Circuits and Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ens.ewi.tudelft.nl/

CAS-2019-4368967

M.Sc. Thesis

Inferring the location of reflecting surfaces from acoustic measurements

Vincenzo Zaccà B.Sc.

Abstract

The response of a sound system in a room primarily varies with the room itself, the position of the loudspeakers and the listening position. The room boundaries cause reflections of the sound that can lead to undesired effects such as echoes, resonances or reverberation. Knowledge of the location of these large reflecting surfaces can help to better estimate the sound field behavior inside the room. This work focuses on exploiting the inherent information present in echoes measured by microphones to infer the location of nearby reflecting surfaces. The investigated application uses a loudspeaker to emit a known signal and record the resulting sound field with a co-located microphone array. A signal model is proposed which provides a relationship between reflector locations and measured microphone signals. The locations of reflections are estimated by fitting a sparse set of modeled reflections with measurements. We present two novelties with respect to prior art. First, the method is end-to-end where from raw microphones measurements it outputs an estimate of the location of reflectors. For the case of a compact uniform circular microphone array, the symmetry can be exploited to reduce the computational complexity of the inference process. Secondly, the model is extended to include a loudspeaker model that is aware of the inherent directivity pattern of the loudspeaker. The performance of the proposed localization method is compared in simulation to the existing state-of-the-art localization methods. An experimental study with real world measurements was also conducted to investigate the performance of the model.

Inferring the location of reflecting surfaces from acoustic measurements

using a compact microphone array collocated with a loudspeaker

---

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Vincenzo Zaccà B.Sc.
born in Delft, Netherlands

This work was performed in:

Circuits and Systems Group
Department of Signals & Systems
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Delft University of Technology

Delft University of Technology
Department of
Signals & Systems

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Inferring the location of reflecting surfaces from acoustic measurements"** by **Vincenzo Zaccà B.Sc.** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 28-10-2019

Chairman: _____
prof.dr.ir. Richard Heusdens

Advisor: _____
prof.dr.ir. Odette Scharenborg

Committee Members: _____
dr. Jorge Martinez

_____
dr. Pablo Martinez-Nuevo

Abstract

The response of a sound system in a room primarily varies with the room itself, the position of the loudspeakers and the listening position. The room boundaries cause reflections of the sound that can lead to undesired effects such as echoes, resonances or reverberation. Knowledge of the location of these large reflecting surfaces can help to better estimate the sound field behavior inside the room. This work focuses on exploiting the inherent information present in echoes measured by microphones to infer the location of nearby reflecting surfaces. The investigated application uses a loudspeaker to emit a known signal and record the resulting sound field with a co-located microphone array. A signal model is proposed which provides a relationship between reflector locations and measured microphone signals. The locations of reflections are estimated by fitting a sparse set of modeled reflections with measurements. We present two novelties with respect to prior art. First, the method is end-to-end where from raw microphones measurements it outputs an estimate of the location of reflectors. For the case of a compact uniform circular microphone array, the symmetry can be exploited to reduce the computational complexity of the inference process. Secondly, the model is extended to include a loudspeaker model that is aware of the inherent directivity pattern of the loudspeaker. The performance of the proposed localization method is compared in simulation to the existing state-of-the-art localization methods. An experimental study with real world measurements was also conducted to investigate the performance of the model.

Contents

List of Figures

List of Tables

1 Introduction

Experiencing sound can be a powerful tool to evoke emotions in the listener. We often regulate our mood by listening to our music of choice. With the vast number of portable loudspeaker systems on the market, the way one can experience sound is evolving. For example, by placing several loudspeakers around a room, so-called (personal) sound zones can be created [2]. The challenge in creating the sound zones is to minimize the interference between two zones that are close together [3]. One thing to consider when controlling the sound zones is that walls tend to echo back sound emitted into them [4]. Thus, setting up acoustic zones within a room requires awareness of the positions of the room walls and their acoustic properties. Similarly, in teleconference one wishes to improve the intelligibility of speech by canceling out the echoes introduced by the room. Another application that has recently gained momentum is that in recent years, formats with object-based audio tracks are standardized [5]. In surround sound systems, the group of loudspeakers can arrange the individual audio tracks that corresponds to instruments in space to create a rich sound stage. Essentially moving the sound mixing away from the studio and into the living room. These methods that compensate or exploit room information are limited by the availability and reliability of the information that can be acquired about the room [6]. Obviously, one could ask the user to provide this information and to configure the device. However, as the complexity of these multiple loudspeaker and multi-microphone systems increase, the burden to optimally configure the device should not be placed at the user.

So how can a smart loudspeaker system automatically infer the room size and shape and its relative positioning? Recently, intelligent loudspeakers that are equipped with inexpensive microphones have entered the consumer market. These systems usually consist of an enclosure with multiple loudspeaker drivers and a microphone array. With the microphone array, the system is capable of recognizing speech commands from users and has created a new interface between human and machine. It has been proposed to use such microphones to also determine the location of the room walls.

The general principle is that when a loudspeaker emits a sound signal into the room, it will be reflected (echoed) by the walls. The microphones will receive these echoes in the form of delayed versions of the transmitted signal (filtered by the loudspeaker, walls, and microphones). The direct path contribution (i.e. the emitted signal received directly by the microphones, without reflection from the walls) is typically know, since the relative position between loudspeaker and microphones is known and constant over time, and can therefore be eliminated. Distances to the walls can then be determined by estimating the precise delay of the echoes from the walls, and by using the relative delay between microphones in the array to determine angles. In practice, such echo detection is rather challenging, as the echoes of the transmitted signal are concealed by the filtering of the loudspeaker, walls, and microphones.

In this thesis a novel measurement model is proposed that provides a relationship

between microphone measurements and reflector locations. The signal model includes the convolution with the excitation signal, a loudspeaker model and the microphone array geometry. The influence of the room is modeled as a spatially varying linear time invariant system that assumes specular reflections. The forward model treats the location of the reflectors as input, and defines the microphone signals as the result of the system (acoustic excitation signal, loudspeaker, microphone array) acting on this input. The locations of the reflectors are identified by solving the inverse problem.

This work in this thesis has two novelties with respect to prior art. First, the method is end-to-end, where from raw microphones measurements it outputs an estimate of the location of nearby reflections. In particular, for the compact uniform circular microphone array, the symmetry is exploited to create an algorithm that is of reduced computational complexity. Secondly, it uses a loudspeaker model that is aware of the inherent directivity of a loudspeaker. It is thus assumed that the loudspeaker under test has been measured in free field conditions from various angles to infer an appropriate model. The directivity aware model is then used to infer the locations of the echoes in a more robust way.

1.1 Research statement and outline

In this thesis, the following general research question is addressed:
How can the location of reflecting surfaces in a room be estimated using a loudspeaker system with co-located built-in compact microphone array? The rest of the thesis continues as follows. Chapter 2 provides an overview of the prior art, presents the contributions of this thesis and provides the necessary background information. In particular at the end of Chapter 2 a general signal model is presented. In Chapter 3 the proposed design is presented. The signal model for the compact uniform circular array is presented that provides a relationship between reflector location and microphone observations. At the end of Chapter 3, an efficient method is presented to locate the reflectors by fitting a sparse set of modeled reflections with measurements. Chapter 4 presents simulation results that evaluate the performance of the proposed methods with state-of-the-art methods and investigates the performance of the model with real world measurements. Finally, Chapter 5 presents the conclusions and future research directions.

1.2 Notation

Throughout this thesis vectors are denoted by lowercase bold letters, i.e.$\mathbf{y}$ and matrices with capital bold letters, i.e. $\mathbf{M}$. Scalars are denoted by lowercase letters as c and integers representing dimensions are denoted by capital letters such as T and L.

(Multivariate) signals are assumed discrete, unless otherwise specified. As an example $x(n)$ is a one dimensional discrete signal of length L. Zero based indexing is used, therefor $x(n)$ is defined for $n = 0, 1, \ldots, L-1$, sampled at f_s. The $*$ denotes the convolution operator and $\otimes$ denotes the Kronecker product. Finally, for some matrices whose structure is important, the names are predefined. In particular we have that $\mathbf{I}_M$

is the identity matrix of size $M \times M$, $\mathbf{F}_M$ denotes the unitary Discrete Fourier Matrix of size $M \times M$. $\mathbf{\Lambda}$ is reserved for diagonal matrices and $\mathbf{T}$, $\mathbf{C}$ are reserved for Toeplitz and Circulant structured matrices respectively.

