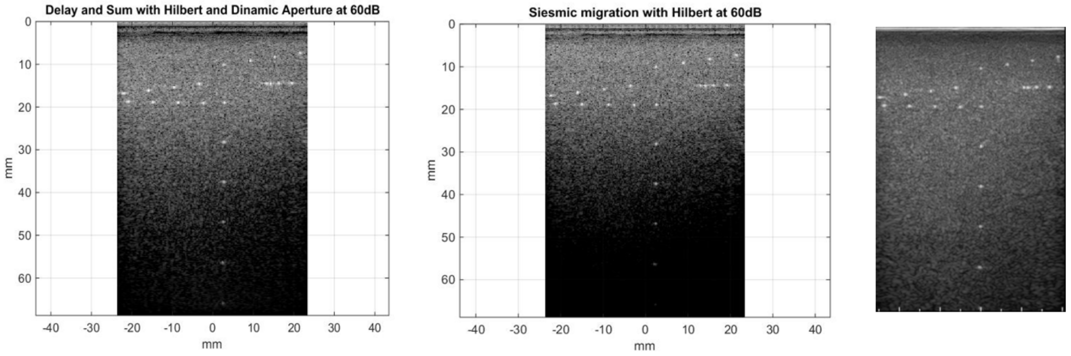

Delay and Sum with Hilbert and Dinamic Aperture at 60dB
Siesmic migration with Hilbert at 60dB
mm
mm
0
10
20
30
40
50
60
-40
-30
-20
-10
0
10
20
30
40

Figure 2.18 Reconstruction obtained via Matlab using a static phantom and linear probe on real RF data with a single shot of a single plane wave processed with the DAS time domain algorithm with dynamic aperture (left). Reconstruction of the same data processed with the seismic migration algorithm (center). Comparison of the results obtained with the two offline software algorithms in the time domain and the frequency one with a standard acquisition obtained from traditional beamforming ultrasound with hardware-based real time reconstruction (right).

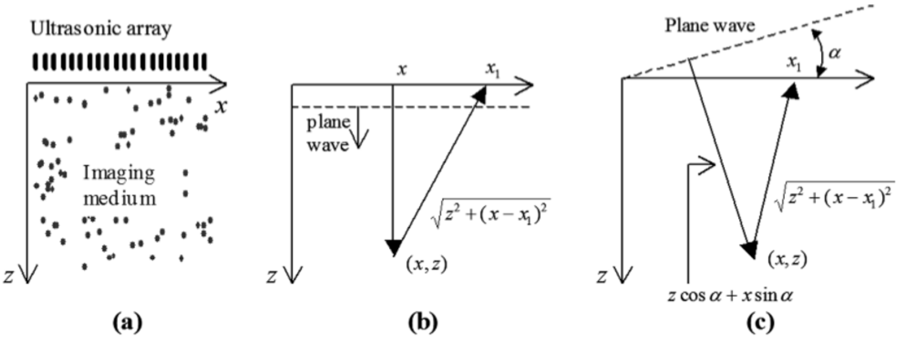


Figure 2.19 (a) Axis convention, (b) time delays for a plane wave insonification, and (c) time delays for a plane wave of angle α

a phantom with a linear probe are shown below: the results of the Matlab offine beamforming algorithm of the DAS are compared in the times with the dynamic opening, the seismic migration and the comparison of the reconstruction performed with the image obtained with standard focused transmission on traditional ultrasound.

2.5 Compounding plane wave algorithms

As can be seen from the previous comparison, although the real acquisitions were made through insonification with a single plane wave at 0° degrees of inclination, the result in terms of signal/noise ratio (S/N ratio) is not very far from the image obtained with real time standard hardware based beaforming [2]. To overcome this drawback (lower S/N but at high frame rate FR), it is possible to acquire successive plane waves at different steering angles and then compose the results obtained. Therefore, the code was also developed in the Matlab environment in order to evaluate the feasibility to extend the state of the art of the implementation of the algorithm in the times of the DAS on plane wave and of the seismic migration by inserting a *coherent compounding* algorithm (i.e. on the data Complex IQ) on a set of plane wave acquisitions characterized by different transmission *steering* angles. In particular this chapter takes into consideration the case of the linear probe. Referring to the diagram for example from [16], in the evaluation of delays in the case of plane wave steering contrary to the standard case, we have to consider the angle of incidence alpha in the geometric evaluation of distances and consequently in that of the "two-ways". In particular, referring to Figure 2.19 we have that the "outward" time to arrive at a point of coordinates

(x,z) within the medium with respect to the angle dependence is:

$$\tau_{ec}(\alpha,x,z) = (z\cos\alpha + x\sin\alpha)/c \tag{2.24}$$

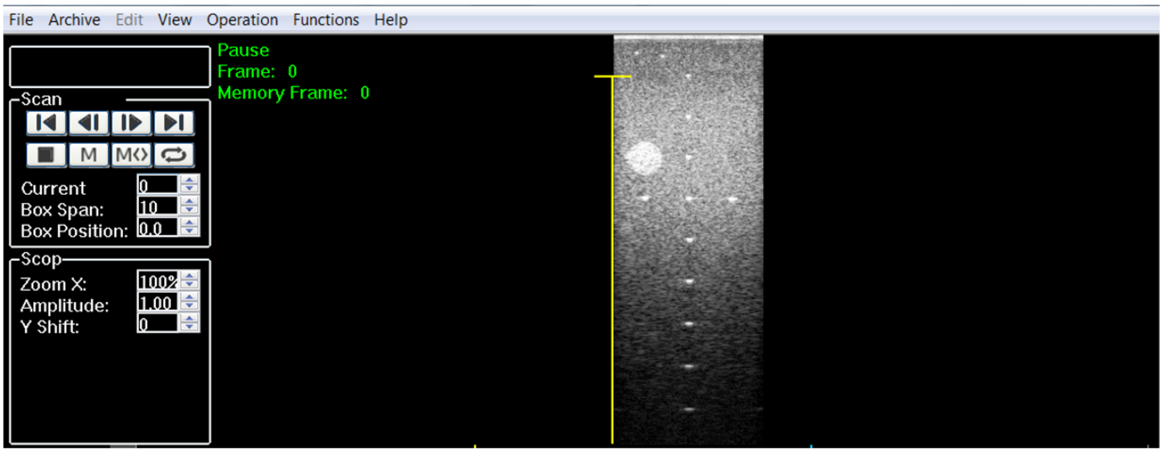
While the "return" time with respect to a transducer / channel located in abscissa x_1 is:

$$\tau_{rec}\left(x_1,x,z\right) = \sqrt{\left(z^2 + \left(x-x_1\right)^2/c\right)}. \tag{2.25}$$

From which it appears that the new delays on which to evaluate the DAS in time are given by the sum

$$\tau\left(\alpha,x_1,x,z\right) = \tau_{ec} + \tau_{rec} \tag{2.26}$$

The final image quality is obtained by making a coherent compounding of the images obtained by the beamforming of the signals received by the plane waves transmitted at different angles α. In accordance with the theory, experimentation was carried out on a static phantom using 15 angles of incidence (plus the wave at angle 0° for a total of 15 plane waves) chosen with the following inclinations (in degrees): $\alpha(\mathrm{i}): -12.0000; -10.2857; -8.5714; -6.8571; -5.1429; -3.4286; -1.7143; 0; +1.7143; +3.4286; +5.1429; +6.8571; +8.5714; +10.2857; 12.0000$. The "focusing files" corresponding to the angles listed above were constructed, consequently evaluating the delays evaluated by the previous expressions and subsequently the 15 series of RF data were acquired on a real ultrasound equipment at two different depths respectively at 96 mm and 103 mm to compare the yield of the algorithm even at high depht. First of all, the acquisition was reconstructed using the standard sequential DAS algorithm with focused transmission present on the machine and placed below as a reference image. Subsequently, the 13 RF data corresponding to the plane waves relating to the above described alpha angles in degrees were acquired The DAS algorithm was therefore extended (but the same applies to the seismic migration algorithm) to understand the dependence on the angle and as a first step we proceeded to a simple sum of the various contributions without particular weightings, given the results more than good. In fact, in the literature there are very refined evaluations that provide an analytical weighting and derivation of the number of angles that must be performed to recover a comparable signal/noise ratio compared to the standard technique. The partial results and the final sum are attached: as you can see from the figures, you can go to cut out the reconstructed images to make a comparison in terms of signal-to-noise ratio with respect to Figure 2.20, or you can add them to enlarge the field of view, circumstance that the new method with plane waves with steering allows to do. Also the feasibility activity of the compounding algorithm for the plane wave on a linear probe has given positive results:



Pause
Frame: 0
Memory Frame: 0

Current 0
Box Span: 10
Box Position: 0.0

Zoom X: 100%
Amplitude: 1.00
Y Shift: 0

Figure 2.20 Standard focused sequential transmission reference image on static phantom and linear probe.

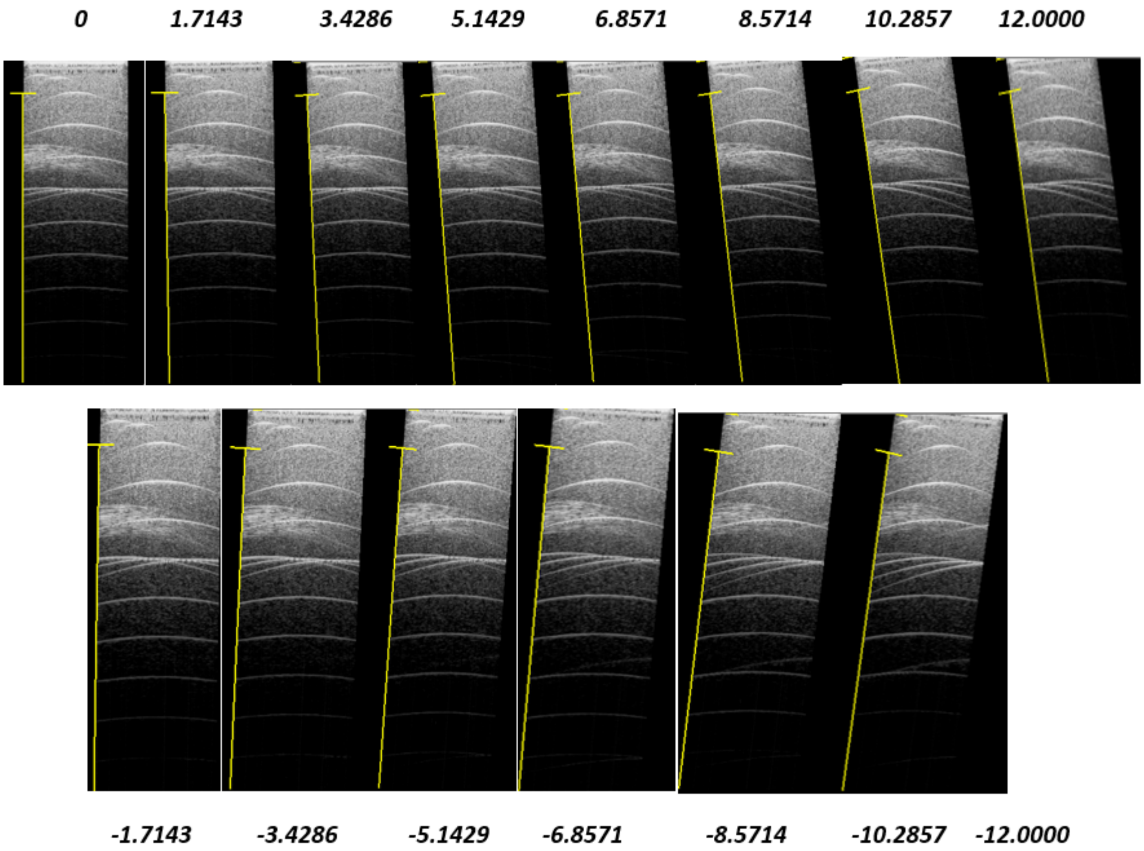
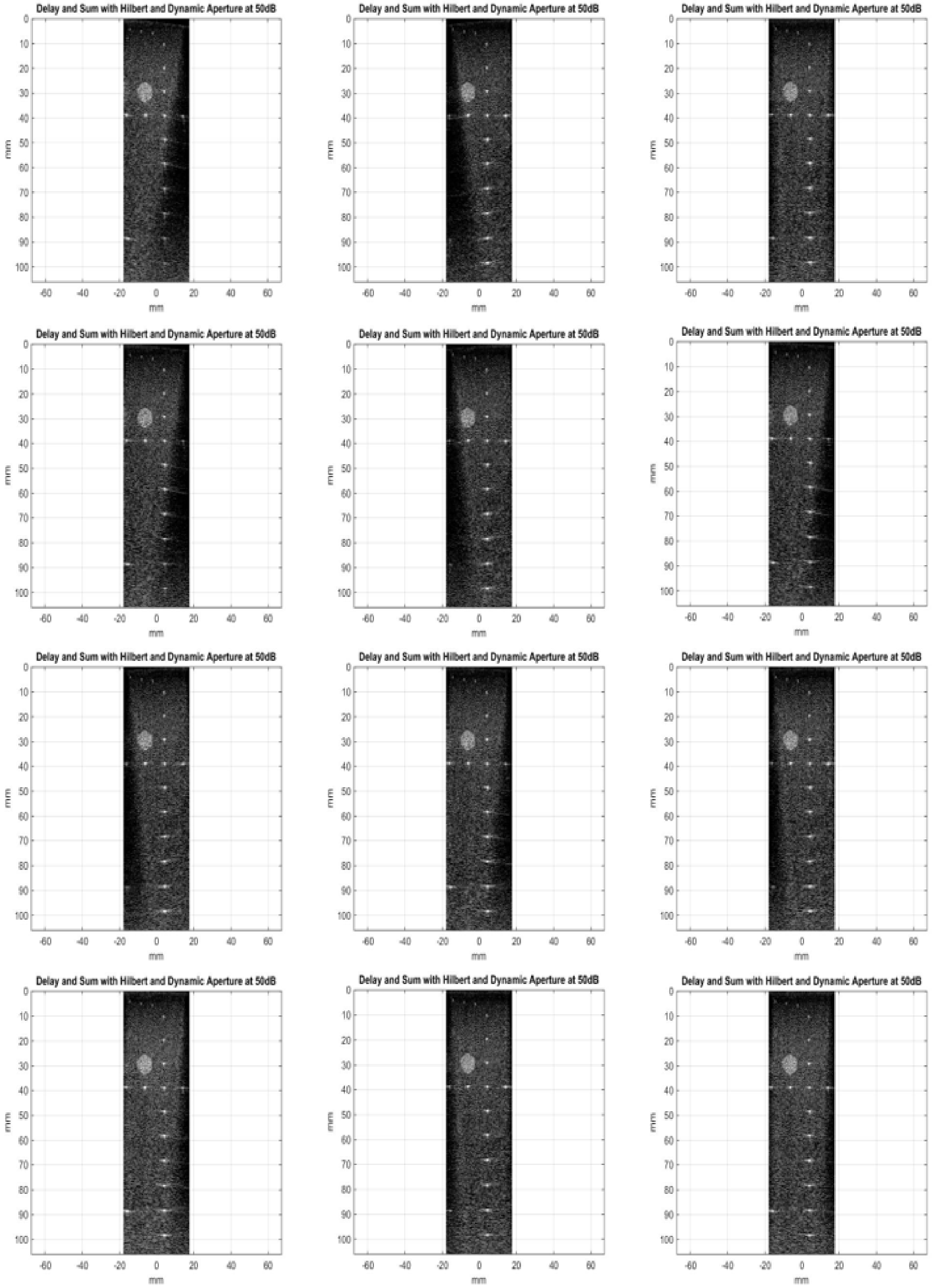


Figure 2.21 RF datasets acquired at the various angles of incidence above.



Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Delay and Sum with Hilbert and Dynamic Aperture at 50dB

Figure 2.22 Partial reconstructions of the images corresponding to the different single steering angles.

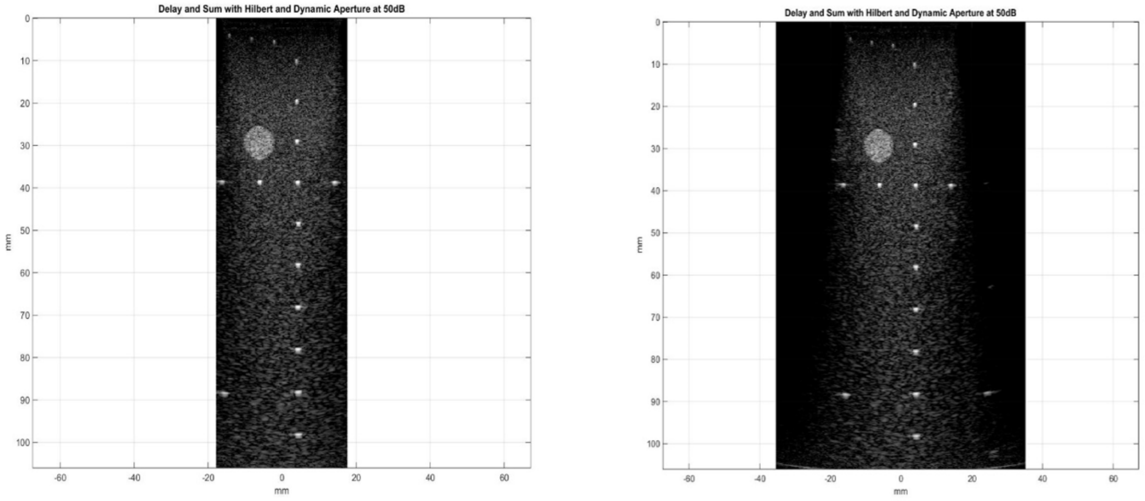

Delay and Sum with Hilbert and Dynamic Aperture at 50dB
0
10
20
30
40
50
60
70
80
90
100
mm
-60
-40
-20
0
20
40
60
mm
Delay and Sum with Hilbert and Dynamic Aperture at 50dB
0
10
20
30
40
50
60
70
80
90
100
mm
-60
-40
-20
0
20
40
60
mm

Figure 2.23 Reconstruction with compounding algorithm of the final image: standard comparison (left) and enlarged FOV (right).

it can be seen that already 13 insonification angles entail an excellent increase in the S/N ratio as can be concluded by comparing Figure 2.20 (standard hardware based sequential focusing) and Figure 2.23 (software off line reconstruction of compounding plane waves at different steering angles). It is therefore also experimentally verified that (as stated by the literature reported above) already with only about ten steering angles, a S/N ratio of the b-mode image is obtained, comparable to the traditional one, with an enormous advantage, however, of increasing the frame. installments that the PWI approach entails. Consequently, all the modalities and algorithms of reconstruction and recombination of beamforming software necessary to build a future software-based ultrasound scanner that works with unfocused plane waves were explored and positively evaded (through offline Matlab code).

2.6 Innovative image reconstruction algorithm based on back-propagation and PWI

Recent ultrasound imaging based application applications (2D elastography, ultrafast doppler, 4D cardio) require the use of high frame rate ultrasound imaging techniques [2; 8; 10; 14; 24; 25]. This implies minimizing the number of transmissions and, in the extreme case, reconstructing the whole image with a single transmission. One way to achieve this is to insonify the entire field of view with a plane (or spherical) wave, collect all the RF data and reconstruct the image. The application of the Delay and Sum in reception is not optimal. On the contrary, the seismic migration algorithm allows a series of advantages (exact delays, optimal intrinsic apodization, etc.) Seismic migration [26; 27]is based on the backward propagation of the acoustic field collected by the probe. In ideal conditions, infinite and continuous probe opening (scatters-points), in the absence of additive noise to the sensors and the absence of due to tissue attenuation, infinite bandwidth of the signal, the reconstruction of the distribution of scatters is correct, because it is based on the solution of waves on a boundary condition. However, such non-ideality make that the mere application of the algorithm can lead to a not optimal solution. The main idea of the patented algorithm is that through appropriate masking applications in the $k_x - \omega$ space in seismic migration algorithm context, you can avoid the non-ideal behaviors described above. Today the production of ultrasound pictures is based on the serial focused transmission of acoustic waves in the field of view and the reconstruction of the final picture is based on Delay and Sum algorithm and processing in the time domain of the RF data received from the probe. The idea for the next generation ultrasound machine is to use a faster and not focused transmission given by the plane wave and to process in a fully software approach of the reconstruction process of the RF data using advanced algorithm in the Fourier domain. Thanks to the potentiality of the new computing platform this approach allows to have faster imaging and to avoid some hardware pre-processing (oversampling and so on), reducing the cost of the equipment and guarantying higher performances. Today this kind of approach is used only in the scientific context (opto-acoustic mainly) and our final goal is to build an innovative and smart processing reconstruction algorithm suitable for a commercial ultrasound equipment and to protect it with the patent a competitive advantage compared to the competitors to have an important role in the next future in the technological commercial and industrial ultrasound trend. The algorithm is based on the solution of the wave equation in a homogeneous medium, given a boundary condition. In particular the pressure field sampled by the probe at depth z = 0, $p(x,0,t)$ can be back-propagated to recover the pressure field at a generic depth $p(x,z,t)$. Hence, the ultrasound image given by the scatterers distribution can be recovered by evaluating $p(x,z,t)$ at given times t dependent from each location (x,z). The solution of the wave equation

$$\frac{\partial^2 p(x,z,t)}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 p(x,z,t)}{\partial t^2} - \frac{\partial^2 p(x,z,t)}{\partial x^2} \tag{2.27}$$

is based on the 2D Discrete Fourier Transform across time and azimuth coordinates of the field sampled by the probe:

$$P(k_x,0,\omega) = \sum_x \sum_\omega P(x,0,t)e^{(-j(k_x x+\omega t))} \tag{2.28}$$

The solution at a generic depth z for a component (k_x, ω) given by:

$$P(k_x, z, \omega) = P(k_x, 0, \omega)\,\mathrm{e}^{jk_z z} \tag{2.29}$$

where

$$k_z = \frac{\omega}{c}\sqrt{1 - \left(\frac{ck_x}{\omega}\right)^2} \tag{2.30}$$

The final pressure field at depth z is given by the inverse 2D fourier transform:

$$p(x, z, t) = \sum_{k_x}\sum_{\omega} P(k_x, z, \omega)\, e^{(j\omega t)} e^{(jk_x x)} \tag{2.31}$$

2.6.1 Seismic migration with "masking" of the k-space: the algorithm

Here we find the fundamental steps of the algorithm of the seismic migration applied to the ultrasound imaging context

- Acquire a matrix r of RF data of size number of samples times number of probe channels, in response to plane wave transmission.
- Perform a 2D FFT on the RF data matrix (along temporal and azimuth axes) and compute the corresponding vectors of temporal frequencies ω and wavenumbers k_x
- For each depth z repeat:
  1. Compute the propagation matrix whose entries are given by $e^{jk_z * z}$ where $k_z = f(k_x, \omega)$
  2. Multiply element wise the propagation matrix by the data matrix R.
  3. Perform an inverse 2D FFT on the resulting matrix.
  4. For every azimuth position x consider the signal in a given range of times as representative of the scatterer intensity at location (x, z).

The innovative idea is to introduce a further matrix, named *masking matrix*, to be multiplied element-wise to the data matrix at each iteration of the algorithm. The masking matrix can change according to the reconstruction depth z. The masking matrices can be computed offline knowing the system parameters, since they do not depend from acquired data. Moreover, they can be pre-multiplied to propagation matrices, without affecting the overall computational load. The masking operation can be used to:

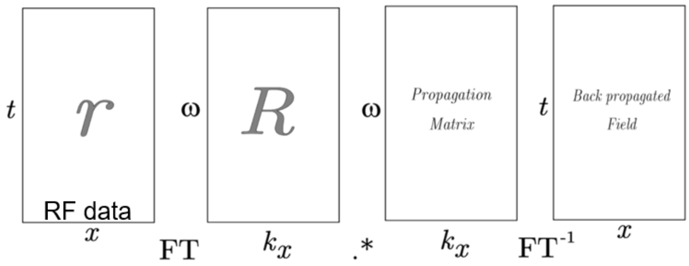


Figure 2.24 Algorithm flow in back-propagation algorithm for ultrasound imaging.

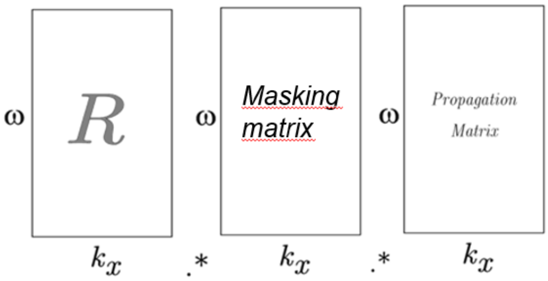


Figure 2.25 Novelty in the algorithm flow of back-propagation algorithm for ultrasound imaging.

1. Delete the contributions of evanescent waves from the k-space.
2. Take into account the distortion of the spectrum of received signals due to frequency-dependent tissue absorption.
3. Take into account the depth-dependent cut off of spatial frequencies in the received signals, due to the probe finite aperture.
4. Cope with the directivity of the individual transducers.
5. Work in a phase-quadrature (IQ) RF domain, thus allowing pixel based beamforming.

Let's go and explain each of the points listed above.

Evanescent waves

Definition of evanescent waves: all the components in the $k_x - \omega$ space for which

$$\frac{\omega^2}{c^2} - k_x^2 < 0 \tag{2.32}$$

correspond to evanescent waves that do not propagate. Consequently they have to be eliminated setting to zero all the points for which the above condition is true. This condition speed up the calculations of the algorithm of the seismic migration in the transformed space.

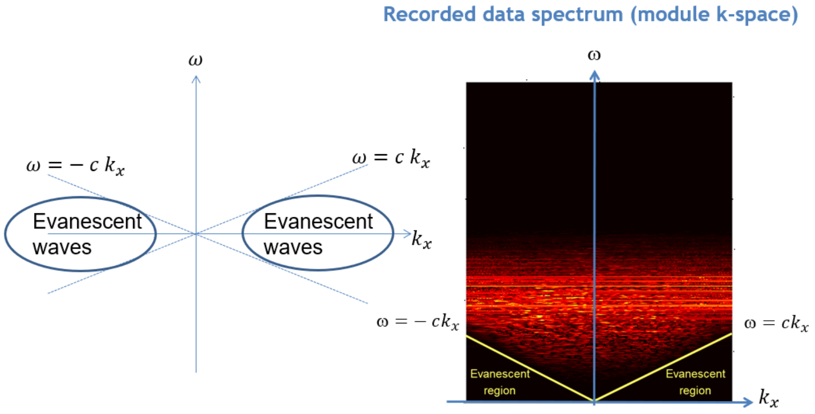


Figure 2.26 Evanescent wake masking in the $k_x - \omega$ space step.

Frequency dependent tissue absorption

The attenuation in the tissues is proportional to the frequency, typically 0.5dB / cm / MHz. As a result the spectrum of the received signals will tend to decrease in bandwidth and move towards a lower central frequency as far as the scatterers depth z increases. As a consequence the useful signal band, where signal to noise ratio (SNR) is acceptable, will change accordingly. In order to keep the overall SNR sufficiently high, in standard delay-and-sum beamforming, variable pass band filters are applied to RF signals. Since the filters parameters vary with depth, this operation is computationally expensive. In seismic migration such filters can be replaced by masking of the proper range of ω according to the current depth z. The binary masking is the simplest case, resulting in a reduction of computational load. However more complex masking can be devised, e.g. smoothing the boundaries of useful band. If the overall SNR is sufficiently high, masking can be also used to partially compensate the frequency dependent attenuation, by enhancing the upper part of the signal spectrum.

Finite aperture of the probe

The finite aperture implies a limitation of the spatial frequencies depending on the depth of the scatterers. The probe finite aperture implies a cut-off of azimuthal spatial frequencies of the received signals. In particular, consider a point like scatterer emitting a spherical wave impinging on the probe surface. The “local” spatial frequency sampled by the transducers will depend form the local angle of incidence of the spherical wave with respect to the probe plane. Since the maximum angle occurs at the probe boundaries, the probe size sets a limit to the spatial frequency content that can be acquired. Increasing the scatterer depth the wave

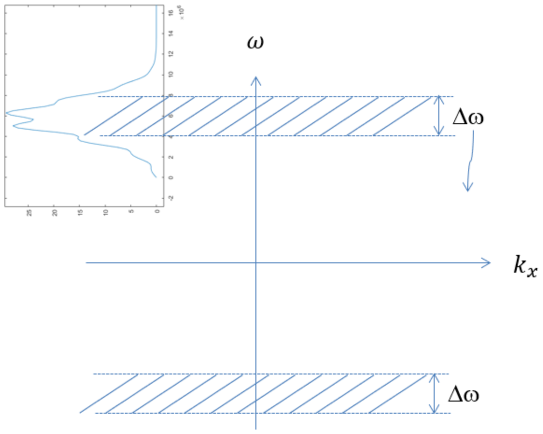


Figure 2.27 In the simplest case, the mask (shadowed region) is given by ones in the rectangle corresponding to the band with intensity above a certain threshold, and zeros outside. The rectangles shrink and move toward the k_x axis is increasing the depth z.

front impinging on the probe becomes flatter and the maximum angle decreases, decreases the maximum spatial frequency as well. A given angle of incidence is mapped to a line crossing the origin in the $k_x - \omega$ space, whose slope tends toward the ω axis as far as the angle decreases. This fact allows to define a region of validity with a shape of a double triangle where the SNR is acceptable. This region can be selected with a binary masking, possibly smoothing the boundaries to avoid ringing effects in the original domain. Notice that increasing the depth z of the scatterers, the triangles shrink as well.

Directivity of individual elements

The effect of the finite size of the transducers can be modeled as a convolution of the continuous wave-field incident on the array with a rectangle and a subsequent spatial sampling. In the $k_x - \omega$ space this operation results in a series of replicas of the spatial components, each one multiplied by a *sinc* function along k_x and invariant along ω. This induces a reduction of the spatial frequency components available for reconstruction, since, increasing the spatial frequencies index k_x, the lower values of the *sinc* multiplied to the signal decrease the SNR value below acceptable levels. To keep high the overall SNR a binary mask can be applied, with ones in the region $|k_x| < \textit{threshold}$ and 0 outside If the noise level is not too high (e.g. for shallow depths) the distortion induced by the *sinc* can be partially compensated for, adopting a mask whose entries are the inverse of the *sinc* function. For intermediate levels of noise an optimal masking can be derived, e.g. with a least squares solution.

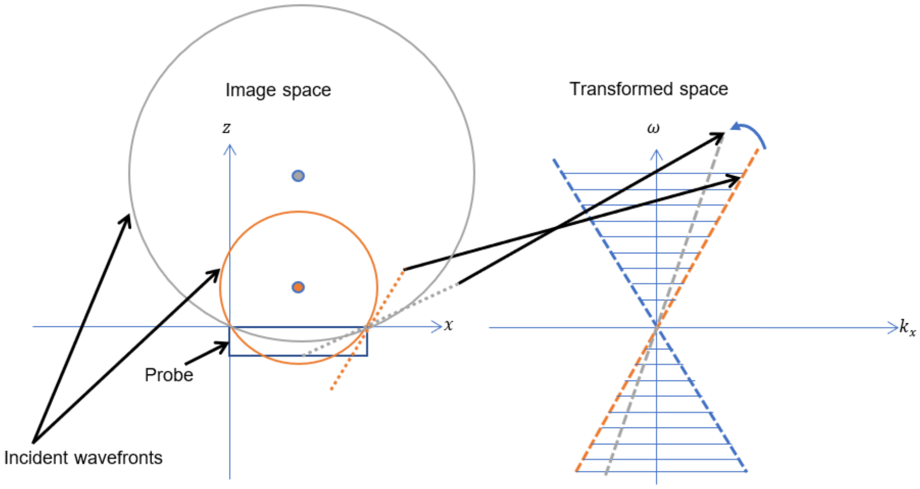


Figure 2.28 Increasing the scatterer's depth, the triangle area, corresponding to the masking equal to 1, is decreased in the transformed space.

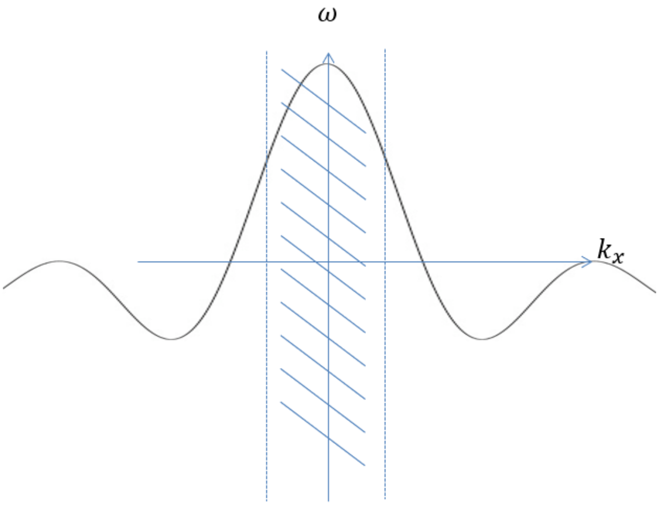


Figure 2.29 The *sinc* function is multiplied across k_x axis and replicated over each ω. In the simplest embodiment the binary mask is equal to 1 in the region of the *sinc* above a certain threshold 0 outside.

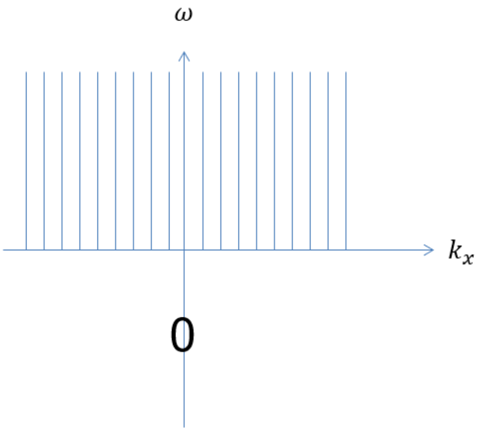


Figure 2.30 Hilbert filtering in $k_x - \omega$ domain mask with 1 for $\omega > 0$ and 0 for $\omega < 0$.

Pixel beamforming - IQ signal

In standard Delay-and-Sum beamforming the beamformed RF signal is typically filtered by an Hilbert filter in order to extract its envelope. From the envelope a subset of samples are visualized as image pixels. An alternative procedure consists in filtering the received RF signals with an Hilbert filter and perform beamforming with the resulting analytic signals (IQ signals). In this way the envelope extraction is reduced to a modulus operation, thus allowing to process just the samples necessary for the pixels to be visualized (pixel beamforming) with a relevant computational saving. Since the spectrum of the Hilbert filter is a $0-1$ step, with zeros on the negative frequencies, it is possible to implement the filtering of RF signals in the $k_x - \omega$ domain, by a simple mask of 1 for $\omega > 0$ and 0 for $\omega < 0$. This obviously halves the number of entries to be processed.

2.7 Conclusions

In this first chapter, I presented some new techniques in order to reconstruct the ultrasound images using plane waves which implies a platform of the equipment *software based* rather then *hardware based*. This new opportunity is possible thanks to the introduction of the GPUs and the new modules able to deal with the computational load required. I highlighted the advantages using algorithm in the frequency domain, which leads to a novel and patented beamforming algorithms based on seismic migration and the masking data in the $k_x - \omega$ space. The future work and natural evolution is to develop and test the algorithm in condition of absence of the homogeneity of the medium (phase aberration conditions and/or image degradation).