Problem Background and Related Work

2

A smart loudspeaker consists of at least one loudspeaker drive and a microphone in the same enclosure. The problem is to sense the environment in which the system is placed. More specifically, in domestic situations, one is interested in locating the boundaries of the room. These walls cause acoustic reflections that influence the received signals in the embedded microphones. Estimating the locations of these reflecting surfaces is typically done in several stages. In the next section, these steps are explained and the prior art for each of these submodule is discussed. In the remainder of this chapter, background information is provided in room acoustics and loudspeaker modeling. At the end of the chapter a general signal model is provided and it is explained how this signal model can be used to locate reflecting surfaces.

2.1 Prior art

This thesis is focused on estimating the location of reflectors by using a loudspeaker with embedded microphones. The literature on this topic can be group according to their approach (and priors) and initial assumptions. To begin with, there is a tendency in literature to group three distinct scenarios: i) the loudspeaker and microphone signals are synchronized and the loudspeaker excitation signal can be freely chosen, ii) the loudspeaker and microphone signals are synchronized and the loudspeaker signal is predermined but observable (eg. a user is playing music) and iii) the loudspeaker and microphone signals are unsynchronized and the loudspeaker signal is unknown. In this thesis, scenario 1 is assumed.

In [10] three popular excitation signals are compared in performance. The exponential sine sweep is used in the coming chapters, one reason for using this over the maximum length sequence is that the non-linear effects of a loudspeaker are removed more easily when excited with a sine sweep [11]. More information about the exponential sine sweep and its inverse is provided in Appendix B.

In general, when assuming that the excitation signal of the loudspeaker is a known signal $x(n)$, the problem of finding the wall locations can be seen as a sequence of steps. In Fig. 2.1 these steps are given by:

1. Channel estimation - The channel of interest is convolved with the excitation signal $x(n)$. To estimate the channel, the excitation signal must be deconvolved from the microphone measurements.

2. Loudspeaker removal - A loudspeaker model is used to remove the influence of the loudspeaker response from the channel estimate. For an omnidirectionally symmetric loudspeake, the influence is removed using deconvolution. If one considers

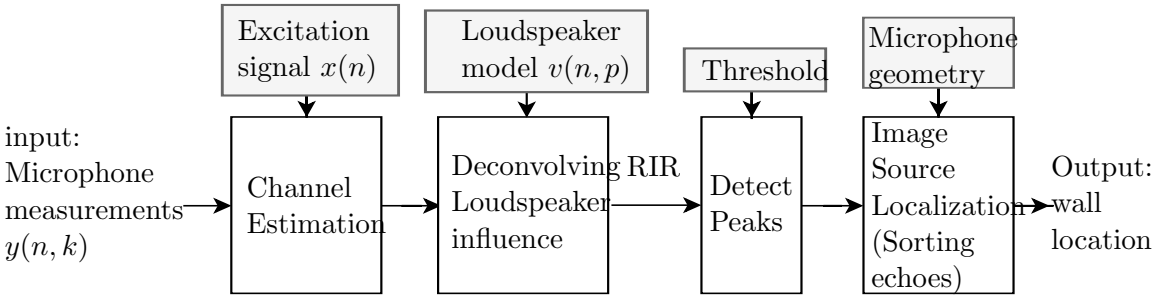


Figure 2.1: Typical processing stages from microphone measurements to wall locations. Here RIR means Room Impulse Response, as explained in Section 2.2.2

a typical loudspeaker, then due to the innherent directivity this step is more complicated and cannot typically be performed as an isolated step. The hope is that this results in the room impulse response (explained in Section 2.2.2).

3. Detect/pick peaks - Assuming that the wall has no influence on the signal, clear distinct pulses should be visible. The peaks from the room impulse responses correspond to reflections and must be detected. This detection problem requires knowledge on the number of peaks or needs an appropriate threshold. Inherently, there is a trade-off between false positives and false negatives.

4. Echo sorting - Since we have multiple microphones, the acoustic echoes must be labeled according to the wall which produced them. This sorting of delays is a combinatorial problem (and NP-hard) [12] and depending on the microphone geometry in the room, may have a large search space.

5. Reflector localization - Since all methods assume that the microphone locations are known, if at least three microphone locations are combined with three times-of-arrival then the origin of the reflection can be localized.

The problem that motivates the writing of this thesis is that in general detecting the peaks in the room impulse response (step 3), that correspond to acoustic reflections, is a major challenge. In domestic rooms one can find many smaller obstacles that are reflective but are not acoustically dominant for most room exploitation strategies. Thus, most of the peaks present in the room impulse responses do not correspond to dominant reflectors, and the ones that do have been distorted. Often in literature white noise is added to synthesize more challenging scenarios; however, in practice the peak picking methods fail, even under high SNR scenarios [6].

For example, methods that seek to estimate the channel between acoustic transducer and microphone based on adaptive filtering have been extensively researched [13, 14]. By assuming a power limited room impulse response, it was found that Tikhonov regularization (ℓ_2-norm regularization) can improve the identification of the channel and even be optimal. However, since these publications neglect to detect the delays from reflectors, it provides no guarantees to be optimal in a processing chain to solve the problem of this thesis.

Similarly, to account for a more complete loudspeaker model, the image source method [15] (explained in Section 2.2.1) has been extended for directive sources [16, 17, 18]. By modeling the loudspeaker's directivity pattern using spherical harmonics, an orthogonal basis is provided to model all loudspeakers, in theory. However, due to computational complexity, the modeling of loudspeaker directivity using spherical harmonics has not been included as step 2 in any processing chain to the best knowledge of the author.

As mentioned previously, the echo sorting problem raised in step 4 can be solved in many ways. First of all, if the loudspeakers are placed randomly in a room, the problem is computationally much more demanding. Whereas if a microphone array is used, beamforming techniques can be used that greedily assign groups of echoes by computing the steered response power for all directions. The methods that seek to solve the combinatorial problem in the echo sorting stage all assume that noiseless echo delay timing information is available [12, 19]. Thus, if a set of delays is given that is geometrically inconsistent, then the method will detect this inconsistency and try a different group labeling of each echo. Fortunately, this means that the methods can detect errors in the timing information. However, the mathematical framework from that area of research cannot easily be adapted to work with noisy echo information nor a probability distribution using a stochastic framework.

Recently, methods that combine the peak detection (step 3) in the multichannel room impulse response with the image source localization (step 4) have been proposed. These methods assume that the microphone array is *compact* with respect to the source distances. This allows for a plane wave assumption for reflections. In [20] uniform circular arrays are used to detect peaks from the multichannel response only if they are geometrically consistent. The ideas developed there are inspirited from image signal processing. The proposed method, however, loses performance as the number of wall reflections increase. In general, the method is biased and performs poorly in separating two wall reflections that arrive in close succession or from similar direction of arrival.

High-resolution techniques to resolve multiple (real) acoustic sources using compact microphone array solves this problem [21, 22, 23]. By assuming a sparse number of sources, the method can separate even closely placed sources by formulating the solution as an inverse problem. The ℓ_1-norm is used as penalization to force a sparse solution. The idea to use the sparsity prior on the room impulse response was an earlier idea. In [24] it is argued that if the room impulse response is constructed from a sparse number of nonnegative reflections, the linear deconvolution performs a pseudo-inverse on a matrix that may be ill-conditioned. Instead given this prior the ℓ_1 penalization is a Bayesian approach, within a probabilistic framework, leading to an expectation-maximization (EM) procedure that infers the optimal regularization parameters..

The sparsity assumption on the image sources is a prior that is also used successfully in more recent work for solving the room impulse response interpolation problem. This problem seeks to extrapolate room impulse responses for different positions in a room. Solutions based on compressive sensing use the ℓ_1 regularization to enforce a sparse number of sound rays have shown to increase performance [25, 26, 27].

In [28] an end-to-end method is proposed that is aware of the loudspeaker directivity model and uses the ℓ_1 penalization to regularize for priors. The method matches the

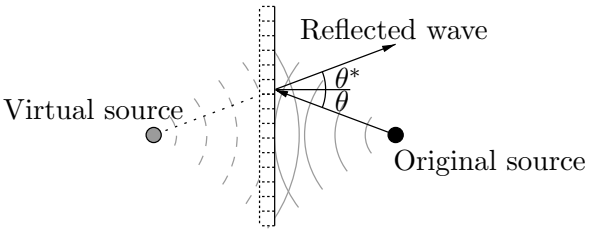


Figure 2.2: Specular reflection is equivalently modeled with a virtual image (copy) of the original source positioned at the far side of the reflective boundary. Source: [1]

measured signal with atoms in a dictionary. The dictionary is constructed from many measurements in an anechoic chamber. The measurements are setup with a single loudspeaker and wall. The loudspeaker is rotated for a total of 240 direction of arrivals.