Chapter 3

Superdirective robust algorithms for linear array audio processing

3.1 Introduction and Motivation

The fundamental technique described in this chapter is based on *Beamforming*. It encompasses a large set of linear processing methodologies whose purpose is to strengthen the signal coming from a given direction, called *steering* direction (desired signal) and attenuate those coming from different directions (interfering signals). In practice, *beamforming* performs a spatial filtering of signals transmitted or received (the latter is the case under consideration) [28]. In fact, if the interfering signals have a spectrum that overlaps that of the desired signal, it is not possible to use temporal filtering to separate them and it is necessary to resort to spatial filtering. This type of filtering is obtained by spatially sampling, by means of the sensors, the wave field in incident propagation of the sensor array, or *sensor array*, and by adding the samples obtained in this way with suitable weight coefficients. This is done by the *beamformer*, whose output is a one-dimensional signal called the *beam* signal. The response of the *beamformer* to a plane wave of unitary amplitude incident on the array is defined as *beam pattern* (BP): it will therefore be a function of the direction of arrival of the wave and its frequency. Frequency-invariant beam patterns are often required by systems using an array of sensors to process broadband signals. In some experimental conditions (small devices for underwater acoustic communication), the array spatial aperture is shorter than the involved wavelengths. In these conditions, *superdirective* beamforming is essential for an efficient system. We present a comparison between two methods that deal with a data-independent beamformer based on a filter-and-sum structure. Both methods (the first one numerical, the second one analytic) formulate a mathematical convex minimization problem, in which the

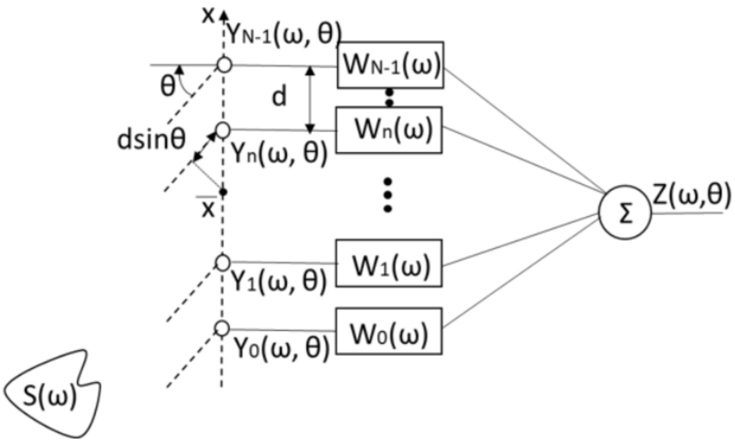


Figure 3.1 Beamforming with a linear microphone array configuration.

variables to be optimized are the filters coefficients or frequency responses. The goal of the optimization is to obtain a frequency invariant superdirective beamforming with a tunable tradeoff between directivity and frequency-invariance. We compare pros and cons of both methods measured through quantitative metrics to wrap up conclusions and further proposed investigations. A beamformer is an important data processing method in different fields (radar, sonar, biomedical imaging, and audio processing) to elaborate the signals coming from an array of sensors to get a versatile spatial filtering [28]. The beamformer can process both narrow-band and broad-band signals. The weights of the coefficients in the case of a data-independent beamformer do not depend on the array data. In filter-and-sum beamforming, each array sensor (microphones in our case) of the array feeds a transversal finite impulse response (FIR) filter (Figure 4.1) [29; 30; 31] and the filter outputs are summed-up by a convolution in the time-domain with an impulse response $w_{n,l}$ (N is the total number of sensors and L is the FIR filter's length) to produce the desired beam signal. The tapped delay line architectures are typically exploited to design a broadband spatial filter [32]. The beam pattern (BP) (Equation (3.2)) represents the beamformer spatial response in the far-field region and is a function of the direction of arrival θ (DOA) and the frequency $\omega/2\pi$. $S(\omega)$ is the input source (plane wave) and $Z(\omega,\theta)$ (Equation (4.3)) is the final output (Figure 4.1).

$$Z(\omega,\theta) = \mathbf{W}^H(\omega)\mathbf{Y}(\omega,\theta) = \sum_{n=0}^{N-1} W_n^*(\omega)Y_n(\omega,\theta) \tag{3.1}$$

We analyzed and implemented two different approaches in superdirective data-independent beamforming design described in literature [33; 34; 35]. To our knowledge, no one in the relevant literature has conducted this important and extremely deep comparison in order to apply further these methods in a real experimental context. *Superdirective* beamformers are sensitive to the errors in the uniform array characteristics. In the first numerical least-squares method proposed [33; 36] a desired beam pattern (BP) profile has to be chosen as input by the user; moreover the frequency and angle independent array characteristics affect the beamformer in a manner close to spatially white noise. Then the white noise gain $WNG(\omega)$ (Equation (3.6)) is the way to measure the performances and the robustness of the beamforming design. The method uses a threshold for its value to assure it, as a constraint for the WNG to lie above this value. In the second approach [34; 37; 38] the BP profile is optimized by the method itself without any user's choice [39], finding the most directive profile compatibly with the frequency invariance of the real BP. The robustness of the solution is guaranteed taking into account in the synthesis process the errors of the array, using the probability density function (PDF) of the microphones firstly introduced by Doclo and Moonen [35; 40], inside the analytic minimization of the cost function. The method uses a cost function written in a closed form, which has an analytic minimum. Both methods deal with data-independent beamformer, which not suffer of problems of signal cancellation in presence of echoes and multipath, as it happens frequently with adaptive methods, so they are very suitable for real applications (phased array for radar, etc.). This chapter is organized as follows. Section 3.2 describes the quantities useful to compare the results and the describes the mathematical background of the two methods implemented, Section 3.3 reports the conditions and the results of the comparison found, Section 3.4 comments the results, and Section 3.5 defines the conclusions and potential new directions of investigation.

3.2 Materials and Methods

Considering an equi-spaced (d is the constant distance between two microphones) linear array of N omnidirectional point-like sensors each connected to a FIR filter composed of L taps (Figure 4.1), the BP can be computed as:

$$B(\theta,\omega,w) = \sum_{n=0}^{N-1}\sum_{l=0}^{L-1} w_{n,l}g_{n,l}(\theta,\omega) \tag{3.2}$$

where

$$g_{n,l}(\theta,\omega) = e^{\left[-j\omega\left(\frac{n*d\sin(\theta)}{c}+l\tau_c\right)\right]} \tag{3.3}$$

$n*d$ is the distance of the n-th sensor to the center of the array, L is the FIR's length, T_c sampling interval of the FIR filter, c the speed of the acoustic waves in the medium, and

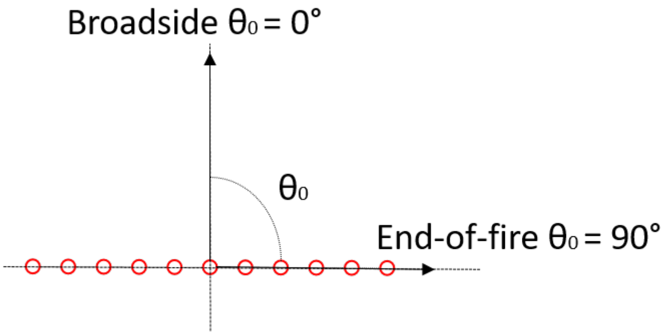


Figure 3.2 Broadside end end-of-fire steering directions (broadside orthogonal to the linear array, end-of-fire parallel to the direction of the array).

the weight $w_{n,l}$ is a real value representing the l-th tap coefficient of the n-th filter. We can rewrite (Equation (3.2)) as:

$$B(\theta,\omega) = \sum_{n=0}^{N-1} e^{\left[-j\omega\left(\frac{n*d\sin(\theta)}{c}\right)\right]} W_n(\omega) \tag{3.4}$$

where

$$W_n(\omega) = \sum_{l=0}^{L-1} w_{n,l} e^{[-j\omega(lT_c)]} \tag{3.5}$$

Superdirective beamfomers algorithms are extremely sensitive to spatial white noise and to small errors in the array characteristics. These errors are nearly uncorrelated from sensor to sensor and affect the beamformer in a manner similar to spatially white noise. Hence, the WNG is a commonly used measure for the robustness. The WNG is given by:

$$WNG(\omega) = \frac{|B(\theta_0,\omega)|^2}{\sum_{n=0}^{N-1} |W_n(\omega)|^2} \tag{3.6}$$

where θ_0 is the steering angle ($\theta_0 = 0°$ i.e., broadside geometry and $\theta_0 = 90°$ for end-of-fire) (Figure 3.2). An assessment of the beamformer performance independent of DOA, is attained through the array gain, which is the improvement in the signal to noise ratio obtained by using the array, when ambient noise is considered as noise. For isotropic noise, the array gain is called $Directivity$ and for a linear array is given by:

$$D(\omega) = \frac{2\,|B(\theta_0,\omega)|^2}{\int_0^\pi |B(\theta,\omega)|^2 \sin\theta d\theta} \tag{3.7}$$

In the methods of synthesis used, at a certain frequency, the higher is the $WNG(\omega)$, the lower is the $D(\omega)$ and the more frequency-invariant is the beampattern generated in the whole bandwidth of interest. Both methods implemented deal with the problem of robustness of super directive beam pattern (BP) against errors in the microphone response due to manufacturing limitations which is the focal and critical point in the real experimental scenario.

3.2.1 Proposed Methods

Robust Least-Squares Frequency-Invariant Beamformer Design (RLSFIB)

In the first method used [33; 41], the data-independent broadband least-squares frequency-invariant beamforming design (LSB), it directly constrains the WNG to lie above a given lower limit by solving a convex optimization problem. The idea behind the design is to optimally approximate a desired response, $B^*(\theta,\omega)$, by $B(\theta,\omega)$ in the least-squares (LS) sense. Typically, a numerical solution is obtained by discretizing the frequency range of the bandwidth into Q frequencies ω_q, where $q = 1, ..., Q$ and the angular range of the DOA into P angles θ_p, where $p = 1, ..., P$ and solving the resulting set of linear equations numerically. Since the number of discretized angles is typically greater than the number of sensors, i.e., $P > N$, the problem is therefore over-determined. A least-squares frequency-invariant beamformer design (LSFIB) is obtained by choosing the same desired response for all frequencies. This design inherently leads to superdirective beamformers for low frequencies if the wavelengths of the signals involved are larger than twice the sensor spacing and is therefore very sensitive to small random errors encountered in real-world applications. Since the WNG is a measure of the robustness of a beamformer the first algorithm design is then obtained by constraining the WNG above a threshold (Equation (3.8)). The algorithm imposes in the LS sense solution of the problem a constraint on the WNG that indirectly makes the array robust to errors in microphones. The idea behind the method is to incorporate a WNG constraint into the LSB design by adding the following quadratic constraint. The least-squares solution to this problem, which gives the smallest quadratic error by definition, is given by:

$$WNG(\omega_q) \geq \gamma \tag{3.8}$$

where γ is the lower bound for the WNG and

$$\min_{\mathbf{w}(\omega_q)} \left\| B(\theta_p,\omega_q) - B^*(\theta_p,\omega_q) \right\|_2^2 \tag{3.9}$$

where

$$\mathbf{w}(\omega) = [W_0(\omega), \ldots, W_{N-1}(\omega)]^T \tag{3.10}$$

Moreover, we impose that the desired signal from a given angle θ_0 (broadside in this case) remains not distorted. This method requires an iterative optimization (sequential quadratic programming *SQP*, *CVX* in Matlab) but is able to reach the global minimum since the problem formulation is convex. In this method, the trade-off between frequency-invariant and directivity is given by γ: the higher is its value, the more the BP pattern will be frequency-invariant and the lower the directivity.

Frequency-Invariant Beam Pattern Design (FIBP)

The second method analyzed [34; 37; 38; 41; 42; 43] uses a cost function over the probability density function PDF of errors (that must be known), granting in such way a solution that is optimal "*on average*". The algorithm proposes a method of FIR synthesis that allows for the design of a robust broadband beamformer with tunable tradeoff between frequency-invariance and directivity, without the need to impose a desired beam pattern [39]. The algorithm uses a cost function $J(\mathbf{w}, \mathbf{d})$ minimized with respect to the FIR filter's coefficient $\mathbf{w}$ (Equation (3.10)) and the values of the $P-1$ vector of desired beam pattern (DBP) D_p:

$$\mathbf{d} = \left[D_1 D_2 \ldots D_{\tilde{p}-1} D_{\tilde{p}+1} \cdots D_P\right] \tag{3.11}$$

The length of the $\mathbf{d}$ vector is $P-1$ because the desired beam pattern (DBP) at the steering angle (index $\tilde{p}$) $D_{\tilde{p}}$ equals 1 by construction and then is not a part of the $\mathbf{d}$ vector. The cost function $J(\mathbf{w}, \mathbf{d})$ to be minimized is given by next equation:

$$J(\mathbf{w}, \mathbf{d}) = \textstyle\sum_{p=1}^{P} \sum_{q=1}^{Q} \left[(1-K)\left|B\left(\theta_p, \omega_q, \mathbf{w}\right) - D_p e^{-j\omega_q \Delta}\right|^2 + K D_p^2\right] \tag{3.12}$$

We impose once again that the desired beam pattern (DBP) from a given angle θ_0 (in our case broadside) equals 1. Such a cost function (Equation (3.12)) is made up of two terms: the first term accounts for the adherence between the obtained BP and the DBP in a least-squares sense, and for all the frequencies and directions of interest, and the second one expresses the DBP energy. The relative weight of the two terms is ruled by the K parameter, whose values belong to the interval $[0, 1]$. Finally, to avoid any distortions of the received signals, the phase of the DBP should be linear, to produce a time delay referred to as Δ. The devised cost function has just one global minimum whose argument can be found in closed form by a computationally inexpensive procedure. The cost function previously described (Equation (3.12)) is based on the hypothesis that the characteristics of the sensors are perfectly known and not subject to deviations from the nominal values. Consequently, a synthesis based on this cost function, if applied to superdirective arrays, would produce a beamforming characterized by a high sensitivity to errors, especially in the gain and in the phase of the sensors. To overcome this drawback, a robust cost function is introduced. The method includes a subsequent modeling of the amplitude and phase errors of the microphones, of which we assume equal numerical values for the standard deviations. The relationships between mean values and standard deviations (variances) (Equation (3.13)) are expressed by the same article [34].

$$\mu_\gamma = \int f_\gamma(\gamma)\cos(\gamma)d\gamma \quad \sigma_\gamma = {\mu_\gamma}^2 \quad \sigma_a^2 = \int f_a(a)\,(a-1)^2 da \tag{3.13}$$

The cost function must be minimized with respect to the **w** coefficients of the FIR filters, and to the DBP values, contained in the vector with each discretized DOA except the steering angle. Considering the constraint on the DBP at the steering angle and the definition of directivity, the minimization of the DBP energy (i.e., the second term of the cost function), is equivalent to the maximization of the DBP directivity calculated in an approximate way on a discrete number of angles. In order to get a robust cost function, the formula (Equation (3.2)) in the previous paragraph can be replaced with:

$$B(\theta,\omega,w) = \sum_{n=0}^{N-1}\sum_{l=0}^{L-1} w_{n,l}A_n g_{n,l}(\theta,\omega) \tag{3.14}$$

where

$$A_n = a_n e^{(-j\gamma_n)} \tag{3.15}$$

The term A_n is included to take into account the gain and phase characteristics of the n-th sensor, supposed to be frequency invariant. The idea is to optimize the average performance, expressed through the weighted sum of the cost functions calculated for all possible combinations of sensor characteristics, using the probability density functions (PDFs) of the sensor characteristics as weights. For this purpose, a total cost function $J^{tot}(\mathbf{w},\mathbf{d})$ can be defined in the following equation:

$$J^{tot}(\mathbf{w},\mathbf{d}) = \int_{A_0} \cdots \int_{A_{N-1}} J\left(\mathbf{w},\mathbf{d},A_0,\ldots,A_{N-1}\right) * f_A\left(A_0\right)\ldots f_A\left(A_{N-1}\right) dA_0 \ldots dA_{N-1} \tag{3.16}$$

Consider initially the non-robust cost function defined by the expression (Equation (3.2)), the expression of the BP can be written as:

$$BP(\theta,\omega,\mathbf{w}) = \mathbf{w}\cdot\mathbf{g}^{T}(\theta,\omega) \tag{3.17}$$

where $\mathbf{w}$ is the row vector of size $M = NL$ and $\mathbf{g}^{T}$ is the column vector containing the complex exponentials that take into account the delays introduced by the propagation of the plane wave and the delay lines of the filters. Entering the previous relation (Equation (3.17)) in the expression (Equation (3.12)) we obtain:

$$J(\mathbf{w},\mathbf{d}) = \textstyle\sum_{p=1}^{P}\sum_{q=1}^{Q}\left\{(1-K)\left[\mathbf{w}\mathbf{g}_{p,q}^{\mathrm{T}}\mathbf{g}_{p,q}^{*}\mathbf{w}^{\mathrm{T}}\ -2D_{p}\mathbf{w}\,\mathrm{Re}\left\{\mathbf{g}_{p,q}^{\mathrm{T}}\right\}\right]+D_{p}^{2}\right\} \tag{3.18}$$

where $g_{p,q} = g(\theta_p,\omega_q)$. We rewrite the cost function in a matrix formulation using the following definitions:

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} & \mathbf{d} \end{bmatrix} \tag{3.19}$$

$$\mathbf{G} = (1-K)\sum_{p=1}^{P}\sum_{q=1}^{Q}\mathbf{g}_{p,q}^{\mathrm{T}}\mathbf{g}_{p,q}^{*} \tag{3.20}$$

$$\overline{\mathbf{g}}_{p}^{\mathrm{T}} = -(1-K)\sum_{q=1}^{Q}\mathrm{Re}\left\{\mathbf{g}_{p,q}^{\mathrm{T}}\right\} \tag{3.21}$$

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} \overline{\mathbf{g}}_{1}^{T}\overline{\mathbf{g}}_{2}^{T}\cdots\overline{\mathbf{g}}_{\tilde{p}-1}^{T}\overline{\mathbf{g}}_{\tilde{p}+1}^{T}\cdots\cdot\cdot\overline{\mathbf{g}}_{p}^{T} \end{bmatrix} \tag{3.22}$$

$$u = Q \tag{3.23}$$

$$\mathbf{U} = u\mathbf{I}_{P-1} \tag{3.24}$$

$$\mathbf{r}^{\mathrm{T}} = \begin{bmatrix} -\overline{\mathbf{g}}_{\tilde{p}}^{\mathrm{T}} \\ \mathbf{0}_{P-1} \end{bmatrix} \tag{3.25}$$

where $\tilde{p}$ indicates the index of the direction of steering, the apices ^{T} the transpose, and * the conjugate complex, $\mathbf{I}_{P-1}$ is the identity matrix of dimensions $(P-1)(P-1)$ and $\mathbf{0}_{P-1}$ is a column vector of size $P-1$ whose elements are all zeros. Introducing the matrix $\mathbf{M}$ of dimensions $(M+P-1)(M+P-1)$ defined as follows:

$$\mathbf{M} = \begin{bmatrix} \mathbf{G} & \mathbf{A}^{\mathrm{T}} \\ \mathbf{A} & \mathbf{U} \end{bmatrix} \tag{3.26}$$

the cost function to be minimized becomes:

$$J(\mathbf{w},\mathbf{d}) = \mathbf{v}\mathbf{M}\mathbf{v}^{\mathrm{T}} - 2\mathbf{v}\mathbf{r}^{\mathrm{T}} + u \tag{3.27}$$

whose *analytic* minimum turns out to be:

$$\mathbf{v}_{\mathrm{opt}} = \mathbf{M}^{-1}\mathbf{r}^{\mathrm{T}} \tag{3.28}$$

The first M elements of $\mathbf{v}_{\mathrm{opt}}$ represent the optimized FIR's coefficients while the last $(P-1)$ elements present the optimized DBPs for all DOAs apart from the steering direction where the DBP is 1 for construction. The $\mathbf{M}$ matrix (Equation (3.26)) is positive defined and therefore *invertible*. The described process can be extended to the minimization of the robust cost function in the following way, replacing the expression of the BP given by the relation (Equation (3.17)), which contains the model of the errors in amplitude and phase of the microphones within the same cost function (Equation (3.12)). After some calculations, the matrix expression of the expected cost function is similar to the relation (Equation (3.27)) with only some modifications of the matrices involved:

$$J_e(\mathbf{w},\mathbf{d}) = \mathbf{v}\widetilde{\mathbf{M}}\mathbf{v}^{\mathrm{T}} - 2\mathbf{v}\widetilde{\mathbf{r}}^{\mathrm{T}} + u \tag{3.29}$$

where

$$\widetilde{\mathbf{r}}^{\mathrm{T}} = \mu_\gamma \mathbf{r}^{\mathrm{T}} \tag{3.30}$$

$$\widetilde{\mathbf{A}} = \mu_\gamma \mathbf{A} \tag{3.31}$$

$$\widetilde{\mathbf{M}} = \begin{bmatrix} \tilde{\mathbf{G}} & \tilde{\mathbf{A}}^{\mathrm{T}} \\ \tilde{\mathbf{A}} & \mathbf{U} \end{bmatrix} \tag{3.32}$$

Finally, the matrix $\widetilde{\mathbf{G}}$ of size M x M is defined as:

$$\widetilde{\mathbf{G}} = \begin{bmatrix} \left(1+\sigma_a^2\right) 1_L & \sigma_\gamma \mathbf{1}_L & \cdots & \sigma_\gamma \mathbf{1}_L \\ \sigma_\gamma \mathbf{1}_L & \left(1+\sigma_a^2\right) \mathbf{1}_L & \cdots & \sigma_\gamma \mathbf{1}_L \\ \vdots & \vdots & & \vdots \\ \sigma_\gamma \mathbf{1}_L & \sigma_\gamma \mathbf{1}_L & \cdots & \left(1+\sigma_a^2\right) \mathbf{1}_L \end{bmatrix} \otimes \mathbf{G} \tag{3.33}$$

The symbol $\otimes$ indicates the element-wise multiplication and $\mathbf{1}_L$ indicates a matrix of size $L \times L$ in which all the elements are equal to 1. The solution is given by:

$$\mathbf{v}_{\text{opt}} = \widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{r}}^{\mathrm{T}} \tag{3.34}$$

Once the synthesis FIR coefficients' of the beam pattern have been obtained, the expected beam power pattern (EBPP) can be computed through the expectation operator $E\{.\}$. The EBPP represents the mean of the actual beam power pattern and is considered the most likely quantity to compared the average performance of different superdirective design methods. The aim is to make a performance average with respect to the randomness of the gain and phase of every array sensor. While the function $B(\theta, \omega, \mathbf{w})$ represents the nominal beam pattern (where all the array sensors have the nominal gain and phase behavior), we can denote by $B_a(\theta, \omega, \mathbf{w})$ the actual beam pattern, i.e., the beam pattern in which the actual gain and phase (to be intended as realizations of random variables) of each sensor composing the array are duly considered. The EBPP, $B_e^2(\theta, \omega, \mathbf{w})$, is defined as the mean of the actual beam power pattern, as follows: $B_e^2(\theta, \omega, \mathbf{w}) = E\{|B_a(\theta, \omega, \mathbf{w})|^2\}$. Working on the expectation operator, the following equation can be reached [34]:

$$\begin{aligned} B_e^2(\theta, \omega, \mathbf{w}) = \sigma_\gamma |B(\theta, \omega, \mathbf{w})|^2 + \left(1 + \sigma_a^2 - \sigma_\gamma\right) \sum_{n=0}^{N-1} |H_n(\omega)|^2 = \\ \sigma_\gamma |B(\theta, \omega, \mathbf{w})|^2 + \eta \sum_{n=0}^{N-1} |H_n(\omega)|^2 \end{aligned} \tag{3.35}$$

where

$$\eta = 1 + \sigma_a^2 - \mu_\gamma^2 \tag{3.36}$$

Another important parameter to evaluate the synthesis' performances is given by the generalized directivity D_G [44] (Equation (3.37)), where k is the wavenumber $k = 2\pi\, f/c$, $u = sin(\theta)$ and $u_0 = sin(\theta_0)$ is the steering direction. The n-th element (microphone) is placed at the position x_n and the m-th at the position x_m.

$$D_G = \frac{\eta \sum_{n=1}^{N} |w_n|^2 + \mu_\gamma^2 \sum_{m=1}^{N} \sum_{n=1}^{N} w_m^* w_n}{\eta \sum_{n=1}^{N} |w_n|^2 + \mu_\gamma^2 \sum_{m=1}^{N} \sum_{n=1}^{N} w_m^* w_n \exp\left[jk\left(x_m - x_n\right) u_0\right] \operatorname{sinc}\left[k\left(x_n - x_m\right)/\pi\right]} \tag{3.37}$$

The generalized directivity D_G represents the directivity of the EBPP obtained by using (Equation (3.7)) and replacing $|B(\theta, \omega)|^2$ with B_e^2.

3.3 Results

Both algorithms and the simulations have been developed and performed under Matlab framework [45; 46]. The geometry used for both simulations is broadside ($\theta_0 = 0°$) with:

- Aperture of the array 12 cm (length for *superdirective* beamforming).
- 8 equi-spaced microphones (good compromise between number of microphones and final price of the array).
- Sampling frequency 16 kHz (twice the maximum frequency of the signals involved to respect the Nyquist's theorem).
- L i.e., FIR's filter length 128 taps.
- Frequency band of interest (100–8000) Hz (typical bandwidth for audio applications).
- Value of the speed of sound $c = 340$ m/s.
- We used a grid of input data dividing the frequency bandwidth of interest $(100; 8000)$ Hz in values with a constant step of 50 Hz, and the range of the angles of the direction of arrival $(-90°;\ 90°)$ in values with a constant step of $1°$.

In these conditions, we find d = 1.71 cm then the limit for the spatial aliasing is 9.9 kHz, outside the band of the interest. To run both simulations we chose $P = 181$ and $Q = 157$, then for the second algorithm the computational load is related to $M = (N * L) = 1024$ iterations. For the first RLSFIB algorithm, we chose $\gamma = -10$ dB as the constraint whereas for the second FIBP algorithm we used K = 0.3 as numerical input as a trade-off between frequency-invariant and directivity. With this setting of parameters we get an *intermediate* synthesis between frequency-variant and frequency-invariant for both algorithms, but this is done to compare, in a fair way, their performances. The PDFs of the microphone gain and phase are assumed to be independent Gaussian functions with a mean equal to 1 and 0, respectively, and with a standard deviation of 0.02 for both. Once synthesized the two FIR coefficient sets for the two algorithms design, we built up and compared directivities (Figure 3.3), *WNGs* (Figure 3.4), directivity and generalized directivity vs standard deviation (Figure 3.5) and BPs (Figures 3.6 and 3.7).

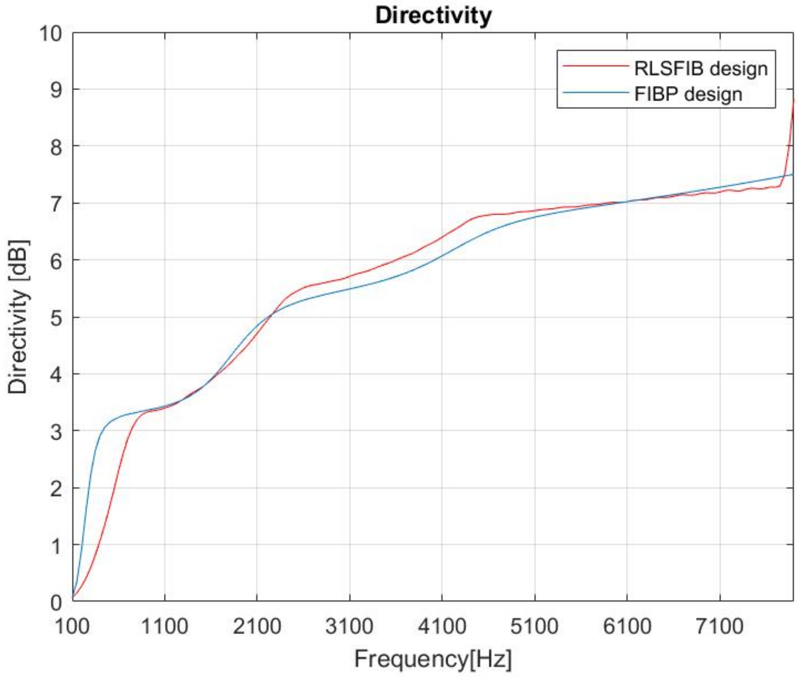


Figure 3.3 Directivity comparison: frequency-invariant beam pattern design (FIBP) blue, robust least-squares frequency-invariant beamformer design (RLSFIB) red.

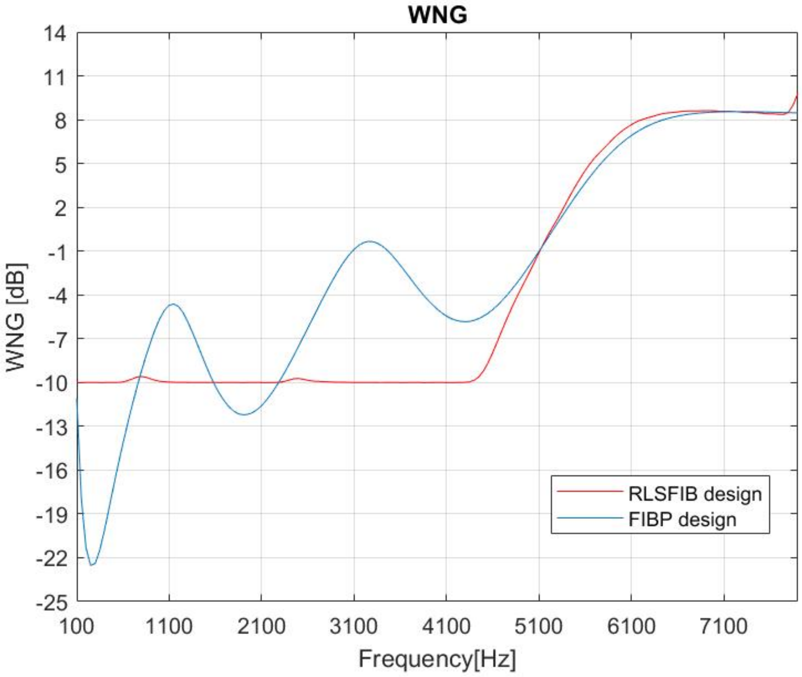


Figure 3.4 WNG comparison: FIBP blue, RLSFIB red.

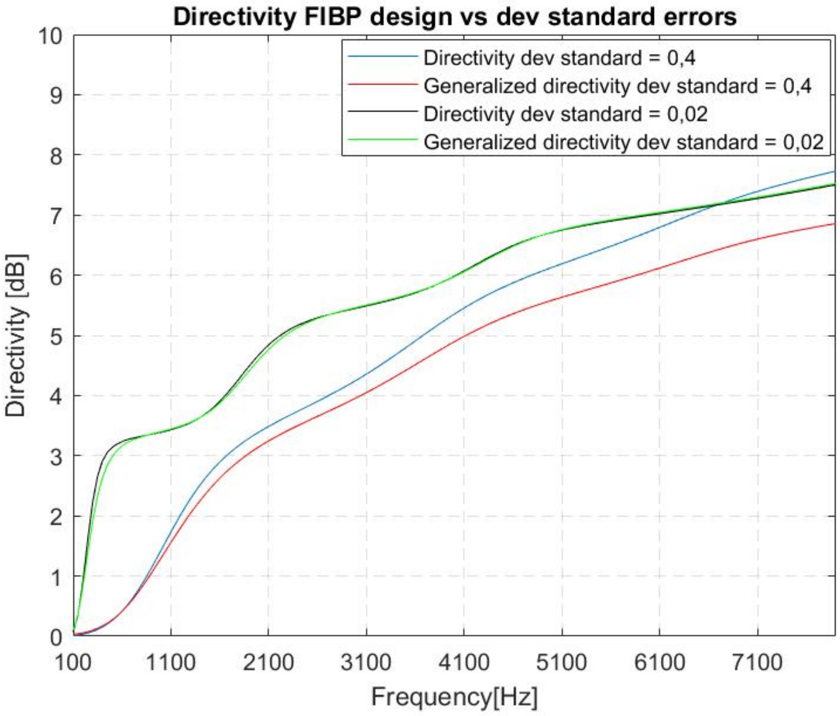


Figure 3.5 FIBP design: directivity and generalized directivity vs standard deviation comparison.

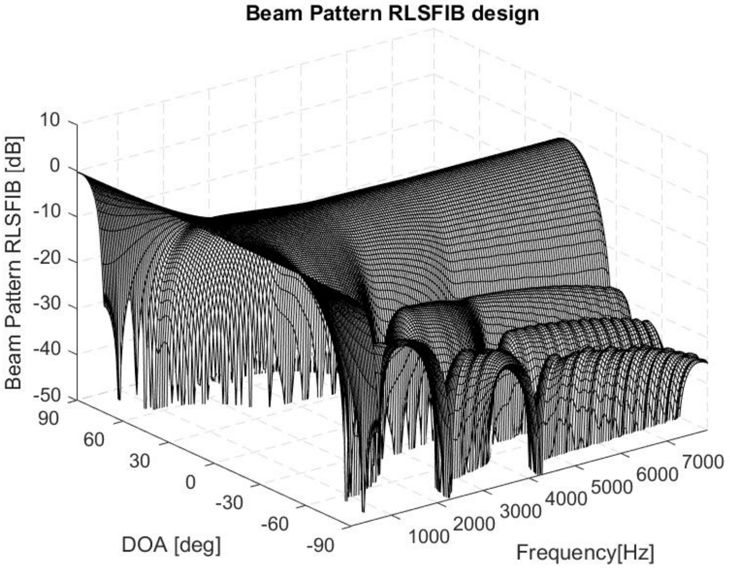


Figure 3.6 Beam pattern RLSFIB design.

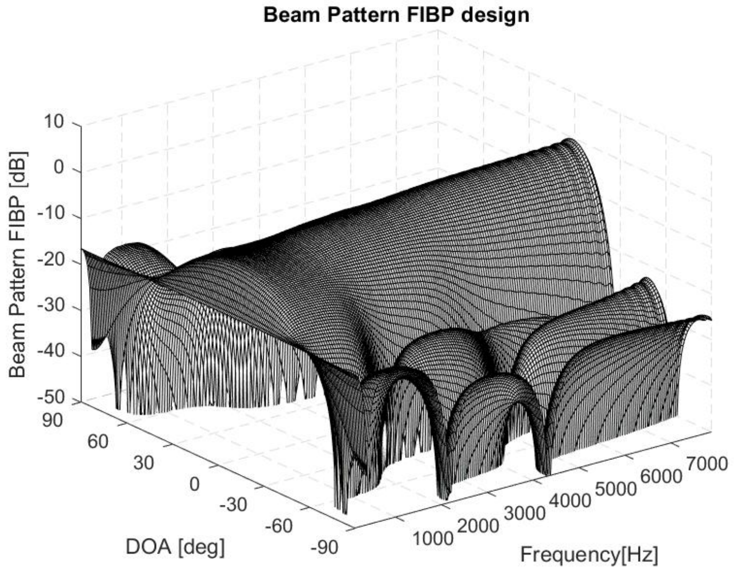


Figure 3.7 Beam pattern FIBP design.

3.3.1 RLSFIB Algorithm

The constrained least-squares problem was shown to be convex and therefore well-established methods for convex optimization, such as the *SQP* methods and *CVX*, may be used to solve the constrained problem. The results shown confirm that the RLSFIB design is capable of controlling the robustness of the resulting beamformer, which underlines the flexibility of this design procedure.

3.3.2 FIBP Algorithm

The robustness of the solution is achieved by taking into account the PDFs of the sensors' characteristics during the design phase. The EBPP (Figure 3.8) has been adopted to assess the performance of the beamformers obtained by the proposed synthesis method in addition to the traditional broadband BP graph and the curves of directivity and white noise gain.

3.4 Discussion

In the described simulations, we have chosen a geometry and a set-up of parameters that allows us to make a fair comparison between the performances of the two different design methods analyzed. In particular, we addressed a small linear array for audio capture with different purposes (hearing aids, audio surveillance system, videoconference system, multi-

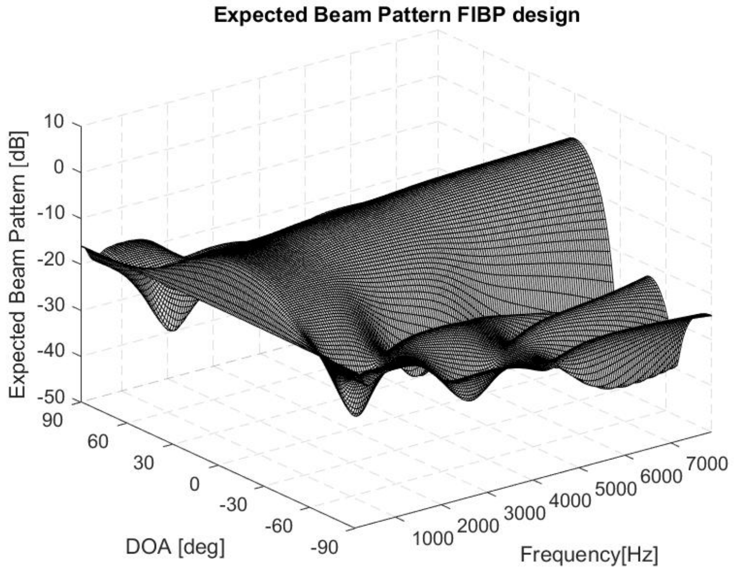


Figure 3.8 Expected beam pattern FIBP design.

media device, etc.). FIBP presents a more frequency-invariant BP and better performances at lower frequencies. With the parameters' choice, directivity and WNG are comparable for the two methods at the higher frequencies, but at lower frequencies WNG has an oscillating behavior for FIBP method, avoided by definition for the RLSFIB design. The oscillating behavior of the WNG in the FIBP method is due to the fact that, in general, $directivity$ and WNG are mutually reciprocals. In fact, when the derivative of directivity in the FIBP method is high (Figure 3.3, blue line), the WNG is low and vice versa. We can see that the first derivative of the directivity of the FIBP method is changing a lot in the range $(100;\ 5000)$ Hz: that is why the WNG is oscillating (Figure 3.4, blue line). The change of the first derivative is less pronounced for the directivity of the RLSFIB method (Figure 3.3, red line). Moreover, the threshold in the WNG forces this parameter to have a flatter and more stable curve (Figure 3.4, red line). The directivity at low frequencies provided by the RLSFIB method, lower than that provided by the FIBP method, justifies why the shape of the beam pattern (Figure 3.6) has a main lobe wider that of the FIBP method (Figure 3.7). The great advantage of the FIBP is the possibility to change the deviation standard of the distribution of the errors to highlight its impact: increasing the standard deviation, the difference between directivity and generalized directivity increases as well (Figure 3.5). This fact for FIBP design is very interesting because the difference between the generalized directivity and the nominal one provides useful insight on the expected impact that a given level of microphone mismatches induces on the system performance. This analysis allows us to choose the microphone accuracy, which is necessary to limit the (mean) performance decay at the level the user requires. A potential future work is to compare the performances of the two approaches for more frequency-variant rather then frequency-invariant synthesis, playing respectively with γ and K, comparing once again both performances of the algorithm of FIR's synthesis, using respectively, a lower γ and a higher K with respect the current values.