An overview of the literature mentioned here is provided in Table 2.1. The papers are categorized by the steps from Fig. 2.1 and a short summary is given.

2.2 Background information on room acoustics

The formal description of the evolution of the sound field in any fluid is given by the *acoustic wave equation*. This equation is in the form of a second order partial differential equation (PDE). In general solving the acoustic wave equation in practical scenarios is challenging.

Numerous acoustic modeling techniques exist, derived from the governing PDE. In particular, in room acoustics, so-called geometric acoustics can be useful. Instead of wave-theory based methods that describe the dynamics of wave propagation and wave refraction derived from the PDE, the idea behind geometrical acoustics is a simplified theoretical description of the wave propagation. It replaces the concept of sound waves with the concept of sound rays. Inspired from geometrical optics, the sound rays propagate on a narrow straight path. This description fails to model room modes at lower frequencies, but for higher frequencies can be a useful model. If the sound ray encounters a surface larger than the wavelength, it is assumed that the sound ray reflects specularly (Fig. 2.2). This is similar to a mirror-like reflection in optics. In typical domestic rooms this assumption holds for frequencies higher than 1000 Hz. In the next section the mirror image source method is explained and the room impulse response is introduced. If the reader is interested in more room acoustics background, then one is referred to work by Heinrich Kuttruff [34] and Finn Jacobsen [35].

2.2.1 Mirror Image Source Method

In 1979 Allen and Barkley introduced the mirror image source method (MISM). This acoustic model is specific to modeling the sound field in a room with rigid walls. The MISM seeks an expression of the sound field that is a superposition of one real source and infinitely many image sources (IS). The image (or virtual) sources are sources that model the mirror-like reflections that occur at room boundaries. The image source method furthermore assumes that the acoustic properties of the wall are uniform. In

Table 2.1: Literature from 20 sources. Most references focus on a single step in the processing chain. Most notably is the difference in loudspeaker modeling.

	Channel Estimation	Loudspeaker model	Echo detecting	Delay sorting
Buchner [13], Waterschoot [14]	Adaptive filtering			
Stan [10]	Linear Deconvolution			
Torras-Rossel [11], Farina [29]	Sweep measurements	Non-linear artifacts removed		
Chen [30]	Generalized Cross Correlation			
Lin [24]	Sparse deconvolution		ℓ_1 regularization	
Antonacci [31]	Linear deconvolution	Matched filter with omnidirectional assumption	Peak picking	Common tangent Estimation of Ellipses
Tervo [6]	Overview of Generalized Cross Correlation		Peak picking	Beamforming: Maximizing steered response power
Brooks 2006 [21], Tiana-Riog [22], Lyllof [23]			Real uncorrelated source localizationℓ_1 regularization	Plane wave decomposition using template mask
Tervo [32]		Highly directional loudspeaker	Only detect peaks that are spatially correlated	Echoes are already grouped in DOAs as the room is probed with highly directional loudspeaker
Samarasinghe[16], Bu2017 [17] , Hafezi 2015 [18]		Spherical harmonics - Image source method		
Dokmanic [12]				Eucledian Distance Matrix based approach
Coutino [19]				Subspace based filtering
Ribeiro [28]		100's of Loudspeaker-wall measurements as dictionary atoms	ℓ_1 regularization	Loudspeaker response in dictionary atoms
Remaggi [33]	DYPSA		Peak picking	Overview on various geometric methods
Torres [20]	Maximum length sequence - Linear deconvolution		Image processing based thresholding	Plane wave decomposition using template mask
Proposed method	Sparse deconvolution	two dimensional loudspeaker impulse response model - Directivity aware model	ℓ_1 regularization	Plane wave decomposition using template mask

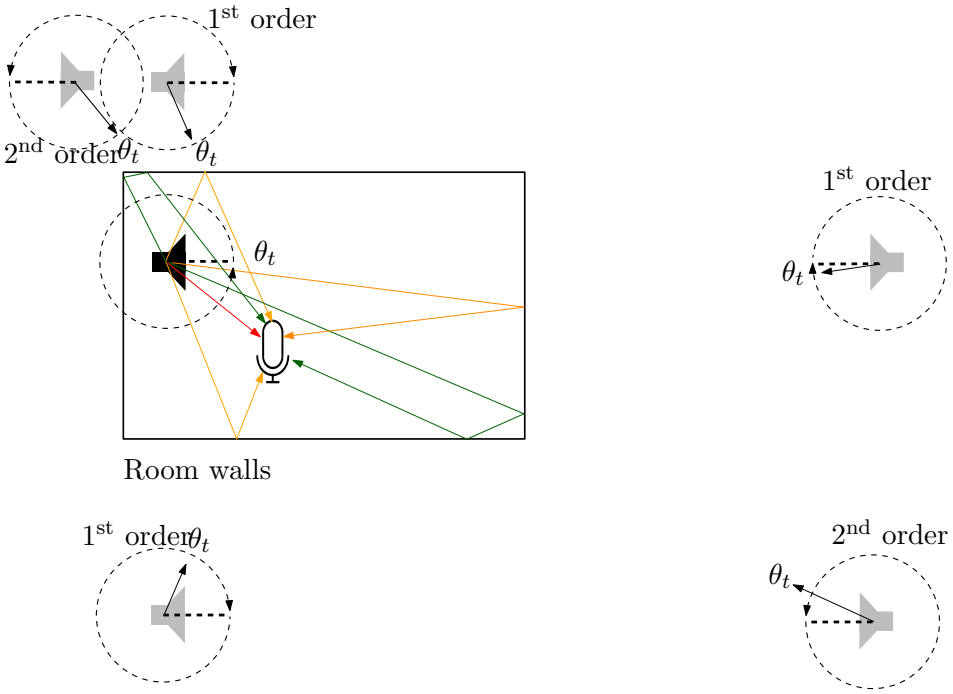
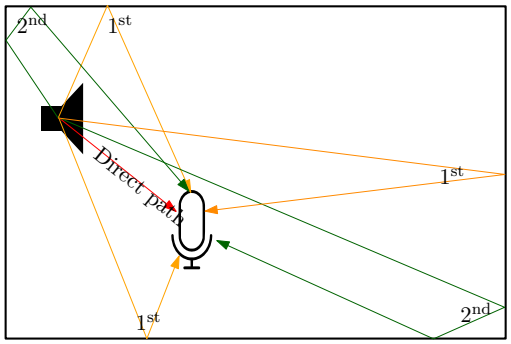


Figure 2.3: Top view of a loudspeaker and microphone in room. The image sources are depicted in gray. The coordinate system of the image sources is reflected. The microphone measurements are convolved with the loudspeaker impulse response evaluated at the transmit angle (black arrow).

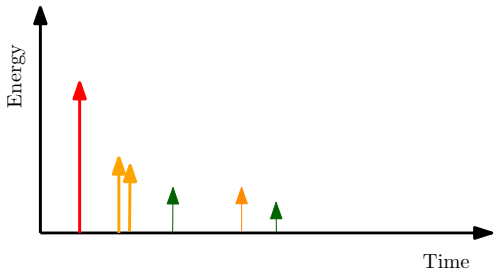
Fig. 2.3 three first and two second order image sources are depicted. Besides the direct path (red), there are multiple paths that sound rays will travel to reach the microphone. The acoustic echoes caused by these distinct reflections are characterized by a delayed that is proportional to the total distance traveled.

The MISM has since been extended for directive sources [16, 17, 18]. In this literature, the directivity of any source is denoted by its spherical harmonic decomposition coefficients. Each interaction with a wall applies a transformation to the coordinate system, such that the coordinate system is reflected with respect to the wall. As can be seen in Fig. 2.3 the transmit angle θ_t is defined at the source location with $\theta_t = 0$ (depicted with the dotted line) to the right of it and by counting counter-clockwise (denoted by the arrow). Since, the coordinate system is reflected for each wall interaction, the resulting axis on which θ_t is defined is depicted for each image source in the Figure. It is important to realize that each sound ray (and consequently each virtual source) in the room has its own transmit angle θ_t. The transmit angle is defined by the vector connecting the image source location and the microphone position; however, the angle must be projected on the coordinate system of the image source.