3.5 Conclusions

In this chapter we presented and compared the metrics of the synthesis of two different methods of simulation, following two different *philosophies*, to get the synthesis of FIR's coefficient filters for an efficient and robust superdirective beamformer to target audio applications in a real experimental scenario using a compact linear array of microphones. The main drawback of the two methods presented is the limitation on the choice of the cost function forced by the convexity conditions. In particular, there is no way to differentiate between main lobe and side lobe region, or to impose a worst-case design by minimizing the maximum of the side lobes. Moreover cost functions are quadratic so that low energy regions of BP are not very weighted in the cost function. Working with different representations (logarithmic) of BP could allow for a better shaping of low energy regions. For all these reasons, for further development of a new and better method of synthesis, the cost function should be modified to lose the convexity property. Then, to face the related problem of local minima, it would be necessary to take into account heuristics algorithms such as genetic algorithms. The simulations presented in this Chapter allow us to point out, for the two compared design methods, the tradeoff between performance (directivity), invariance, robustness (WNG), and sensor accuracy. They represent a starting base for further investigations the reader can perform, providing an insight on the parameters to modify in order to achieve the desired performance.

Chapter 4

Room Impulse Response evaluation with audio processing algorithms

4.1 Introduction and motivation

Given an unknown audio source, the estimation of time differences-of-arrivals (TDOAs) can be efficiently and robustly solved using blind channel identification and exploiting the cross-correlation identity (CCI). Prior "blind" works have improved the estimate of TDOAs by means of different algorithmic solutions and optimization strategies, while always sticking to the case $N = 2$ microphones. But what if we can obtain a direct improvement in performance by *just* increasing N? (Figure 4.1)

In this chapter we try to investigate this direction, showing that, despite the arguable simplicity, this is capable of (sharply) improving upon state-of-the-art blind channel identification methods based on CCI, without modifying the computational pipeline. Inspired by our results, we seek to warm up the community and the practitioners by paving the way (with two concrete, yet preliminary, examples) towards joint approaches in which advances in the optimization are combined with an increased number of microphones, in order to achieve further improvements.

Sound source localisation applications can be tackled by inferring the time-difference-of-arrivals (TDOAs) between a sound-emitting source and a set of microphones. Among the referred applications, one can surely list room-aware sound reproduction [47], room geometry's estimation [48; 49; 50; 51; 52], speech enhancement [53; 54] and de-reverberation [1; 55; 56]. Despite a broad spectrum of prior works estimate TDOAs from a known audio source [57; 58; 59], even when the signal emitted from the acoustic source is *unknown*,

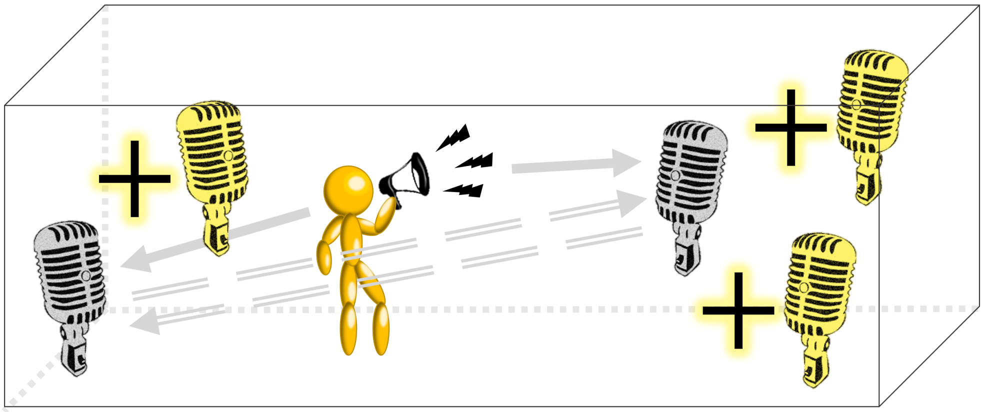


Figure 4.1 We are given an unknown sound-emitting source, where in the actual applicative scenario that we encompass, we have no prior knowledge about the sound source and can be therefore arbitrary. We are interesting in (robustly) inferring TDOAs in an (unknown as well) environment given a pool of microphones, using the the following principle. Given the pair of grey microphone, the audio that each of them acquires from the source (solid arrow) must "agree" with the other. That is, if any of the two mic could "hear" the other, the registered signal has to be the very same (dashed arrows). This is called *cross-correlation identity* and it was empirically studied in the case $N = 2$, only. In this Chapter we answer to what happens then if $N > 2$? Can we improve in robustness and/or accuracy in the estimate, for instance, by adding the yellow microphones?

TDOAs can be inferred by comparing the signals received at two (or more) spatially separated microphones [1; 55; 60; 61; 62] using the notion of **cross-corrlation identity** (CCI) - see Fig. 4.1. This is the key theoretical tool, not only, to make the ordering of microphones irrelevant during the acquisition stage, but also to solve the problem as *blind* channel identification [1; 55; 60; 61; 62], robustly and reliably inferring TDOAs from an unknown audio source (see Sec. 4.2).

However, when dealing with natural environments, such "mutual agreement" between microphones can be tampered by a variety of audio ambiguities such as ambient noise. Furthermore, each observed signal may contain multiple distorted or delayed replicas of the emitting source due to reflections or generic boundary effects related to the (closed) environment. Thus, robustly estimating TDOAs is surely a challenging problem and CCI-based approaches cast it as single-input/multi-output blind channel identification [1; 55; 60; 61; 62]. Such methods promotes robustness in the estimate from the methodological standpoint: using either energy-based regularization [60], sparsity [1; 55; 62] or positivity constraints [55], while also pre-conditioning the solution space [1].

In this Chapter, we posit that there is a much easier practical strategy to ensure robustness while inferring TDOAs: *the possibility of exploiting a larger pool of microphones*. In fact, it is surprising to observe that, in prior state-of-the-art methods based on CCI, experimental evidences are provided for the case $N = 2$ microphones only [1; 55; 60; 61; 62]. Despite such a number is the bare minimum to solve the problem, it remains elusive whether $N > 2$ can, *by itself*, boost the estimate of TDOAs in accuracy/robustness, without requiring any changes in the computational pipeline. In fact, since all methods [1; 55; 60; 61; 62] can *theoretically* accommodate for $N > 2$, why not test them in such a regime?

The purpose of this work is to answer this question and back up the investigation of state-of-the-art methods based on CCI [1; 55; 60; 61; 62] in handling the case $N > 2$. Our goal is to understand whether an increase in the number of microphones will translate into an improved TDOAs estimate.

Our contributions. Among all state-of-the-art methods based on CCI [1; 55; 60; 61; 62], we consider the most effective one: IL1C [1]. Despite, in fact, recent advances were essentially devoted in estimating TDOAs given a known audio source in how to exploit the TDOAs [57; 58; 59], the problem of achieving the very same task while being blindly unaware of which audio source was deployed can be still efficiently and effectively solved using methods such as [1; 55; 60; 61; 62] out of which IL1C [1] is the best in terms of robustness and efficacy. IL1C infers TDOAs by solving a stack of convex problems through a weighted sparsity promoting (ℓ^1) constraint, leveraging the non-negativity of the Acoustic Impulse Response (AIR), from which TDOAs are easily estimated using peak finding [1]. To guarantee robustness while inferring TDOAs, in addition to sparsity, IL1C [1] takes advantage of a pre-conditioning mechanism to better initialize the AIRs using a data-driven initialization.

We setup a broad experimental validation, measuring the performance of IL1C on a variety of audio signals, going well beyond the experimental evidences provided in [1]. That is, on the one side, we test the effectiveness of this method on many more audio signals: synthetic (pink and white) noise and a list of natural audio sources (two different plastic rustles – obtained from either scraping a bag or compacting a bottle before thrashing, adult male voice, dog barking, stapler and hand-clapping). On the other side, differently to [1], we do not only consider the case $N = 2$, but we also consider a bigger number of microphones $N = 3, 4, 5, 10$ motivated by encompassing the scenario of multiple microphones.

As our experimental evidences show, we stably register improvements in either the robustness (towards outliers) or the accuracy in retrieving the peaks of the AIRs. We evaluate on that by exploiting two well known performance metrics as in prior work [1; 55; 60; 61; 62], and, although there are (sound-specific) cases in which one of the two indicators show a damaged performance, still the other one shows improvements. In fact, we can demonstrate that, across the wide number of different audio sources that we consider, the general trend is that, while averaging the absolute improvement across different choices for $N = 3,4,5$ or 6 over the baseline case $N = 2$, we score positive signed improvements (see Table 4.3) which seems not to be effected on whether the source is emitting synthetic or natural sounds. At the same time, we register a very positive trend if we are enriched by an oracle knowledge of the optimal number of microphones that have to be arranged before the acquisition stage. In such a case, we *always* register positive improvements over the baseline $N = 2$, which are, in the worst case, by $+3\%$, while achieving more than $+28\%$ as well.

Inspired by our evidences, in Section 4.5, we attempt to warm up future research directions towards optimization approaches which explicitly account for the case $N > 2$. Although proposing a new paradigm which falls inside this new family of methods is out of scope for us, we still deem interesting to inform practitioners about the effect of two straightforward modifications of IL1C [1], using either an incremental pre-conditioning or an ensemble strategy - see Section 4.5. Regardless of the scores results (in which the ensemble strategy is better than the incremental pre-conditioning, while also improving the baseline IL1C method [1]), we deem our effort to be effective in stimulating the research towards methods which explicitly account for the case $N > 2$ when dealing with an unknown audio source.

4.2 Problem Statement & Related Work

Let us formalize the problem of inferring TDOAs, so that we can easily refer to prior related works. Let us consider a given enviroment (e.g., a room) of unknown geometry in which an audio source emits together with N microphones (Figure 4.2): the task is to reconstruct TDOAs.

Let $\vec{h}_n$ represent the AIR (Acoustic Impulse Response) from a fixed audio source and the n-th microphone, $n = 1,\dots,N$. The signal $\vec{h}_n$ is sampled into a fixed number of temporal bins $\vec{h}_n(k)$. The signal $y_n(k)$ received at microphone n can be written as the discrete convolution between the transmitted signal $x(k)$ and the n-th AIR:

$$y_n(k) = \vec{h}_n(k) * x(k) + \nu_n(k), \quad n = 1,\dots,N \tag{4.1}$$

where $\nu_n(k)$ is an additive noise term. The ultimate goal of the problem is leveraging the measurements $y_n(k)$ to recover the AIRs $\vec{h}_n(k)$ without knowing the transmitted signal $x(k)$.

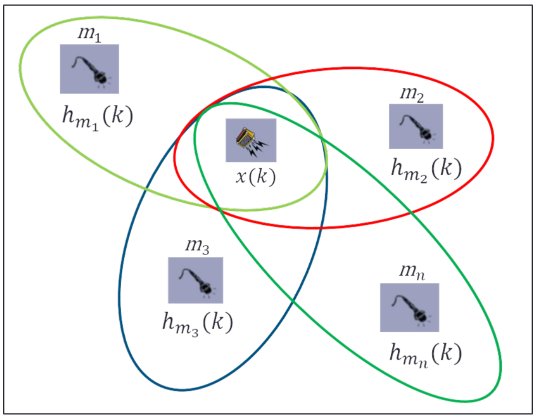


Figure 4.2 Multiple Cross-Correlation Identities: graphical interpretation.

Cross-correlation identity. When multiple microphones are recording the same audio source, the acquisition should be independent of the order of the microphones according to the following constraint:

$$\vec{h}_m(k) \ast \vec{h}_n(k) \ast x(k) = \vec{h}_n(k) \ast \vec{h}_m(k) \ast x(k), \tag{4.2}$$

for every pairs of microphones m and n. In turn, using eq. (4.1), we rewrite eq. (4.2) as $\vec{h}_m(k) \ast y_n(k) = \vec{h}_n(k) \ast y_m(k)$. Hence, by using the well-known fact that the convolutional operator $\ast$ is linear, we obtain

$$Y_n\vec{h}_m = Y_m\vec{h}_n, \quad m,n = 1,\dots,N \tag{4.3}$$

where $\vec{h}_n$ is the column vector which stacks the AIRs $\vec{h}_n(k)$ by columns, while Y_n is the diagonal-constant matrix with first row and column given by $[y_n(k-K+1), y_n(k-K), \dots, y_n(k-K-L+2)]$ and $[y_n(k-K+1), y_n(k-K+2), \dots, y_n(k), 0, \dots, 0]^\top$ respectively, with K and L being the signal length and channel length.

In order to solve for (4.3), a number of prior approaches have took advantage of regularization [1; 60]. For instance, Tong *et al.* [60] have framed the problem of TDOAs estimation as the following regularized Least Squares fitting

$$\min_{\vec{h}_1,\dots,\vec{h}_N} \sum_{m \neq n} \|Y_n\vec{h}_m - Y_m\vec{h}_n\|_2^2 \quad \text{s.t.} \sum_i \|\vec{h}_i\|_2^2 = 1 \tag{4.4}$$

to ensure robustness by means of regularization. Clearly, adding a regularization term is fundamental to avoid the optimization to converging towards the trivial solution $\vec{h}_n = 0$ for every $n = 1,\dots,N$. Remarkably, the real problem is choosing a proper regularization term.

In fact, when using ℓ^2 regularization - as in eq. (4.4), the solution can be computed in closed-form by means of eigenvalue decomposition [60]. Unfortunately, L^2 regularization neglects some crucial physical properties of the expected solution - such as non-negativity [61; 63].

Additionally, requiring $\sum_i ||\vec{h}_i||_2^2 = 1$ as in (4.4) makes the AIRs to be co-prime[64] and constraint each of them to have a fixed norm - each of such requirements are likely to introduce numerical instabilities and artifacts during the optimization process. As a remedy for this, sparsity priors have been successfully applied to a broad spectrum of prior work in TDOAs estimation [47; 48; 49; 50; 51; 52] [64], while also encompassing speech enhancement [65] and de-revereberation [55]. Therefore, as to impose sparsity in the reconstructed $\vec{h}_n$, replacing the L^2 regularization in eq. (4.3) with a L^1 counterpart [1; 55; 62] seems an appealing solution. Precisely, in [62] a L^1-norm penalty was added to eq. (4.4), yielding

$$\min_{\vec{h}_1,..\vec{h}_N} \sum_{m \neq n} ||Y_n\vec{h}_m - Y_m\vec{h}_n||_2^2 \text{ s.t. } \begin{cases} \sum_i ||\vec{h}_i||_2^2 = 1 \\ \sum_i ||\vec{h}_i||_1 < \varepsilon \end{cases} \tag{4.5}$$

Unfortunately, a quadratic optimization subject to mixed quadratic and linear constraints do not preserve the convexity of (4.4). Hence, the method as in (4.5) is prone to local solutions.

To cope with this issue, we can relax eq. (4.3) into

$$\min_{\vec{h}_1,..,\vec{h}_N} \sum_{m \neq n} ||Y_n\vec{h}_m - Y_m\vec{h}_n||_2^2 \text{ s.t. } \begin{cases} |\vec{h}_1(a)| = 1 \\ \sum_i ||\vec{h}_i||_1 < \varepsilon \end{cases} \tag{4.6}$$

where the fixed index a is an anchor constraint [55] which makes the optimization in eq. (4.6) convex and more robust towards spectrum holes of $x(k)$ if compared to eq. (4.4).

However, the anchor constraints $|\vec{h}_1(a)| = 1$ together with $\sum_i ||\vec{h}_i||_1 < \varepsilon$ penalizes all the peaks intensities but one, often leading to peak cancellations in noisy conditions. The approach of [55] has been modified in [61] adding an ancillary non-negativity constraint on the AIRs

$$\min_{\vec{h}_1,..,\vec{h}_N} \sum_{m \neq n} ||Y_n\vec{h}_m - Y_m\vec{h}_n||_2^2 \text{ s.t. } \begin{cases} |\vec{h}_1(a)| = 1 \\ \sum_i ||\vec{h}_i||_1 < \varepsilon \\ \vec{h}_1,..,\vec{h}_N \geq 0 \end{cases} \tag{4.7}$$

where, for each n, $\vec{h}_n \geq 0$ means $\vec{h}_n(k) \geq 0$ for each k. Non-negativity yields increased robustness against noise by further regularizing the problem [66; 67], but it is arguably limited in addressing the limitations of the anchor constraints.

To directly tackle the latter problem, Crocco *et al.* [1] replaced the anchor constrained $|\vec{h}_1(a)| = 1$ by means of the introduction of a slack variables $\vec{p}_1, \ldots, \vec{p}_N$ such that

$$\min_{\vec{h}_1,..,\vec{h}_N} \sum_{m \neq n} \|Y_n \vec{h}_m - Y_m \vec{h}_n\|_2^2 \text{ s.t. } \begin{cases} \vec{p}_n^\top \vec{h}_n = 1 \\ \sum_i \|\vec{h}_i\|_1 < \varepsilon \\ \vec{h}_1, .., \vec{h}_N \geq 0 \end{cases} \tag{4.8}$$

In this way, all the components of the AIRs are equally taken into account without privileging the a-th of the $\vec{h}_1$. At the same time, differently from eqs. (4.5), (4.6), the constraints as in eq. (4.8) are differentiable, since $\vec{h}_1, .., \vec{h}_N \geq 0$ implies $\sum_i \|\vec{h}_i\|_1 = \sum_i \sum_a \vec{h}_i(a)$. The optimization problem as in eq. (4.8) is convex with respect to $\vec{h}_n$ while fixing the slack variables $\vec{p}_n$ and vice-versa. Inspired by this consideration, Crocco *et al.* [1] proposed an alternated iterative scheme in which, $\vec{p}_n$ are firstly initialised as the AIRs computed using Tong *et al.* method's [60], while cycling between: 1) optimizing for $\vec{h}_1, .., \vec{h}_N$ in (4.8) given $\vec{p}_1, .., \vec{p}_N$ and 2) use the newly computed AIRs to update $\vec{p}_n$ for every n. As discussed in [1], although the proposed initialization introduces a distortion in the amplitude of the AIRs, then the iterative procedure is able to compensate. More crucially, initializing $\vec{p}_n$ at the first iteration by using [61] makes the slack variable sparse. Therefore, the first two constraints as in eq. (4.8) make the computed AIRs sparse again. Such property is preserved during optimization because of the updating scheme in which slack variables at a given iteration are selected as the solution of eq. (4.8) as in the prior iteration.

A sharp limitation of prior blind methods. None of the prior methods [1; 55; 60; 61; 62] was generalized to the case $N > 2$. Despite $N = 2$ has the appealing formal property of achieving minimality among the number of microphones necessary to solve the optimization problem, still it remains elusive from a practical standpoint whether allocating for a bigger number N of microphones can effectively boost the estimate of TDOAs. And, in the likely event of this case effectively happening, are we improving upon robustness towards outliers or in accuracy as well? The scope of the present work is to answer this question.

4.3 Multiple Cross-Correlation Identities

In this Section, we evaluate the effect of increasing the number of microphones when tackling the problem of inferring TDOAs by means of well established notion of cross-correlation identity (CCI) [1; 55; 60; 61; 62]. In details, we focus on IL1C [1], the best out of such class of approaches: we optimize equation (4.8) for the case of $N = 2, 3, 4, 5, 10$. By doing so, we are capable of starting from the minimal setup from which the problem can be solved ($N = 2$): note that this is the experimental playground analysed in prior works [1; 55; 60; 61; 62]. Differently, for the sake of inspecting whether a higher number of microphones can provide an improvement in the estimate of TDOAs, we also consider the cases $N = 3,4,5$ up to the $N = 10$ microphones. This range of variability in N is, in our opinion, a good trade-off between having a sufficiently large number of acquisition devices, while still framing a scenario which can be still useful from the applicative standpoint.

Let us briefly introduce the types of source signals considered in this study, as well as the reproducibility and implementation details about our evaluation protocol and the error metrics to check on performance. The results of our analysis are reported in Tables 4.1 and 4.2, while showing relative and absolute improvements in Table 4.3. An extended discussion on our findings is reported in Section 4.4.

The different types of source signals we considered. We considered two types of synthetic audio signals *white noise* and *pink noise*, which differ among each others in the considered frequencies of their spectrum (all vs. only wide ones, respectively). We also encompass a broad list of natural sounds as audio source: ***plastic rustle no. 1 (bag)***, ***plastic rustle no. 2 (bottle)***, ***adult male voice***, ***dog barking***, ***stapler*** and ***hand-clapping***, all of them characterized by a narrow frequency spectrum.

Evaluation. We run experiments by considering any of the source audio signals described in the prior paragraph located in the same environment analyzed in [1]. We model the Acoustic Impulse Response (AIR) for each microphones as seven different peaks, corresponding to one direct path source-microphones, together with six (first-order) reflections. In details, we applied the simulating image method as in [65], using a reflection coefficient of 0.8. We also introduce another degree of variability, by considering different Noise-to-Signal ratios (s). This is done by injecting additive Gaussian white noise on the output microphones according to the following specs: 0 dB, 6 dB , 14 dB, 20 dB and 40 dB. This induces a signal-to-noise ratio $s = 10^{-\mathrm{dB}/20}$ from the following inverse relationship $\mathrm{dB} = 20\log_{10}(1/s)$. When running the optimization (4.8) of IL1C [1], we take advantage of the official code directly shared by authors, while following the same pre-processing and evaluation techniques as in the referred prior work. In addition, as done (4.8), we perform model selection by doing cross-validation on the threshold ε which controls the sparsity-promoting constraint.

Error metrics. Once the AIRs have been computed through (4.8), we apply the peak finding method of [1] and we evaluate performance by means of two standard error metrics: the Average Peak Position Mismatch ($\mathscr{A}_{\mathrm{PPM}}$) and the Average Percentage of Unmatched

Table 4.1 Average Peak Position Mismatch ($\mathscr{A}_{\text{PPM}}$)) metrics for IL1C [1] when $N = 2, 3, 4, 5, 10$. Synthetic source noise are denoted in italic, while bold italic refers to the natural source signal considered in this study. For each source signal considered, we provide an histogram visualization to better perceive the variability of the error metrics: the range of variability of each data bar is normalized within each different source signal. A better performance corresponds to a lower ($\mathscr{A}_{\text{PPM}}$)) value or, equivalently, to a lower bar. The value s quantifies the impact of the additive Gaussian noise on the registered signal: we span the case $s = 0.01$ (easier) to $s = 1$ (harder), while transitioning on the intermediate cases $s = 0.1, 0.2$ and $s = 0.5$.

Method	N	Setup	*white noise*					*pink noise*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[10]	0.2153	0.2636	0.3102	0.7642	2.0156	4.984	4.5005	4.8002	4.5834	5.2774
IL1C	3	ours	0.2238	0.222	0.2528	0.8388	1.6932	4.3063	5.5322	4.2378	5.2365	4.7675
IL1C	4	ours	0.2398	0.2617	0.4049	0.9531	2.1781	4.3561	5.7132	5.2493	5.2118	5.4812
IL1C	5	ours	0.2415	0.2585	0.3318	1.1083	2.1126	4.3109	5.2371	4.76	5.59	6.1503
IL1C	10	ours	0.2495	0.2815	0.4609	1.0902	2.1065	4.529	4.7427	4.7853	6.0846	6.0842

Method	N	Setup	*plastic rustle no. 1 (bag)*					*plastic rustle no. 2 (bottle)*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[10]	0.2489	0.2419	0.4454	1.4199	2.8224	4.5856	2.8086	3.8703	4.1446	4.3346
IL1C	3	ours	0.2519	0.4724	0.2879	1.2378	2.8866	4.362	4.2216	4.7789	5.045	5.614
IL1C	4	ours	0.2598	0.254	0.9009	1.2666	2.7452	4.5136	5.0302	4.107	4.44	6.0028
IL1C	5	ours	0.2581	0.3368	0.5515	1.3383	3.1889	3.483	4.7622	4.3169	5.2023	5.8363
IL1C	10	ours	0.2731	0.2766	0.3143	1.24	2.3357	5.8363	5.8825	5.9941	5.8367	5.9526

Method	N	Setup	*adult male voice*					*dog barking*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[10]	0.2654	0.5665	0.6481	3.3806	2.93	0.2378	0.5777	1.5409	2.2086	4.5964
IL1C	3	ours	0.2728	0.4416	0.3726	2.0912	3.229	0.2618	0.5487	1.1899	2.2948	3.9943
IL1C	4	ours	0.2636	0.358	0.8295	1.6993	2.9215	0.2563	0.2802	1.0446	2.0303	3.058
IL1C	5	ours	0.2641	0.4972	1.0297	1.6344	2.0912	0.2833	0.4283	0.5584	1.6376	3.2308
IL1C	10	ours	0.2906	0.4313	0.6133	1.6644	2.3752	0.2744	0.3589	0.6838	1.7842	2.7217

Method	N	Setup	*stapler*					*hand-clapping*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[10]	3.7404	4.9421	3.8708	3.479	4.478	3.3215	4.7464	3.7643	3.8144	4.8355
IL1C	3	ours	3.5635	4.5142	3.9549	3.6192	5.2086	3.5635	4.5142	3.9549	3.6192	5.2086
IL1C	4	ours	2.458	3.588	4.3245	3.6958	4.7859	4.8498	3.658	3.7566	5.3835	4.9137
IL1C	5	ours	3.2887	3.3599	3.2826	3.2742	5.5281	3.6248	5.0005	4.8891	5.4968	6.1206
IL1C	10	ours	3.3701	3.5617	4.0655	4.1534	5.4883	3.6174	4.3315	4.6524	5.6151	6.2386

Table 4.2 Average Percentage of Unmatched Peaks ($\mathscr{A}_{\text{PUP}}$) metrics for IL1C [1] when $N = 2,3,4,5,10$. Synthetic source noise are denoted in italic, while bold italic refers to the natural source signal considered in this study. For each source signal considered, we provide an histogram visualization to better perceive the variability of the error metrics: the range of variability of each data bar is normalized within each different source signal. A better performance corresponds to a lower ($\mathscr{A}_{\text{PUP}}$)) value or, equivalently, to a lower bar. The value s quantifies the impact of the additive Gaussian noise on the registered signal: we span the case $s = 0.01$ (easier) to $s = 1$ (harder), while transitioning on the intermediate cases $s = 0.1, 0.2$ and $s = 0.5$.

Method	N	Setup	*white noise*					*pink noise*				
			s = 0.01	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**	**s = 0.01**	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**
IL1C	2	[10]	0	0.0014	0.0114	0.0557	0.2	0.7679	0.7179	0.6429	0.6	0.5964
IL1C	3	ours	0	0.0019	0.0095	0.0714	0.179	0.75	0.7238	0.6643	0.5381	0.5476
IL1C	4	ours	0	0.0043	0.0293	0.105	0.225	0.725	0.7125	0.5893	0.5125	0.5696
IL1C	5	ours	0	0.0046	0.016	0.0983	0.2514	0.74	0.6771	0.5886	0.5114	0.5057
IL1C	10	ours	0	0.0058	0.0265	0.1075	0.2367	0.7429	0.6886	0.5721	0.455	0.5086

Method	N	Setup	***plastic rustle no. 1 (bag)***					***plastic rustle no. 2 (bottle)***				
			s = 0.01	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**	**s = 0.01**	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**
IL1C	2	[10]	0	0	0.025	0.15	0.2964	0.7393	0.75	0.7107	0.6	0.5821
IL1C	3	ours	0	0.0262	0.0095	0.1381	0.2952	0.7238	0.7405	0.6524	0.5333	0.531
IL1C	4	ours	0	0	0.0857	0.1357	0.2804	0.725	0.7036	0.6214	0.5196	0.5304
IL1C	5	ours	0.0029	0.0071	0.0329	0.12	0.3343	0.7271	0.68	0.61	0.4743	0.4786
IL1C	10	ours	0	0	0.0043	0.1414	0.28	0.5461	0.5411	0.5396	0.5311	0.5296

Method	N	Setup	***adult male voice***					***dog barking***				
			s = 0.01	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**	**s = 0.01**	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**
IL1C	2	[10]	0	0.0321	0.075	0.3214	0.375	0	0.0214	0.1357	0.3321	0.4464
IL1C	3	ours	0	0.0095	0.0357	0.2833	0.3524	0	0.0286	0.1071	0.2619	0.4119
IL1C	4	ours	0	0.0161	0.0536	0.2036	0.4143	0	0.0018	0.0946	0.3089	0.3464
IL1C	5	ours	0	0.02	0.0757	0.1871	0.31	0	0.0286	0.0614	0.1886	0.3857
IL1C	10	ours	0	0.0157	0.0436	0.1829	0.2986	0	0.0157	0.0529	0.1786	0.3064

Method	N	Setup	***stapler***					***hand-clapping***				
			s = 0.01	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**	**s = 0.01**	**s = 0.1**	**s = 0.2**	**s = 0.5**	**s = 1**
IL1C	2	[10]	0.6643	0.7036	0.6321	0.6071	0.6036	0.6964	0.6964	0.6393	0.5571	0.6143
IL1C	3	ours	0.6048	0.6429	0.5643	0.4833	0.5524	0.6048	0.6429	0.5643	0.4833	0.5524
IL1C	4	ours	0.5607	0.5714	0.5607	0.5732	0.575	0.7	0.6571	0.625	0.5929	0.5786
IL1C	5	ours	0.61	0.57	0.5529	0.4571	0.4614	0.7486	0.7229	0.6186	0.4643	0.4943
IL1C	10	ours	0.6393	0.575	0.5464	0.4243	0.4621	0.7543	0.6979	0.6421	0.4736	0.52

Peaks ($\mathscr{A}_{\text{PUP}}$) [64]. To ensure statistical robustness towards the random generation of reflections using [65], we performed $Z = 50$ random repetitions of the experiments using Monte-Carlo simulation [1]. A ground truth peak is considered to be unmatched if the closest estimated number is more than a fixed number of samples aways from it (we follow [1] in setting this value equal to 20). In formulæ, we compute $\mathscr{A}_{\text{PPM}}$ and $\mathscr{A}_{\text{PUP}}$ in the following manner

$$\mathscr{A}_{\text{PPM}} = \frac{1}{Z}\sum_{i=1}^{Z}\sum_{p=1}^{\bar{P}_i} \frac{|\tau_{p,i} - \widetilde{\tau}_{p,i}|}{\bar{P}_i} \tag{4.9}$$

$$\mathscr{A}_{\text{PUP}} = \frac{1}{Z}\sum_{i=1}^{Z} \frac{K - \bar{P}_i}{K} \tag{4.10}$$

where $\bar{P}_i$ is the number of ground truth peaks for which a matching has been found among the estimated ones: such value is indexed over the Monte-Carlo simulations $i = 1,\ldots,Z$. For every i and given an arbitrary $p = 1,\ldots,\bar{P}_i$ $\tau_{p,i}$, in eq. (4.9), $\tau_{p,i}$ and $\widetilde{\tau}_{p,i}$ are the p-th ground truth peak location and its corresponding estimate, respectively. In eq. (4.10), K denotes is the number of ground truth peaks of the source signal.

By means of such metrics, we can decouple the effect of the outliers (quantified by $\mathscr{A}_{\text{PUP}}$) from the overall peak position accuracy (expressed by $\mathscr{A}_{\text{PPM}}$), ultimately better evaluating on the robustness with which TDOAs are estimated.

4.4 The proposed Test-Case: a Discussion

Performance differences across variable s values. An increasing value for s will make the acquired signal noisier, so that, in Tables 4.1 and 4.2 the case $s = 0.01$ is (much) easier with respect to $s = 1$. This visually translates into errors (and histogram bars) which increase when moving from left to right in the referred error tables. A sharp increase of errors is registerd on *white noise* (synthetic) and ***bag plastic ruslte***, ***adult male voice*** and ***dog barking*** (natural). Differently, on either pink noise (synthetic) or stapler, hand-clapping (natural), we can see that already the case $s = 0.01$ is challenging per se. We posit that a reason for that is the highly oscillatory natura of those sounds that, if compared to other cases, make them less influeced by the additive Gaussian noise (since they behave as if they were intrinsically noisy)

Differences between synthetic and natural sound-emitting sources. Let us comment on whether the usage of a synthetic versus a natural source emitting sound can have an impact on the final performance. According to the experimental results reported in Tables 4.1 and 4.2, while also inspecting the signed absolute/relative improvements of Table 4.3, we can get that there seems not to be a sharp difference in performance between these two categories. In fact, we did not register any drop/raise when swapping from white/pink noise to the other sounds considered in this work. We deem this a valuable property of the cross-correlation identity (CCI) which can naturally accommodate for a variety of applicative scenarios where the audio source is unknown.

Does adding microphones improves upon performance? We are intended in enriching this discussion with a detailed analysis on the ultimate question that our work is trying to respond. We believe that the findings of Tables 4.1 and 4.2 are plain: the honest answer to the aforementioned question we are intended to respond is neither positive nor negative, *in general*. In fact, there is a quite number of cases in which the addition of microphones is not clearly beneficial, on the contrary damaging performance: for the sake of brevity, let us report the worst cases for the two metrics. That is, the cases $s = 1$, $N = 10$ and ***hand-clapping*** for $\mathscr{A}_{\text{PPM}}$ (-1.4031 absolute improvement) and $s = 1$, $N = 4$ and ***adult male voice*** for $\mathscr{A}_{\text{PUP}}$ (-0.0393 absolute improvement). These are definitely failure cases and, specifically, ***hand-clapping***, $s = 1$ for $\mathscr{A}_{\text{PPM}}$ is clearly not positive since the trend is that performance drops while N increases. Albeit these cases are surely negative, let us observe that there are actually no cases where *concurrently* the two metrics deteriorate. In fact, in the worst cases, only one of the two is damaged: we either loose in effectiveness on how we handle outliers or in how accurately we retrieve the peaks. But, globally the case $N > 2$ is never inferior to the baseline $N = 2$ with respect to both metrics concurrently.

At the same time, let us observe that these failure cases are limited since, in the majority of the (remaining) cases, the performance is either stable (therefore adding microphones is not detrimental) or better (and thus addding microphones actually help). The fact that performance is stable when varying the number of microphones is true for the (less noisy) cases $s = 0.01$, *pink noise*, for $\mathscr{A}_{\text{PPM}}$; $s = 0.01$, ***adult male voice***, for both $\mathscr{A}_{\text{PPM}}$ and $\mathscr{A}_{\text{PUP}}$; $s = 0.01$, *pink noise*, for $\mathscr{A}_{\text{PUP}}$; $s = 0.01$ ***dog barking***, for $\mathscr{A}_{\text{PPM}}$.

Finally, let us concentrate on the ideal cases, where the performance improves when N raises. This happens for (the more challenging) cases such as $s = 1$ ***dog barking***, for $\mathscr{A}_{\text{PPM}}$; $s = 0.5$ ***adult male voice***, for $\mathscr{A}_{\text{PPM}}$; $s = 0.1$, ***stapler***, for $\mathscr{A}_{\text{PPM}}$ and $s = 0.5$, ***adult male voice***, for $\mathscr{A}_{\text{PUP}}$, $s = 0.1$ and $s = 0.2$, ***plastic rustle no. 2 (bottle)*** for $\mathscr{A}_{\text{PUP}}$; $s = 0.2$, ***dog barking*** for $\mathscr{A}_{\text{PUP}}$.

Given the alternate nature of the results, when switching from one error metric to another and while varying different s and N values, we deem necessary to summarize the highlights of our findings in the next part of our discussion.

A summary of the improvements. In Table 4.3 (bottom), we report the average absolute signed improvement δ_{avg} over the two error metrics $\mathscr{A}_{\text{PPM}}$ and $\mathscr{A}_{\text{PUP}}$: the overall majority of the cases show a superiority of the case $N > 2$ with respect to the baseline case $N = 2$ of IL1C [1]. This is exemplified from the fact that the signed improvement is positive ($\delta_{\text{avg}} > 0$) for 5 out of 8 different audio signals, in terms of $\mathscr{A}_{\text{PPM}}$, and 7 times out of 8, in terms of $\mathscr{A}_{\text{PUP}}$. Despite of their sign, the absolute value of such improvements is controlled (it never exceeds 0.5). This trend is explained from the fact that, there are high fluctuations, sometimes, between different configurations inside the case $N > 2$ for an unknown audio source.

To better investigate this trend, we also consider the signed relative improvements Δ^O of the error metrics $\mathscr{A}_{\text{PPM}}$ and $\mathscr{A}_{\text{PUP}}$ (Table 4.3, top). In this case, we allow for an oracle selection of the best number N of the microphone configuration so that we can understand what is the "upper" bound on the improvement that we can expect to register. The results are extremely encouraging: we *always* have significant positive improvements. In the worst cases (***plastic rustle no. 2 (bottle)***, $\mathscr{A}_{\text{PPM}}$), we get a +2.8% while, in the most favorable case (***adult male voice***, $\mathscr{A}_{\text{PPM}}$), the relative improvement sharply raises, reaching +28.4%.

4.5 Future Perspectives

In shed of the results of our test-case (Table 4.3), we deem now reasonable for practitioners to start investigating the regime $N > 2$ (unknown source) with computational methods which take advantage of this scenario in explicit terms. Although this actual effort is beyond the scope of the present submission, we are nevertheless interested in warming up the research in this direction by considering what are, to our opinion, the easiest modification that can be applied to the state-of-the-art method IL1C [1]. In the rest of the present Section, we will present two computational variants of IL1C whic are either based on an *incremental pre-codintioning* or an *ensemble mechanism*.