The image source method is often used to motivate that the room impulse response consists of a sparse set of delays in the early part. In the next section, the room impulse response is explained in further detail.



(a) Top view of a room. A sound source and microphone are placed. The room boundaries reflect the sound rays to create multiple paths to the microphone. Red: direct path, orange: first order reflection, green: second order reflection.



(b) Model of the early part of the room impulse response according the the MISM. The wall reflections cause delays, whose energy decreases with distance/time.

Figure 2.4: Early part of the room impulse response modeling.

2.2.2 Room Impulse Response

The room impulse response (RIR) can be defined by assuming that the room influence on a measured signal can be modeled as a linear time invariant (LTI) system. In the literature, sometimes the Acoustic Transfer Function (ATF) is used instead. The ATF is the Fourier Transform of the RIR and is often used by acousticians. Measuring the room impulse response using a loudspeaker and microphone is a challenge in itself, as it is not trivial to remove all influence from the microphone and loudspeaker. Those are typically not flat-frequency nor omni-directional. In early works, the room impulse response was measured by exciting the room with a balloon pop or a toy gun. If a loudspeaker is used instead, a known excitation signal can be used to identify the impulse response of the LTI room system. An overview of pilot signals and methods is given in [10, 29] and in Appendix B the exponential sine sweep method is explained.

More often, however, the room impulse response is used to impose some signal model on measurements. The room impulse response can model the sound field in enclosures using spatially varying LTI-systems. The RIR will change as a function of wall locations, loudspeaker position and listening position. To understand how the early part of the room impulse response relates to the geometry, in Fig. 2.4 an example is given (not to scale). The length of the sound rays determines the delay of the echo in the RIR. The energy of the echo is inversely proportional to the distance and tends to decrease at each reflection, as the wall absorbs some energy. It must be noted that second-order reflections may arrive earlier than some first-order reflections, as is depicted in the example. The early part of the RIR can be seen as a superposition of pulses, each of which corresponds to an attenuated delayed version of the original signal. Let a denote the room impulse response, consisting of the direct path (dp) and a superposition of Image Sources a_{IS}. The location of these image sources are denoted by R, θ_r. Furthermore, as depicted in Fig. 2.3, each IS has a particular direction facing the microphone. This is needed to assign the correct loudspeaker impulse response

for directive loudspeakers and is denoted as the transmission direction θ_t. Let $a^{(\mathrm{dp})}(n)$ denote the direct path channel, assuming that there are S image sources we have

$$a_{\mathrm{room}}(n) = a^{(\mathrm{dp})}(n) + \sum_{i=0}^{S-1} \rho_i a_{\mathrm{IS}}^{(R_i,\theta_i^t,\theta_i^r)}(n). \tag{2.1}$$

From Eq. (2.1) one can see that following this model, the early part of the room impulse response is a superposition of the direct path and S virtual sources. The contribution of each virtual source is denoted to be a function of the source location R_i, θ_i^r and the attenuation(s) of the wall(s) $\rho^{(i)}$. The reason for including θ_i^t is to model the correct loudspeaker response, aware of directivity. This is explained in further detail in the next section.

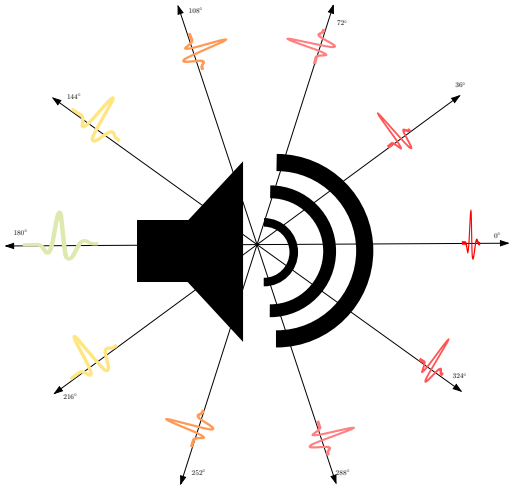
Of interest for the problem of this thesis is to estimate R_i, θ_i^r, from noisy microphone measurements. If one assumes that the loudspeaker enclosure with the built-in microphones is constant, then the direct path is time-invariant. Throughout this thesis, it is assumed that the direct path channel $a^{(dp)}(n)$ can be measured a priori and is LTI and its influence can be removed perfectly.

2.3 Loudspeaker modeling

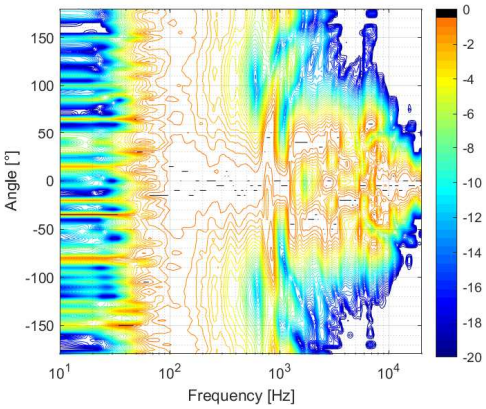
A loudspeaker is an electroacoustic device that connects the realm of electronics with the world of sound. A loudspeaker can convert an electric signal into pressure changes in the air around it. Often of much interests is the frequency response of a loudspeaker. The rule of thumb is that to reproduce music; one needs the full auditory band of 20Hz - 20kHz or even higher. Measuring the loudspeaker impulse response can be done by placing the loudspeaker under test in an anechoic room. A pilot signal is emitted and the response is measured with a microphone. The loudspeaker is assumed to be a linear time-invariant system, whose impulse response is causal and finite. To identify this system, typically a known signal is used to excite the loudspeaker, this is explained in further detail in Appendix B

In the application of reflector localization, using a loudspeaker model is useful for predicting the contribution of a reflector at a particular location, on the measured microphone signal. The more precise the loudspeaker model, the better the prediction and thus the inverse problem of estimating the locations given measurements are also improved. One naive way to do so is to include a measured loudspeaker impulse response, as done in [31]. However, as is shown in Fig. 2.5 the loudspeaker impulse response is not equal in each transmitted direction. In Fig. 2.5b it is shown that if a microphone is circled around the loudspeaker in the horizontal plane, the magnitude frequency response bandwidth is maximum when directly in front of the loudspeaker cone. One can also observe that at the back of the loudspeaker, the bandwidth is reduced. This makes the localization of the corresponding image source more challenging [6].

If we assume that the loudspeaker can be modeled as a linear time invariant system, then the loudspeaker impulse response can be used to define the input/output relationship of such a system. Before doing so, it must be noted that the loudspeaker response



(a) Les formal drawing. Top view of a single driver loudspeaker. Model is shown with $P = 10$ transmit angles. In general the impulse response has higher power (red) in the on-axis direction and lower in the back (green). Furthermore the on-axis loudspeaker impulse response is more spiky as it has higher frequencies.



(b) Loudspeaker absolute frequency response (dB) measured at 1 meter away for various angles of transmission. One can see that the on-axis angle 0° has a high response for high frequencies above 10Khz. For off-axis angles the bandwidth of the loudspeaker decreases substantially.

Figure 2.5: Figures depicting the inherent directivity that loudspeakers have

$v(n)$ is a function of listening position $v(n, \mathbf{r})$, in other words of the direction of transmission and distance [36]. Furthermore, by our construction the loudspeaker response $v(n)$ does not include the propagation delay. Therefore, if the distance r_0 is sufficiently far, such that a far-field assumption can hold, then we have that for distances further away $r_* \geq r_0$ we have

$$v_{\mathrm{c}}(n, r_*, \theta_*) = \frac{r_0}{r_*} v(n, r_0, \theta_*) \tag{2.2}$$

where c denotes a continuous signal in the spatial dimension. It must be noted that the far-field distance is proportional to the wavelength. As a result, in broadband scenarios, the far-field assumption may not hold for the lower frequencies. If one can disregard a scalar ambiguity, then this far-field assumption can be used to model a loudspeaker only as a function of the direction of transmission. It can therefore be argued that in the far-field $v_{\mathrm{c}}(n, r_0, \theta_i)$ for some grid of θ_i is sufficient to model the signal for that source at any position in the room further away than r_0. In the next chapter, the loudspeaker model is given by a two dimensional **discrete** signal $v(n, p)$ for $n = 0, \ldots, K-1$ and $p = 0, \ldots NP - 1$. Where p samples the direction of arrival uniformly, much like in Fig. 2.5a.

In Eq. (2.1) one can see that the room impulse response can be decomposed in the direct path $a^{(\mathrm{dp})}(n)$ and one contribution for each mirrored image source $a_{\mathrm{IS}}^{(R_i, \theta_i^t, \theta_i^r)}(n)$. The angle of arrival θ_i^r is used to compute the correct relative delays for each microphone in the array, as is explained in the next section. The room impulse response and loudspeaker model can be combined, by including the loudspeaker impulse response

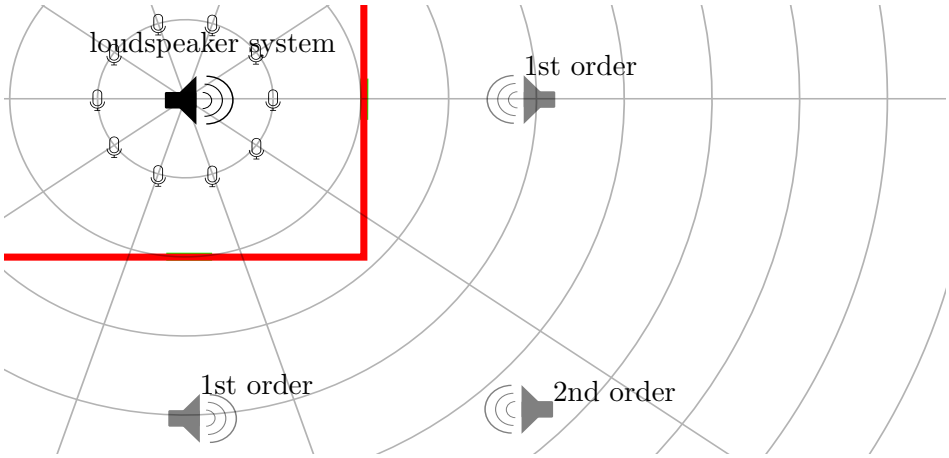


Figure 2.6: Top view of compact microphone array and real loudspeaker placed in a corner. The image sources are reflected at the wall. Because the walls form a 90° angle, the second order reflection is received and transmitted at $\theta_r = \theta_t = \frac{7}{8}\pi$.

in a_{IS}. The number of wall reflections for an image source determines the order of that source and may change the relationship between the direction of arrival θ_r and loudspeaker transmit angle θ_t. In general, for first-order image sources, the situation is similar to Fig. 2.6. No matter how the wall is oriented, the wall reflection always occurs at the normal vector of the wall. This is a direct consequence of placing the loudspeaker at the center of the microphone array. The same does not hold for second-order reflections. However, if one assumes orthogonal walls, ie. a rectangular room, then it holds that $\theta_t = \theta_r$ for all second-order reflections. Remember that we defined $v(n, p)$ as a two-dimensional discrete signal, and we sample the angle NP times. If we can assume some smoothness over the transmit angle θ_t such that the off-grid errors are small, we have

$$v_{\text{c}}(n, \theta_*) \approx v\left(n, \left\lceil \frac{\theta_* NP}{2\pi} \right\rfloor\right). \tag{2.3}$$

Finally, the observation model for a virtual source can be parameterized by R_i and θ_i

$$a_{\text{IS}}^{(R_i,\theta_i)}(n) = \delta\left(n - \frac{R_i}{c}\right) * v\left(n, \left\lceil \frac{\theta_i NP}{2\pi} \right\rfloor\right). \tag{2.4}$$

2.4 Compact Microphone Array response

Consider N microphones in an array near the origin of a polar grid and a single source at $\mathbf{r}_s = \left[R_s, \theta_s\right]^{\mathsf{T}}$ as depicted in Fig. 2.7. The Figure also illustrated that each microphone is bounded to be closer to the origin than r. In literature if r is sufficiently small with respect to the source distance, than the microphone array is *compact*. In this subsection, the linear response for the microphone array is defined. This thesis is focused on the delay-and-sum beamformer idea. It is based on delaying the signals captured at each

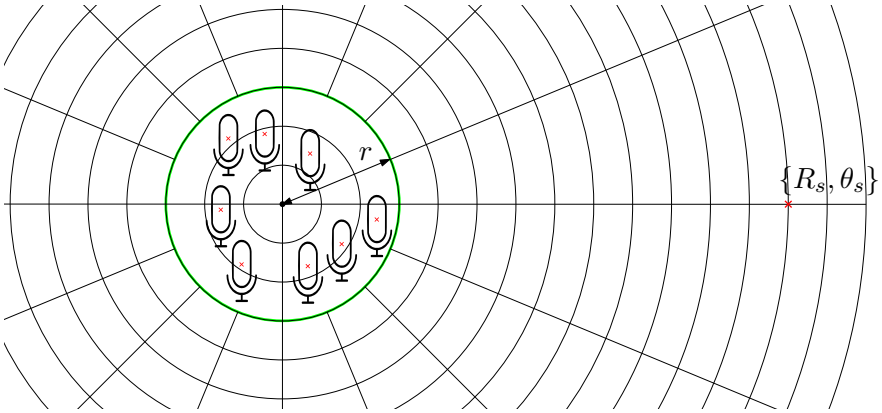


Figure 2.7: Top view of compact microphone array on the left and a single source on the right. The compact microphone array is bounded by the circle of radius r.

microphone by a specific amount and adding them up to focus the system to a specific direction in space. In general the total delay is proportional to the distance from each microphone to that point in space. Let the speed of sound be denoted by c, then the time to travel a distance d is given by

$$t = \frac{d}{c}. \tag{2.5}$$

To focus the discrete microphone measurements $y(t, i)$ for $t = 0, \ldots, T$ and $i = 0, \ldots N-1$ into the point in space of $\mathbf{r}_s$ a multichannel filter is used and the outputs of each filter are summed.

$$q(\mathbf{r}_s) = \sum_{i=0}^{N-1} \delta\left(t - \frac{d_i}{c}\right) * y(t, i) \tag{2.6}$$

Where $*$ is the linear convolution operator in time and d_i is defined as the distance between $\mathbf{r}_s$ and the ith microphone.

The trick that is often applied in beamforming is to assume that the source is in the far field, ie $R_s >> r$. Then the beamformer delay d_i can be decomposed in a constant factor for all i and a relative component. We now define the relative distance $\Delta d_i(\theta_s) = d_i(\theta) - R_s$. Please note that on average for half the microphones this value is negative. As a consequence the angle θ and R_s are decomposed. By using the property that $\delta(t-a) * \delta(t-b) = \delta(t-a-b)$ we have

$$q(\mathbf{r}_s) = \delta(t - \frac{R_s}{c}) * \sum_{i=0}^{N-1} \delta\left(t - \frac{\Delta d_i(\theta_s)}{c}\right) * y(t, i) \tag{2.7}$$

The computation is reduced, as the impulse response that has to be applied for each microphone is of low maximum order. This is because the maximum delay $\frac{\Delta d_i(\theta_s)}{c}$ is bounded by the microphone radius r as $\frac{r}{c}$.

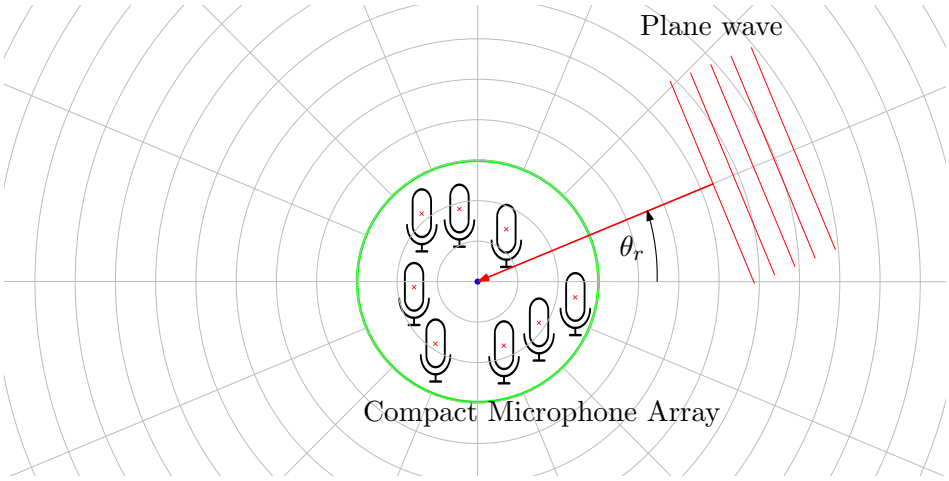


Figure 2.8: A source in the far field causes a plane wave on the compact microphone array. Depicted is a plane wave arriving from θ_r. The relative delays for the delay-and-sum beamformer are only a function of θ_r.