Incremental pre-conditioning. Given the core contribution of pre-conditioning the solution that IL1C introduced, we can think about an *incremental* preconditioning in which we gradually introduce one microphone, intertwining this operation with a fine-tuning of the AIRs. That is, we start from a pair of microphones and we optimize for it. Then, we use the solutions of IL1C for that pair to pre-condition the solution when solving for a third

Table 4.3 Signed improvements for the metrics $\mathscr{A}_{\mathrm{PPM}}$ and $\mathscr{A}_{\mathrm{PUP}}$ when comparing $N > 2$ with the baseline $N = 2$ using the state-of-the-art method [1]. *Top*: we provide the percentage relative improvements Δ^O using the oracle selection for microphone number's configuration (reported as a superscript). *Bottom*: We provide the mean absolute improvement δ_{avg} across *all* cases $N = 3,4,5,10$ with respect to the baseline $N = 2$. *Top and Bottom*: We report the aforementioned statistics for the more challenging noise-to-signal ratio $s = 1$.

		$\Delta^O\,(\mathcal{A}_{\mathrm{PPM}})$	$\Delta^O\,(\mathcal{A}_{\mathrm{PUP}})$
white noise	*synt*	+16.0 $\%^{(N=3)}$	+5.9 $\%^{(N=10)}$
pink noise	*synt*	+9.6 $\%^{(N=3)}$	+11.2 $\%^{(N=5)}$
plastic rustle no. 1 (bag)	***nat***	+26.8 $\%^{(N=10)}$	+16.2 $\%^{(N=10)}$
plastic rustle no. 2 (bottle)	***nat***	+2.8 $\%^{(N=5)}$	+9.8 $\%^{(N=5)}$
adult male voice	***nat***	+28.4 $\%^{(N=5)}$	+25.2 $\%^{(N=5)}$
dog barking	***nat***	+23.4 $\%^{(N=4)}$	+20.6 $\%^{(N=10)}$
stapler	***nat***	+8.1 $\%^{(N=3)}$	+19.8 $\%^{(N=5)}$
hand-clapping	***nat***	+5.7 $\%^{(N=3)}$	+14.6 $\%^{(N=5)}$

		$\delta_{\mathrm{avg}}\,(\mathcal{A}_{\mathrm{PPM}})$	$\delta_{\mathrm{avg}}\,(\mathcal{A}_{\mathrm{PUP}})$
white noise	*synt*	-0.02	-0.01
pink noise	*synt*	-0.20	+0.04
plastic rustle no. 1 (bag)	***nat***	+0.12	+0.00
plastic rustle no. 2 (bottle)	***nat***	-0.40	+0.01
adult male voice	***nat***	+0.14	+0.02
dog barking	***nat***	+0.47	+0.04
stapler	***nat***	+0.25	+0.04
hand-clapping	***nat***	-0.35	+0.02

microphones: we the update also the AIRs for the first two microphones. The procedure iterates until the N-th microphones is added (so that the $N-1$ AIRs of the other microphones are fine-tuned, at least one time). Let us formalize the prior argument in the following pseudocode.

1. Sample two random microphones m_1, m_2.

2. Optimize eq. (4.8), using the *standard* pre-conditioning [1], thus obtaining the AIRs for m_1 m_2.

3. Add a third microphone m_3: optimize eq. (4.8) again but now changing the preconditioning. The AIRs of m_1 and m_2 will be the ones obtained at the previous stage, while the AIR of m_3 will be initialized using the standard approach [1].

4. Update the AIRs for all solved microphones.

5. Keep adding microphones, following the same procedure, until all N ones are covered

Results & Discussion. We did not register any substantial improvement using this sequential addition, to the point that even the case $N = 2$ is superior in performance. For the sake of brevity, let us report a glance of the scored results, providing a peculiar case which is aligned with the general trend which we do not report for the sake of brevity. For *white noise*, the results of incremental strategy describe above are 0.0036 ($s = 0.01$), 0.0357 ($s = 0.1$), 0.09 ($s = 0.2$), 0.1536 ($s = 0.5$) and 0.2343 ($s = 1$) for $\mathscr{A}_{\text{PUP}}$ and 0.2658 ($s = 0.01$), 0.5866 ($s = 0.1$), 1.0023 ($s = 0.2$), 1.7345 ($s = 0.2$) and 2.2391 ($s = 1$) for $\mathscr{A}_{\text{PPM}}$ – all error values refer to the case with $N = 4$, while averaging over $Z = 50$ random extraction of the sequence with which microphones are incrementally added. We explain this lack of improvement with the fact that, despite adding microphones *in a single solution* maybe beneficial, their sequential addition can be detrimental since, albeit on the one side the case $N > 2$ is providing more cues than the baseline $N = 2$, the sequential addition of microphone would lead to "over-fitting" the AIRs of some of the microphones, ultimately damaging the final performance.

Ensemble mechanism. Let us observe that the inference stage of IL1C [1] is based on peaks finding, a method which is known to suffer when spurious peaks are present. To accommodate for that, let us take advantage of the following approach. We can split the case $N > 2$ into several $N = 2$ sub-problems, by pairing microphones into couples. We therefore create a number of playgrounds with 2 microphones only (unknown source) - so that we match the operative conditions on which IL1C [1] was originally tested. We therefore create some redundancy in the estimate of the AIRs: this is because one microphone can belong to several pairings at the same time, so there will be multiple candidate solutions for the same AIRs - two candidates, referring to two different microphones, from each artificial pairing. We solve for this redundancy by averaging out all different candidates referring to the same microphone. We deem this approach to be arguably simple, perhaps rough, but still effective in handling a well known computational issue which damages peak findings algorithm. In fact, the presence of spurious (noisy) peaks surely affect the estimate of TDOAs. We attempt to mitigate this problem by exploiting the well known smoothing and regularizing properties of averaging as our ensemble mechanism.

Results & Discussion. The reader can refer to Table 4.4 for the quantitative evaluation of our ensemble strategy applied to IL1C [1] evaluated in the test-case $N = 10$. We are expecting to register a very interpretable phenomenon out of a simple strategy such as averaging multiple candidate solutions corresponding to the same AIR: we should expect to register a regularizing effect which smooths out the AIRs, removing spurious peaks due to, for instance, numerical instability. This explains the improvements achieved from our proposed ensemble mechanism versus the IL1C [1] baseline: once spurious peaks have been removed, we expect that a peak finding algorithm such that the one applied in [1] can be more effective in finalizing the estimate of TDOAs. This consistently happen in the cases $s = 0.01$, $s = 0.1$ (for both $\mathscr{A}_{\text{PPM}}$ and $\mathscr{A}_{\text{PUP}}$) and $s = 0.2$ (only for $\mathscr{A}_{\text{PUP}}$), while, when

Table 4.4 The ensemble mechanism. We the report the performance of IL1C [1] ($N = 10$, *white noise*) versus the ensemble mechanism in which couples of microphones are solved, first, and the aggregated by averaging across the redundancy of AIRs referring to the same microphones. We denote a better performance in bold, across different signal-to-noise values s.

	$\mathscr{A}_{\text{PPM}}$				
	$s = 0.01$	$s = 0.1$	$s = 0.2$	$s = 0.5$	$s = 1$
IL1C [1]	2.2250	2.0199	**2.2215**	**4.1515**	**4.1766**
Ensemble (*us*)	**1.6982**	**1.8995**	2.2643	4.4532	4.4647

	$\mathscr{A}_{\text{PUP}}$				
	$s = 0.01$	$s = 0.1$	$s = 0.2$	$s = 0.5$	$s = 1$
IL1C [1]	0.3750	0.3543	0.3971	**0.7186**	**0.7214**
Ensemble (*us*)	**0.2157**	**0.2414**	**0.2550**	0.7421	0.8250

considering the "more difficult" cases $s = 0.5$ and $s = 1$ we do not see a sharp improvement of the ensemble method. This is probably due to the fact that the candidate solutions that are averaged are, each of them, noisier. Therefore, the averaging effect produces an excessive over-regularization which excessively smoothens the peaks, damaging the performance of the peak finding. Nevertheless, the regularizing effect of averaging can be inspirational for practitioners in exploiting a large number of microphones to better estimate TDOAs.

4.6 Conclusions

In this work, we generalized the traditional experimental playground in which the notion of cross-correlation identity (CCI), applied to the estimation of TDOAs using blind channel deconvolution methods [1; 55; 60; 61; 62], switching from the case $N = 2$ to $N > 2$. Our analysis shows that, by simply allowing for a increased number of microphones, the very same state-of-the-art method ILC1 [1] can be sharply boosted in performance (see Tab. 4.3) without requiring any change in the computational pipeline.

We deem that our findings open up to a novel research trend in which CCI identities are better combined with the case $N > 2$, so that improvements in the error metrics can come from two different, yet complementary, factors: advances in the optimization standpoint and multiple CCI relationships. We warm-up the research efforts in this directions with two simple modifications of IL1C, showing that, with respect to an incremental addition of the microphones, the practitioners should preferred a late fusion ensemble mechanism - which has the understandable property of easing the peaks finding-based inference stage of [1].

Chapter 5

Dual Cam device: audio and image detection and classification

5.1 Introduction and motivation

Acoustic imaging is an imaging modality that exploits the propagation of acoustic waves in a medium to recover the spatial distribution and intensity of sound sources in a given region. Well known and widespread acoustic imaging applications are, for example, sonar and ultrasound. There are active and passive imaging devices: in the context of this thesis we will consider a passive imaging system that does not emit any sound but acquires it from the environment. In an acoustic image each pixel corresponds to the sound intensity of the source, the whose position is described by a particular pair of angles and, in the case in which the beamformer can, as in our case, work in *near-field*, from a distance on which the system is focused. There are several advantages that the use of an array can bring compared to the use of a single sensor:

- the greater physical dimensions that an array of sensors has compared to a single transducer allow, with the same wavelength λ considered, a greater capacity for spatial discrimination of the directions of origin of the signals, and therefore a greater resolution, referred to in this context as angular resolution. Usually the dimensions of an array are quantified by evaluating its spatial opening D defined as the maximum distance that separates two elements belonging to it. Therefore the spatial discrimination capacity of an array coincides with a value proportional to D/λ.

- an array has properties of flexibility unattainable by the single sensor with the same implementation simplicity. In fact, in many applications it may be necessary to modify

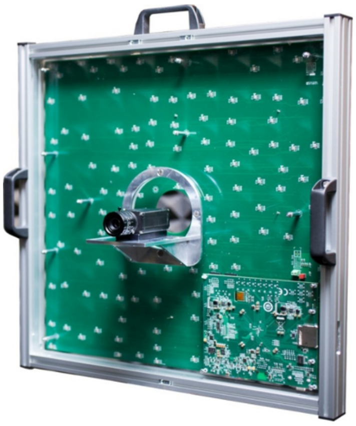


Figure 5.1 Acoustic images are obtained from a geometrically structured array in which microphones are embedded in a single compact device called *Dual Cam*. Dual Cam is a prototype made of 128 microphones, 50 cm x 50 cm, working in real time over a wideband (500 Hz–6400 Hz). Microphone layout and processing parameters are optimized according to a patented method. Frame rate: 12 frames per second. Maximum field of view: 90° elevation, 360° azimuth (tunable according to the video camera field of view). State of art beamforming algorithm: robust, superdirective filter-and-sum beamforming.

> the spatial filtering function in real time to maintain an effective attenuation of the interfering signals to the advantage of the desired ones. This becomes essential in imaging applications in which the pointing direction changes constantly in order to scan all possible directions of arrival of the signal. This change, in a system that adopts an array of transducers, is achieved simply by varying the way in which the beamforming combines the data coming from each sensor in a linear fashion; in the case of a single transducer, the change is impractical as it would be necessary to act directly on the physical characteristics of the sensor.

In the case of isotropic noise, i.e.noise (intended as an interfering signal) that comes uniformly from all directions, the gain given by the array in terms of Signal to Noise Ratio with respect to the single sensor is precisely given by the directivity formula. Now the more directive is an array, the more it has a narrow main lobe (and secondary low lobes) and therefore at a qualitative level it is more able to discriminate signals that come from different directions. Given a set of acoustic sources arranged in a 3D environment, an acoustic image is given by a 2D map in which each point or pixel encodes the intensity of the acoustic signal coming from a certain direction. Beamforming can be used to generate an *acoustic image* since, once a pointing direction is set, the energy of the resulting beam signal will be representative of the signal coming from the pointing direction itself. Consequently, by performing beamforming for a set of pointing directions, corresponding to a grid that covers a certain solid angle, it is possible to obtain the sound intensity for each direction and make it correspond to a pixel in the final acoustic image. For each chosen distal direction, the beam signal is obtained by carrying out the filter-and-sum beamforming on a predefined time window of the 128 acquired signals, setting the delays according to the chosen steering direction. Subsequently the signal is filtered with a band pass filter in order to select the band of interest. Finally, the energy of the filtered signal is calculated, and the value obtained is subjected to a thresholding operation, normalized and converted into heatmap values to be displayed as pixels. These operations are repeated for each pointing direction selected on the basis of a predefined value grid. An acoustic image encodes, for each pixel, the sound intensity coming from a given direction, enabling localization of multiple sound sources in the same temporal frame. An acoustic image can be geometrically overlapped with a standard video image from a camera, allowing easy identification of targets.

5.2 Materials and Methods

In this chapter, we propose the use of a new modality characterized by a richer information content, namely acoustic images, for the sake of audio-visual scene understanding. Each pixel in such images is characterized by a spectral signature, associated to a specific direction in space and obtained by processing the audio signals coming from an array of microphones. By coupling such array with a video camera, we obtain spatio-temporal alignment of acoustic images and video frames. This constitutes a powerful source of self-supervision, which can be exploited in the learning pipeline we are proposing, without resorting to expensive data annotations. However, since 2D planar arrays are cumbersome and not as widespread as ordinary microphones, we propose that the richer information content of acoustic images can be distilled, through a self-supervised learning scheme, into more powerful audio and visual feature representations. The learnt feature representations can then be employed for downstream tasks such as classification and cross-modal retrieval, without the need of a microphone array. To prove that, we introduce a novel multimodal dataset consisting in RGB videos, raw audio signals and acoustic images, aligned in space and synchronized in time. Experimental results demonstrate the validity of our hypothesis and the effectiveness of the proposed pipeline, also when tested for tasks and datasets different from those used for training.

Humans perceive and interpret the world by combining different sensory modalities. However, designing computational systems able to emulate or surpass human capabilities in

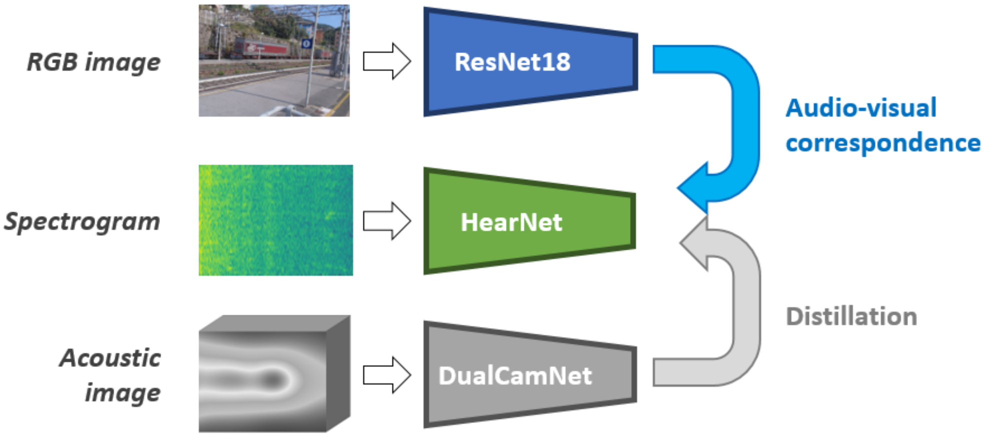


Figure 5.2 We consider three modalities aligned in time and space: RGB, (monaural) audio signal (here in the form of spectrogram), and acoustic images. We exploit such correspondence to jointly learn audio-visual representations. We improve audio models with knowledge transfer from the acoustic image model.

this respect, although of utmost importance from both scientific and applicative standpoints, is still a far-reaching goal.

Among all modalities, vision and audio are surely the most commonly used and important ones that both humans and machines can use to sense the world. This is also caused by the fact that they are often quite correlated, temporally synchronized, and support each other for interpretation tasks. In fact, sound helps to pay attention and visually focus on situations of interest, and may complement noisy or low-quality visual information, ultimately aiming to improve the interpretation of a scene. In such cases, humans take advantage of the spatial localization of the produced sound (obtained thanks to the binaural configuration of our auditory system), to shift visual attention to the event that generated the sound.

Unfortunately, artificial systems mimicking human performance are not so common, especially because video data typically comes with a monaural (single microphone) acoustic signal only. Hence, spatial localization is lost, and reliably recovering it is a difficult and only partially solved problem [68; 69]. Thus, in order to have the possibility to emulate human performance by also exploiting spatially localized audio data, one needs to resort to an array of microphones positioned in special geometrical (e.g., planar) configuration, and able to provide an enriched audio description of a scene – an acoustic image – being formed by properly combining the signals acquired by all microphones. In an acoustic image, each pixel is characterized by the spectral signature corresponding to the audio signal coming from the corresponding direction, so, overall, allowing effectively to visualize the acoustic landscape of the sensed scene (see Fig. 5.3).

In particular, we take advantage of an audio-visual sensor composed by a microphone array coupled with a video camera, jointly calibrated, in order to get a sequence of acoustic images and associated video frames, aligned in space and time [70; 71]. Examples of sample video frames overlaid with the energy map of the sound obtained from the corresponding acoustic images are shown in Fig. 5.3. The peculiar nature of this data, i.e., the spatial alignment and time synchronization of the data produced by such sensor, opens the door to the adoption of *self-supervised learning* approaches for model training. The motivation for this choice lies in the fact that such methods do not require data annotations. This specifically suits to our case, since acoustic images would results quite expensive to fully annotate (i.e., assigning pixel-level or bounding box annotations to the same objects in both video frames and acoustic images while listening the signals coming from different directions). Instead, self-supervised methods just exploit the implicit supervision inherent in the signals themselves. For example, we can train audio-visual networks by simply looking and listening to a large number of unlabelled videos, and exploiting their natural alignment as a supervisory signal. More in detail, in deep self-supervised learning schemes, a network is trained to solve a so-called *pretext task*, and the quality of learned features is then assessed on a variety of *downstream tasks*, which are usually supervised (e.g., classification), showing beneficial effects [72].

More specifically, in this chapter we investigate whether we can obtain more powerful features for downstream tasks by training audio-visual models with a self-supervised framework exploiting audio-visual correspondence. We also employ acoustic image modality as privileged information [73] used at training time in a knowledge distillation [74] framework (see Fig. 5.2) to enhance such audio-visual self-supervised features. The distillation framework was already exploited in the literature for classification tasks in several scenarios [74; 75; 76; 77; 78], but always in *supervised* settings. Instead, here we are proposing a novel *self-supervised distillation* framework, which does not require any time-consuming annotations, and allows to train audio and video models together. To the best of our knowledge, privileged information was never exploited before in a self-supervised learning pipeline. After training, individual models can be used as feature extractors for the sake of audio and video classification and cross-modal retrieval as downstream tasks.

To show the potentiality of acoustic images to improve feature learning, we collected a new multimodal audio-visual dataset, composed by RGB video frames, acoustic images and monaural audio signals[1]. This dataset contains 10 classes of real sound acquired outdoors in the wild, is bigger than AVIA dataset [78] and more suitable for self-supervised learning. With this novel dataset we carry out an accurate ablation study; subsequently, in two different benchmark datasets publicly available [78; 79], we show that, when augmented with privileged information distilled from acoustic images, the obtained feature representations are more powerful than in the case of just training audio and visual models with the audio-visual correspondence task. In the end, acoustic images proved to have notable characteristics to be effectively transferred to other domains and tasks, when distilled by our training mechanism.

In summary, the main contributions of this work can be summarized as follows:

- We propose a multimodal deep learning framework to learn audio-visual models considering a novel modality, acoustic images, which is heavily under-explored in computer vision. This framework embeds a novel self-supervised distillation mechanism to transfer the information extracted from an acoustic image model to an audio model for learning more powerful feature representations.
- We collect and release a new multimodal dataset of aligned audio (single microphone), RGB images and acoustic images, bigger than [78].
- Using this dataset for model training, we show the effectiveness of our framework for downstream tasks such as 1) audio and video classification, and 2) cross-modal retrieval. In particular, we prove that the features obtained by the distillation of acoustic images perform better than those obtained without using such privileged information, not only on our dataset, but also on other public benchmarks [78; 79].

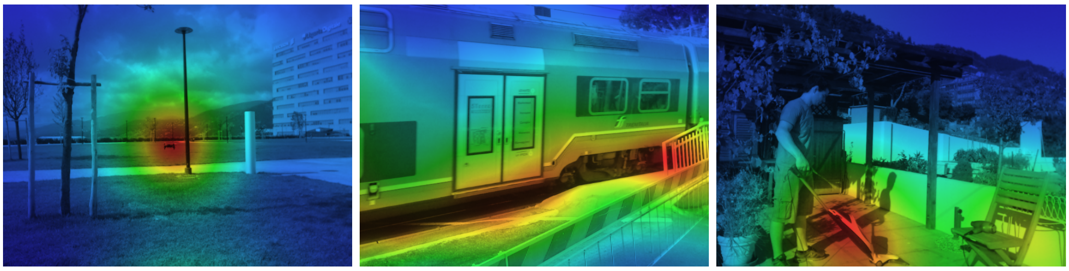


Figure 5.3 Three examples from the collected dataset. We visualize the acoustic image by summing the energy of all frequencies for each acoustic pixel. The resulting map is overlaid on the corresponding RGB frame. From left to right: drone, train, and vacuum cleaner classes.

The rest of the chapter is organized as follows. We review the state of the art and highlight the main differences with respect to our work in Section 5.3. Section 5.4 introduces

[1] This dataset will be publicly released upon acceptance.

our new dataset and briefly presents the sensor used. Section 5.5 explains the proposed self-supervised training method and, in Section 5.10, we evaluate our learning strategy and report the performance of the experiments in the downstream tasks. Finally, in Section 5.11, conclusions are drawn.

5.3 Related works

Our work lies at the intersection of two broad topics, namely self-supervised learning and knowledge distillation. In this section we give an overview of relevant works in both fields, mainly in the context of audio-visual learning, and discuss how our method relates to them. We also review literature dealing with acoustic images.

Audio-visual self-supervised learning. Multimodal learning takes advantage of data from different modalities [80] aiming at obtaining better semantic representations than those learned by segregated modalities.

There has been increased interest in using perception-inspired audio-visual (fusion) deep learning models because the correspondence between the visual and the audio streams is ubiquitous and free in unlabeled consumer videos.

Vision and sound are often informative about the same concept of the world. As a consequence of their correlations, concurrent visual and sound data provide a rich supervisory self-training signal that can be used to jointly learn useful audio and video representations. Early approaches trained single networks on one modality only, using the other one to derive a sort of supervisory signal [81; 82; 83; 84]. For example, [81; 82] train an audio network using pre-trained visual architectures as teachers. Instead, [83; 84] directly predicts sound from video, thus using ambient sound as a supervisory signal for video.

Other works [85; 86; 87; 88; 89; 90; 91] jointly train visual and audio streams, aiming at learning multimodal representations useful for many applications, such as classification, cross-modal retrieval, sound source localization, speech separation, and on/off-screen audio-visual source separation. As in [86], we also use audio-visual correspondence verification task: networks are trained to determine whether a video frame and a short audio chunk overlap in time. Learned representations are then tested in a classification task. Within a similar framework, [85] uses hard samples, i.e., slightly out-of-sync audio and visual segments sampled from the same video in a self-supervised curriculum-based learning scheme. [87] enforces the alignment of features extracted by audio and visual networks by computing the correspondence score as a function of the Euclidean distance between the normalized visual and audio embeddings, hence making them amenable to retrieval.

The common factor in all these works is the natural *temporal* synchronization between (single) auditory signal and visual images, which is used to train the several models in a self-supervised manner. In our case, the intrinsic *temporal synchronization*, but also the *spatial alignment* of visual and acoustic images can be exploited as a supervisory signal. Our method takes inspiration from [87] and [89]: we force audio-visual agreement between feature maps to find aligned shared representations, however, both the task and the mechanism we propose for training are different, since they involve knowledge distillation and an extremely different modality.

Knowledge distillation. Our work is related to knowledge distillation, which can be coarsely and generally defined as the class of approaches trying to indeed 'condensate' knowledge gained in a learning task and feed another learning task or another model [75]. Such framework was later unified with the privileged information framework [73] into the so-called generalized distillation theory [74], and recently exploited in the context of multi-modal learning with missing modalities at test time [76; 77]

In our audio-visual learning scenario, a former knowledge distillation mechanism can be considered as that of [81], which capitalizes on the natural synchronization between vision and sound to train a sound classification model using a teacher-student setup. Knowledge is transferred from vision (ImageNet and Places pre-trained networks) into sound by means of the Kullback-Leibler divergence, using unlabeled videos as a bridge between the 2 modalities. However, the video teachers are models pre-trained with large labelled datasets.

Our approach is quite different. In our case, the privileged information is represented by the acoustic images which, as discussed, are more problematic to acquire, and can thus be missing in a real-world testing scenario. They are thus leveraged only at training time and used to build audio models that, at test time, outperform those learnt without this additional information. Besides, while traditional generalized distillation framework are applied in a *supervised* setup, since exploiting cross-entropy loss and teacher's soft predictions [74], we are here in the self-supervised scenario, where labels are missing. We can thus only leverage embeddings as additional information from the teacher. Furthermore, the teacher network itself is also trained with self-supervision.

Acoustic images. Acoustic images are obtained using an array of microphones, typically distributed in a planar configuration, by properly combining the audio signals acquired by every microphone using an algorithm called beamforming [92]. To the best of our knowledge, acoustic image processing with deep learning methods was only preliminary explored in [78], which proposed an architecture able to classify acoustic images in a multimodal action dataset in a supervised way. Furthermore, it also showed how to distill acoustic image information to audio models, still in a *supervised* way. The substantial difference of our work with respect to [78] is that we use here a self-supervised learning approach, for which, as also above highlighted, the canonical supervised distillation [74] cannot be applied. Other applications of acoustic images regarded only the tracking of sound sources [70; 71]. In the end, no other works are present in the literature aimed at using such unique source of information in a *self-supervised learning* setting.

5.4 ACIVW: ACoustic Images and Videos in the Wild

We acquired a multimodal dataset containing 5 hours of videos outdoors in the wild, using an acoustic-optical camera. The sensor captures both raw audio signals from 128 microphones acquired with a sampling frequency of 12.8 kHz and RGB video frames of 480×640 pixels, using a planar array of microphones located according to an optimized aperiodic layout [93] and a webcam placed at the device center. Audio data is acquired in the useful bandwidth 500 Hz – 6.4 kHz and audio-video sequences are acquired at a frame rate of 12 frames per second (fps).

$36 \times 48 \times 512$ multispectral acoustic images are obtained from the raw audio signals of all the microphones combining them through the beamforming algorithm [92], which summarizes the audio intensity for every direction and discretized frequency bin. The acquisition of the latter modality is aligned not only in time with optical images, but also in space: each acoustic pixel corresponds to 13.3×13.3 visual pixels. Among the raw audio waveforms, we choose one microphone for training monaural audio networks.

We selected 10 classes of interest: drone, shopping cart, traffic, train, boat, fountain, drill, razor, hair dryer, vacuum cleaner. Figure 5.3 shows three sample RGB frames overlaid with the energy of the corresponding acoustic image. More examples and videos are provided in the supplementary material. We acquired data for half an hour for each class, in different locations and viewpoints. This implies more than 21,000 RGB and acoustic images. The data is split in training, validation and test in the proportion 70%, 15% and 15%.

We use the training split of this dataset for the pretext task of learning correspondences. We then test for the downstream tasks of cross-modal retrieval and classification on the test set. For classification, we also test on two publicly available datasets proposed in [78] and [79].

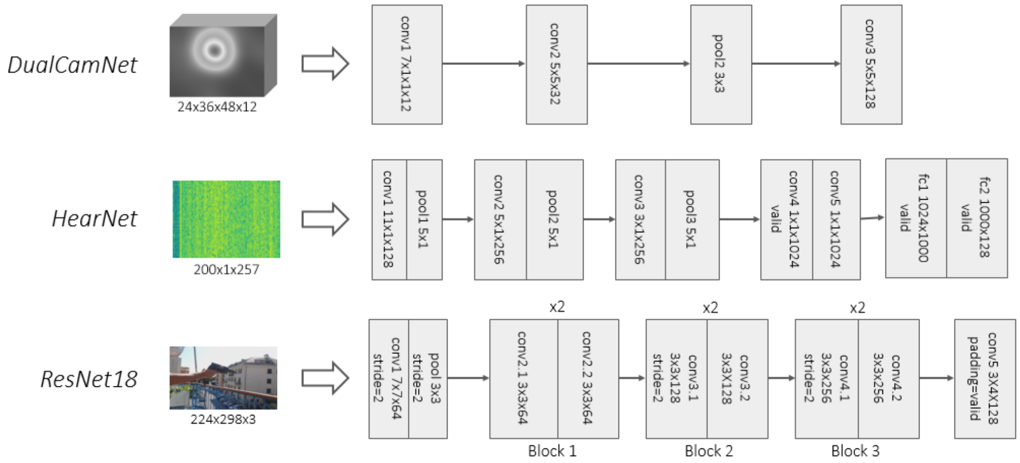


Figure 5.4 The adopted models for the 3 data modalities. In convolutional layers stride=1 and padding=SAME unless otherwise specified.

5.5 The method

As mentioned above, we consider three data modalities, namely audio, acoustic images and RGB images, and we adopt a different stream network for each modality as they are extremely heterogeneous, as shown in Fig. 5.4.

Our aim is to train two models at a time using audio-visual correspondence pretext task: first, we train the acoustic images' stream jointly with the RGB stream, and, second, the audio stream with the RGB stream. After that, we exploit the trained acoustic image stream to distill additional knowledge to a new audio stream, trained again using the same pretext task as illustrated in Figure 5.5. We then compare the performances of audio and video models trained with and without the aid of the self-supervised pre-trained acoustic image stream. In Section 5.6, we provide a description, some statistics and examples about the ACIVW dataset. Section 5.7 presents an analysis of the classification performance using different modalities. In Section 5.8, we report implementation details, in particular regarding the setting of the several hyperparameters used. Finally, in Section 5.9, we show some cross-modal retrieval examples.

5.6 ACIVW Dataset

As described in the main chapter, we used an array of 128 microphones with a webcam in the device center to collect a big dataset of three modalities aligned in time and space: RGB frames, audio and acoustic images. Planar arrays are very sensible to echos that are usually

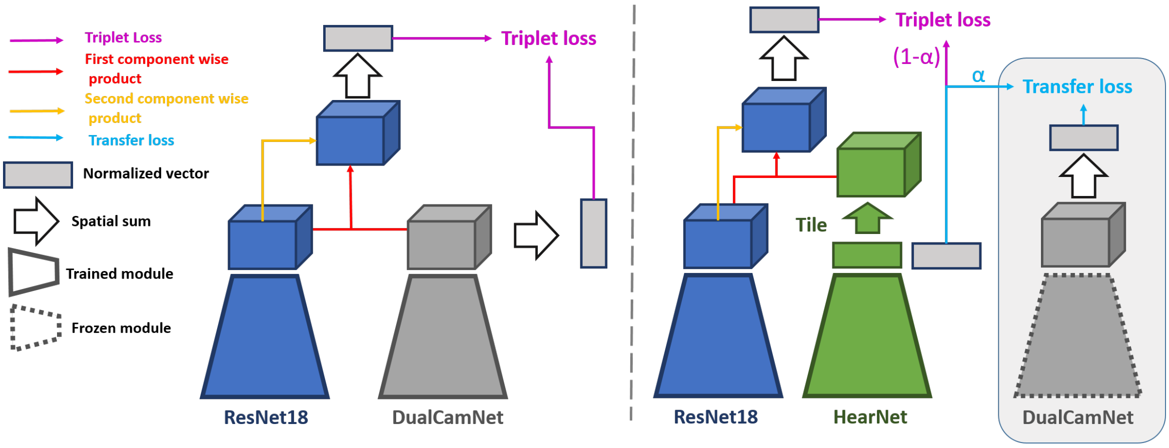


Figure 5.5 Proposed distillation method. Left: self-supervised learning of the teacher DualCamNet. Right: The pre-trained teacher network contributes to the self-supervised learning of the monaural and video networks. Note that setting $\alpha = 0$ the audio network is trained without distillation.

present in indoor environments, so we collected all the dataset outdoors to record good quality sounds exploiting the planar array features. Statistics about dataset are in Table 5.1. We show examples of the three modalities for each class in Figures 5.6, 5.7: on the left RGB image, in the center energy of the corresponding acoustic image overlaid on RGB frame and on the right the spectrogram obtained from one single microphone.

Number of classes	10
Number of videos per class	60/26.80/10
Total number of videos	268
Length of videos in seconds	256/68.95 /2
Length of class in seconds	1898/1847.8/1794
Total length of videos in seconds	18478

Table 5.1 ACIVW dataset statistics. Where there are three entries in a field, numbers refer to the maximum/average/minimum.

We accompany this pdf with some videos with acoustic image energy overlaid on video frames. Each video comes in two versions:

1. With the audio from a single omnidirectional microphone, which is fixed for all videos
2. With the audio coming from the the direction of the sound source. This is obtained by isolating the virtual directional microphone corresponding to the acoustic pixel where the source is located.

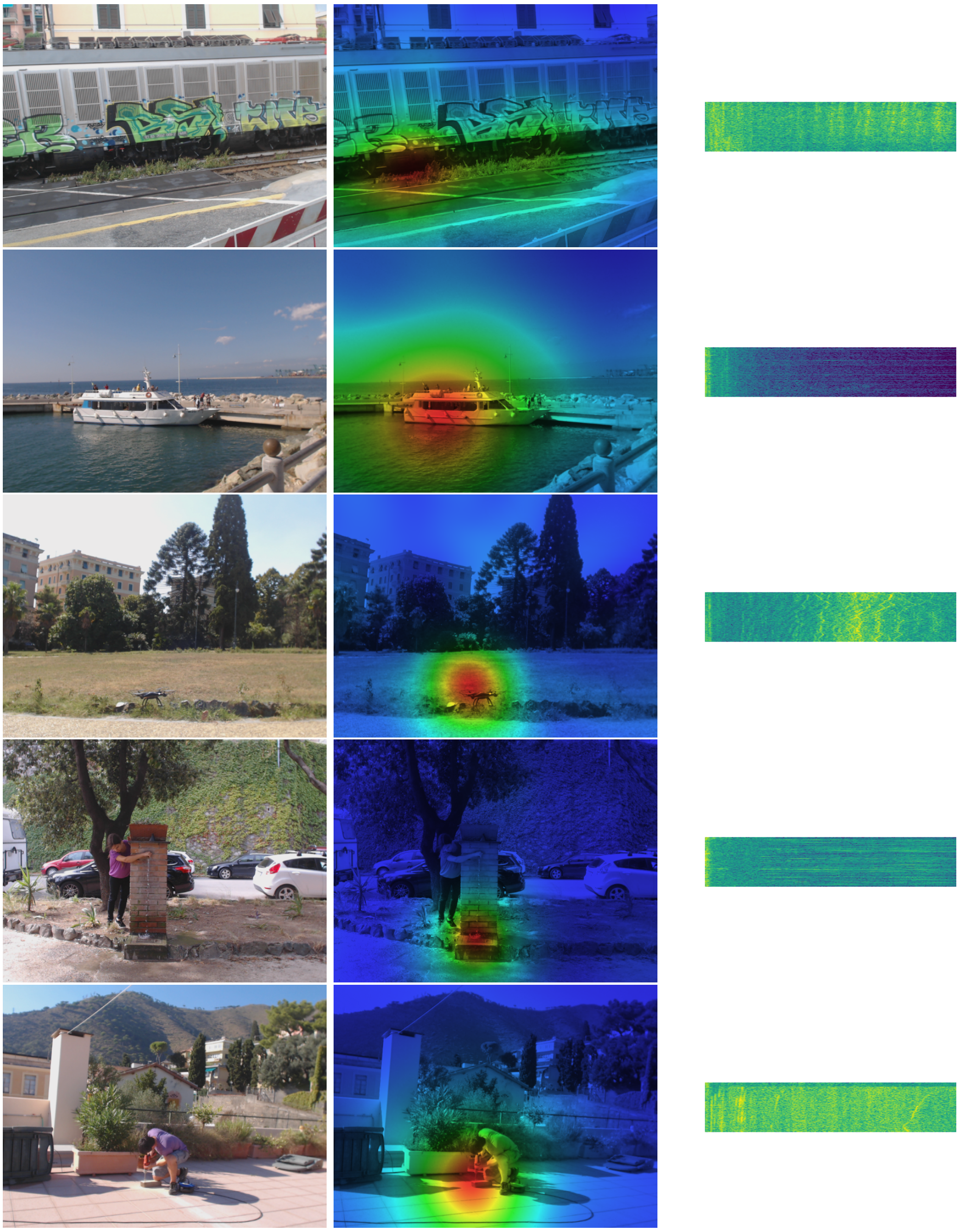


Figure 5.6 Examples of ACIVW Dataset from the classes: from top to bottom *train*, *boat*, *drone*, *fountain*, *drill*. Left: RGB frame, center: acoustic energy map overlaid on the acoustic frame, right: single microphone spectrogram.

The latter is obtained through Inverse Fast Fourier Transform (IFFT) of the FFT of the acoustic pixel obtained with the beamforming algorithm. You can notice the difference between omnidirectional and directional sound.

Both the dataset and the code will be released after the conclusion of the review process.