2.5 Signal model

In the preliminaries, the concept of geometric acoustics and the image source method were introduced. In this section, a complete measurement signal model is presented. Figure 2.9 depicts the components in the signal model that are going to be considered. The Figure is a top view. The ceiling and floor are not depicted in the Figure and are disregarded in the signal model presented here.

Initially, the digital to analog converter generates an analog pilot signal $x(n)$ for the loudspeaker. The (active) loudspeaker is excited by this signal and emits a filtered version of this signal to its surroundings. The loudspeaker impulse response provides a model to calculate the acoustic output signal of the loudspeaker (in the far-field).

We make the fundamental assumption that source and receivers lie on the same plane and the lying plane of the reflector is orthogonal to this plane. In this scenario, the geometry of the acoustic scene is described by the plane in which sources and receivers lie in two dimensions. In general, the impulse response function of the loudspeaker depends on the angle of transmission θ_t. Since this model is only concerned with modeling vertical walls, the azimuth angle is disregarded here. Once the sound ray hits the wall, it is convolved with the wall impulse response, that is a function of wall material and wall position. The signal is then convolved with a delay, to account for the transmission distance. In other words, the delay contains information about the distance of the wall and is characterized by a peak in the channel response. Lastly, the sound ray arrives at the compact microphone array with angle θ_r. The array geometry determines the relative measurements for each microphone.

Previously, in Eq. (2.4) the reflection of a single source is given as a function of loudspeaker transmit angle. If we combine this Equation with the *plane wave response* on a compact microphone array, given in Eq. (2.7) we obtain a measurement model for

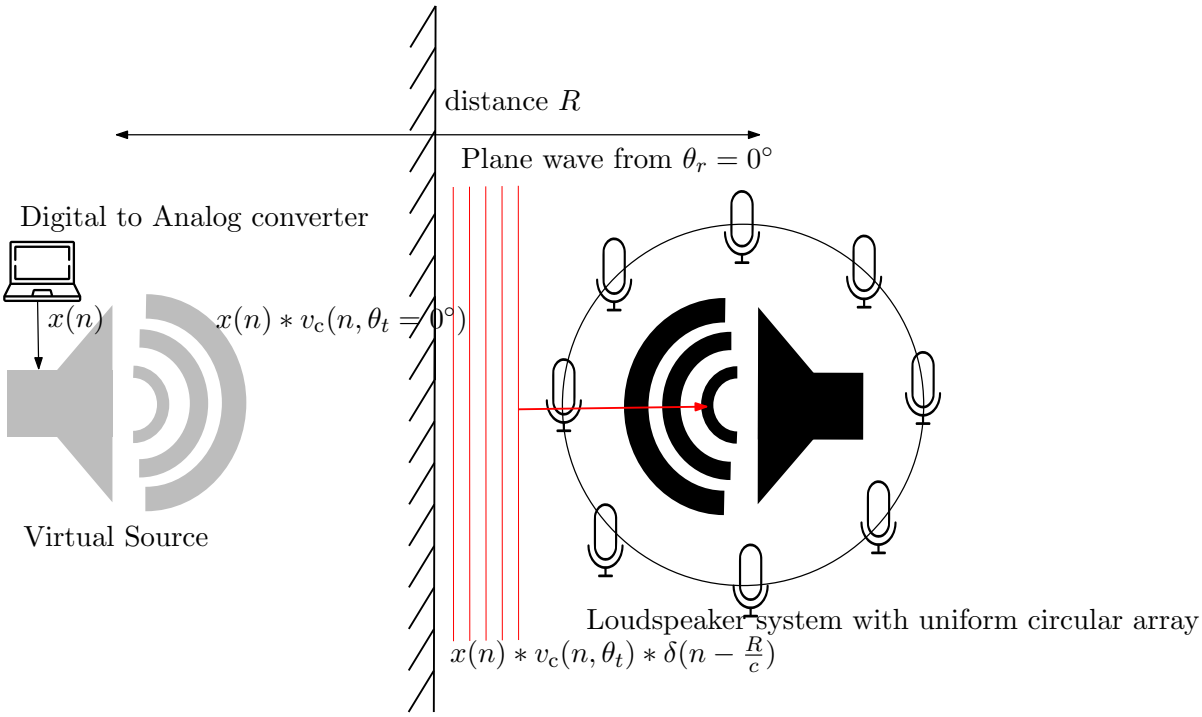


Figure 2.9: Top view of a loudspeaker with compact microphone array close to a wall. The mirrored virtual source is depicted in gray. The measurement model for the plane wave of this image source is given by the loudspeaker response in the on-axis direction $v_c(n, \theta_t)$ delayed by R/c, where R is the total distance. The array response for the plane wave is neglected here.

the kth microphone at time n:

$$y(n,k) = \underbrace{x(n)}_{\text{Excitation signal}} * \left[a^{dp}(n) + \sum_{i=0}^{S-1} \underbrace{v\left(n, \left\lceil \frac{\theta_i NP}{2\pi} \right\rfloor\right)}_{\text{Loudspeaker model}} * \underbrace{\delta\left(t - \frac{R_i}{c}\right)}_{\text{Propagation delay}} * \underbrace{\delta\left(t - \frac{\Delta d_k(\theta_i^r)}{c}\right)}_{\text{Mic Array response}} \right] \tag{2.8}$$

where $a^{dp}(n)$ is the direct path channel between the loudspeaker and built-in microphones, $v\left(n, \left\lceil \frac{\theta_i NP}{2\pi} \right\rfloor\right)$ denotes the loudspeaker impulse response transmitted at θ_i in the far-field, R_i is the distance of the ith image source and $\Delta d_k(\theta_i^r)$ is the relative delay for the kth microphone for a plane wave arriving from θ_i and is determined by the array geometry. It must be noted that it is assumed that $\theta_r = \theta_t$ which holds for geometries as depicted in Fig. 2.9. In the next chapter, Eq. (2.8) is evaluated for a specific compact microphone array: The uniform circular array. It is shown that the symmetry can be exploited, to efficiently compute the measurement model for many R_i and θ_i on a uniform polar grid by restating the microphone array response as a two dimensional (circular) convolution. Evaluating Eq. (2.8) for many candidate locations is essential, as the problem is solved as an inverse problem. As explained in greater detail in Section 2.6.

2.6 Sparse deconvolution

Sparse deconvolution is a deconvolution technique that uses prior knowledge in the model to increase performance. Suppose there is a LTI system that one wishes to estimate . A pilot signal excites this system and the output is observed. One can write the signal output as the convolution of the input with the system's impulse response function $h(n)$

$$y(n) = \sum_{k=0}^{k=L} x(k)h(n-k) \tag{2.9}$$

Where x is of length L, h is of length T and y is of length $T + L - 1$. Since the convolution is a linear operation, it can also be written down in matrix-vector notation as

$$\mathbf{y} = \mathbf{T}_x\mathbf{h}, \tag{2.10}$$

where we have that $\mathbf{y}$ and $\mathbf{h}$ are defined as

$$\mathbf{y} = \big[y(0), y(1), \ldots, y(T+L-2)\big]^\intercal \in \mathbf{R}^{T+L-1}, \tag{2.11}$$

$$\mathbf{h} = \big[h(0), h(1), \ldots, h(T-1)\big]^\intercal \in \mathbf{R}^{T}, \tag{2.12}$$

and $\mathbf{T}_x$ is an overdeterimned linear system of equations with a Toeplit structure of size $T + L - 1 \times T$ and is given by

$$\mathbf{T}_x = \begin{bmatrix} x(0) & 0 & \ldots & \ldots & \ldots & 0 \\ x(1) & x(0) & 0 & \ddots & & \vdots \\ \vdots & x(1) & x(0) & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ x(L-1) & x(L-2) & x(L-3) & \ldots & \ldots & x(0) \\ 0 & x(L-1) & x(L-2) & \ddots & \ddots & x(1) \\ \vdots & 0 & x(L-1) & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & x(L-1)) \end{bmatrix}. \tag{2.13}$$

The Toeplitz matrix can also be defined as a column subset of a Circulant matrix. A Circulant matrix must be constructed with a zero padded version of $\mathbf{x}$. The relationship with the Circulant matrix is useful, as it reveals the use of the Fast Fourier Transform to perform the matrix-vector multiplications in $O(n\log_2(n))$ flops instead of $O(2n^2)$. More information about Toeplitz and Circulant matrices and their properties can be found in Appendix A. The least squares estimate of $\mathbf{h}$ is given by

$$\hat{\mathbf{h}} = \underset{\mathbf{h}}{\operatorname{argmin}} \left\|\mathbf{y} - \mathbf{T}_x\mathbf{h}\right\|_2^2, \tag{2.14}$$

which has the closed form solution of

$$\hat{\mathbf{h}} = \left(\mathbf{T}_x^\dagger \mathbf{T}_x\right)^{-1} \mathbf{T}_x^\dagger \mathbf{y}, \tag{2.15}$$

where $(.)^\dagger$ denotes the Hermitian of the matrix. If the excitation is poor, then the the matrix is ill-conditioned. In general, the bandwidth of the excitation signal x determines the singular values of the Toeplitz matrix $\mathbf{X}$.

Another method for channel estimation is the Matched Filter, which approximates the matrix inversion. It is seen in practice when the pilot signal is too long to compute the pseudoinverse. It is assumes that $\mathbf{T}_x^\dagger \mathbf{T}_x \approx \alpha \mathbf{I}$. In Appendix B it is shown that for some pilot signal (the exponential sine sweep), the Matched Filter approximates the pseudoinverse very well.

Suppose now, that from the T filter taps, at most S are non-zero and $S << T$. This prior information can be used to improve the estimate significantly.

$$\hat{\mathbf{h}} = \begin{array}{ll} \underset{\mathbf{h}}{\text{argmin}} & \|\mathbf{y} - \mathbf{T}_x \mathbf{h}\|_2^2 \\ \text{subject to} & \|\mathbf{h}\|_0 \leq S \end{array} \tag{2.16}$$

This approach is referred to as sparse deconvolution and has application in channels consisting of distinct multi-paths.

3 Proposed Design

Consider a system that has at least one loudspeaker and a compact microphone array with N microphones and known geometry. The system is placed in an unknown room and the loudspeaker can actively probe the room by emitting a known signal. We assume that the transmitter and receiver are synchronized and coincide geometrically. The proposed design is based on a geometric acoustic model. The acoustic waves are modeled as sound rays. For first order reflections we thus have that the reflection points on the wall are orthogonal to the sound ray. Figure 3.1 depicts a top view of a system placed in a rectangular room. The first order reflections probe the wall in the green areas and it is assumed that the wall extends linearly with similar acoustic characteristics.

Much like in [28] the microphone measurements are used in an ℓ_1-regularized least squares to fit with synthetically generated echoes, where the reflector localization is solved as an inverse problem. The measurement model is the *forward* model, that can predict the microphone measurements for any room, given the wall locations. In Eq. (2.8) one such model is provided. In this chapter, the forward signal model is formulated to have as input a signal, rather than a summation over all the image sources.

This chapter will first define the geometrical setup in the room, after which the microphone array response is given for the uniform circular array in Section 3.1. the *forward measurement model* is then presented and in order to solve the inverse problem using known least squares matching programs, the *candidate locations* are discretized. It is shown in Section 3.3 that if the candidate locations are defined on an uniform polar grid, then the symmetry in the array can be exploited to efficiently compute the forward model. Later, in Section 3.5 the forward measurement model is rewritten in matrix-vector notation. Finally in the last section of this chapter, the convex optimization problem that resolves wall locations from measurements is presented.

3.1 Problem scenario

Consider a loudspeaker system and a uniform circular array (UCA) of radius r with N microphones. In Fig. 3.1 one can see such a system. We define the coordinate system such that loudspeaker point source and center of the microphone array are at the origin and $\theta = 0^\circ$ corresponds with the direction at which the main loudspeaker transmits on axis. Consequently, the sampled point on the wall is at half the distance $\{(\frac{R_i}{2}, \theta_i\}$.

The proposed method generalized for any number of vertical walls. Take as an example the shoe-box shaped room, where we have four first order image sources and four more second order image sources. One can observe that in Fig. 3.1 the systems location with respect to the bottom left corner of the room is given by $\mathbf{p} = [5, 2]$m and

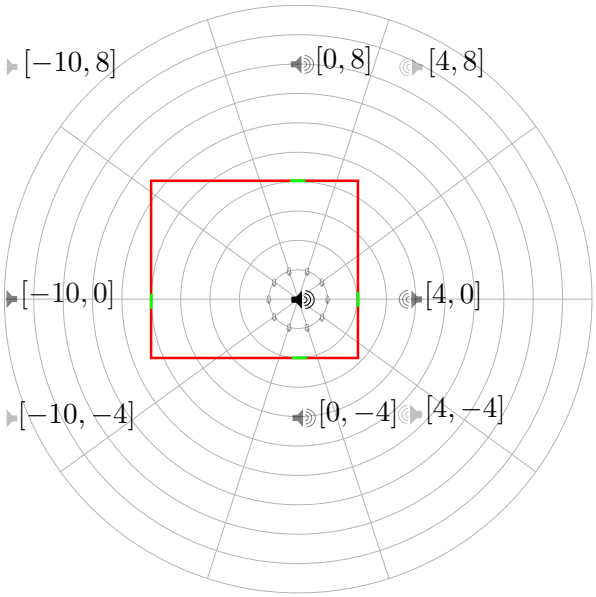


Figure 3.1: Top view of loudspeaker system with $N = 10$ microphones in a rectangular room. The polar coordinate system originates at the center of the array. The red line indicate the room boundaries. In gray all first and second order image sources are depicted. Following the sound ray model, the first order reflections only probe the wall at the green locations.

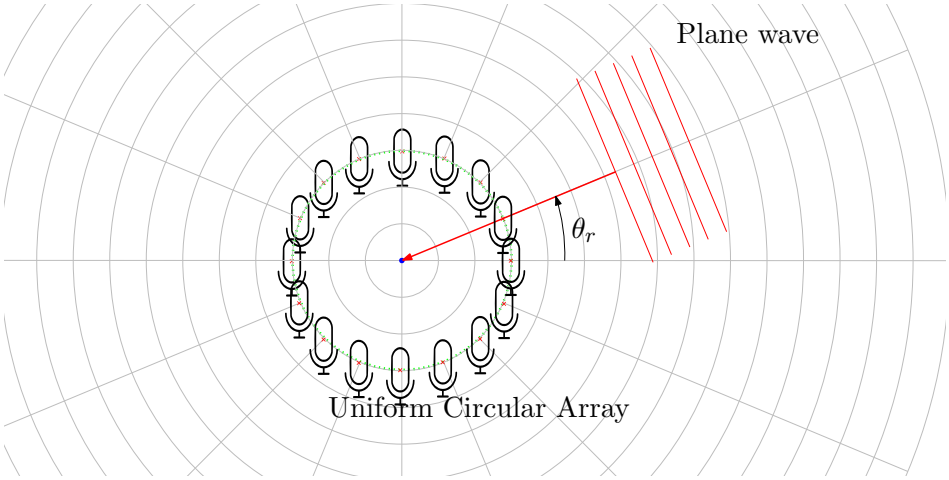


Figure 3.2: Uniform Circular Array with a plane wave arriving from θ_r

furthermore observe that the room size is $\mathbf{L} = [7, 6]$m, we have the following Cartesian coordinates for the image sources:$[-2p_x, 0]$, $[0, -2p_y]$, $[2(L_x - p_x), 0]$, $[0, 2(L_y - p_y]$. The second order images sources are given by $[-2p_x, 2(L_y - p_y)]$, $[-2p_x, -2p_y]$, $[2(L_x - p_x), -2p_y]$, $[2(L_x - p_x), 2(L_y - p_y)]$.

3.2 Plane Wave on Uniform Circular Array

Remember how in Eq. (2.7) the general equation for a plane wave on a compact microphone array is given. By construction, the *plane wave response* of a compact microphone array assumes that the source signal is in the far field. Therefore, the attenuation of the signal is approximately equal for all microphones. This *plane wave signal* must only be delayed appropriately to account for the array geometry. This is encapsulated in the relative delays for each microphone if at $t = 0$ the plane wave hits the array. In general these relative delays $\Delta d_i(\theta)$ are only a function of angle of arrival. This situation is depicted in Fig. 3.2. In this section it is shown that the Uniform Circular Array has symmetry in θ_r that can be exploited to compute the array response for many image sources on a uniform grid using the Fast Fourier Transform.

Consider an uniform circular array (UCA) consisting of N microphones. Remember that the center of the microphone array denotes the origin of the coordinate system. The microphone locations are denoted by $\{r_m, \theta_m\}$ and we have a plane wave arriving from θ_r at $t = 0$. This is depicted in Fig. 3.2. Where for a circular array, we have

$$r_{m,i} = r \forall i = 0, \ldots, N-1 \tag{3.1}$$

$$\theta_{m,i} = \frac{2\pi i}{N} \forall i = 0, \ldots, N-1 \tag{3.2}$$

Consider now a single source, whose location is $\{R_s, \theta_s\}$. The distance from each microphone to the source is given by

$$d(R_s, \theta_s, r_m, \theta_m) = \sqrt{R_s^2 + r_m^2 - 2R_s r_m \cos(\theta_s - \theta_m)} \tag{3.3}$$

Combining this Equation with the constraints from Eqs. (3.1) and (3.2) we obtain

$$d(R_s, \theta_s, r, i) = \sqrt{R_s^2 + r^2 - 2R_s r \cos\left(\theta_s - \frac{2\pi i}{N}\right)} \tag{3.4}$$

The goal here is to obtain an expression for the relative delays Δd, which is independent of source distance R_s. The reason for doing so, is that the input signal h already captures the delay caused by R_s. Of interest, is to compute the microphone array response for NP uniform steps between 0 and 2π, as is explained in further detail in Section 3.3. We can decompose the total distance from source to microphone as $R_s + (\Delta d_i - r)$ where only the second part is a function of microphone index i. The resulting Δd will explain a plane-wave event on the microphone array, which will only depend on the angle of arrival θ_s. In order to compute the response for a plane wave from angle θ_s, the constant R_s may be subtracted and r is added to avoid negative numbers (useful when we use this to define a discrete Finite Impulse Response (FIR) filter)

$$\Delta d = d - R_s + r \tag{3.5}$$

If we furthermore assume that the source is in the far field, such that we have $R_s >> r$

$$\begin{aligned}\lim_{R_s\to\infty} \Delta d(R_s,\theta_s,i) &= \lim_{R_s\to\infty} \sqrt{R_s^2 + r^2 - 2R_s r\cos\left(\theta_s - \frac{2\pi i}{N}\right)} - R_s + r \\ &= r\left(1 - \cos\left(\theta_s - \frac{2\pi i}{N}\right)\right)\end{aligned} \tag{3.6}$$

The proof of is provided in Appendix C. So now we can approximate Eq. (3.4) by

$$d(R_s,\theta_s,r_m,i) \approx R_s + r\left(1 - \cos\left(\theta_s - \frac{2\pi i}{N}\right)\right) \tag{3.7}$$

And so finally we obtain the *relative measured delay's* for a plane wave arriving from θ_s, which is not a function of source distance, but only of the source angle.

$$\Delta\tau_i\left(\theta_s, r_i\right) = \frac{r}{c}\left(1 - \cos\left(\theta_s - \frac{2\pi i}{N}\right)\right) \tag{3.8}$$

Observe that the maximum relative delay between two microphones is bounded by $\frac{2r}{c}$. We assume sampling rate in time of f_s, thus the maximum length of a discrete finite impulse response filter that captures the differences in delays $\Delta\tau$ has $W = \left\lceil \frac{2rf_s}{c} \right\rceil$ taps.

The key observation is that one must remember that the microphone measurements $y(n,k)$, can be interpreted as a two dimensional sampled signal. Where n samples in time and k samples in the *microphone dimensio*n. This microphone dimension is the green dotted circle in Fig. 3.2 and is indeed uniformly sampled. A closer look at Eq. (3.8) from the perspective of the ith *microphone sample*, shows that this is only a function of the difference $\theta_s - \frac{2\pi i}{N}$. If we wish to use the convolution theorem, we must evaluate θ_s with uniform intervals. This creates a shift-invariant *steering function* that only depends on the difference between the ith microphone and the source angle [23].

Similarly as in [20] define a template mask matrix $\mathbf{M} \in \{0,1\}^{\left\lceil f_s \frac{2r}{c} \right\rceil \times NP}$. At this point it must be noted that although there are N microphones, if one wishes to have NP candidate angle locations, then one must obtain a *higher resolution* template mask. A two dimensional circular convolution with h and m, will then explain the plane wave for NP microphone channels.

$\mathbf{m}$ is defined as follows

$$m_{n,p} = \begin{cases} 1 & \text{if } n = \left\lceil f_s \frac{r}{c}\left(1 - \cos\left(\frac{2\pi p}{NP}\right)\right)\right\rceil \\ 0 & \text{elsewhere} \end{cases} \forall n,p \tag{3.9}$$

As one can observe, the matrix $\mathbf{m}$ is essentially a delay and sum filter bank that is steered in $\theta = 0$. However, by circularly permutating(?) the columns of $\mathbf{m}$ one can steer into NP directions (in uniform steps).

By using the farfield assumption, $\Delta\tau$ has been constructed in such a way that it is independent of source distance R_s. As a result, the two dimensional convolution

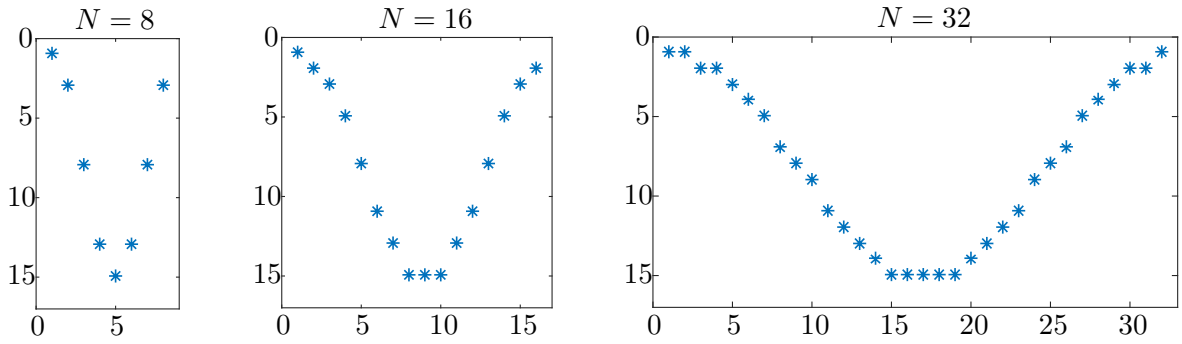


Figure 3.3: Generated masks for $c = 340$m/s $f_s = 48$kHz, $r = 0.05$m (gives $W = 15$) and varying microphone N ($P = 1$ assumed). Blue dots are the non-zero entries of the matrices. Each column has exactly one non-zero entry.

with the template, **can be computed as the product of two convolutions**. One convolution is delay (temporal translation) that is proportional to the source distance. The second convolution permutes the mask, such that any plane wave direction can be modeled. Specifically, one can now write a circular convolution in the microphone index dimension and a linear convolution in the microphone time dimension as a product of two convolutions

$$f(t_1, j) \triangleq \sum_{\alpha=0}^{NP-1} \sum_{d=0}^{T-1} h(d, \alpha) m_{t_1-d,[j-\alpha]_{\text{mod } NP}} \tag{3.10}$$

Observe how f is now defined for NP microphone channels, even though not all microphone channels will be used in the next section. In Eq. (3.10) there is a circular convolution in the *discrete microphone index* dimension j and a linear convolution in the *microphone time* dimension t_1. A physically motivated interpretation of this convolution is that it maps source directions of arrival to the microphone index dimension, which in our case is an uniform circular array. In other words it has input/output dimensions:
(source direction of arrival $\times$ source distance)$\rightarrow$(microphone channel $\times$ time).
Figure 3.3 depicts examples of this $\mathbf{M}$ mask for varying number of microphones.

3.3 Define a uniform polar grid of candidate image source locations

The signal model given in Eq. (2.8), although complete, has to iterate over all reflector locations R_i, θ_i to evaluate the contribution of each echo. If one wishes to solve the inverse problem, it may be beneficial to discretize R and θ and to define a grid of *candidate locations*. The general idea is to create a large dictionary of Rotated Image Source Impulse Responses (RISIR). The inverse problem is then solved by fitting a sparse number of these RISIR in the dictionary, with the microphone observations. Once the RISIR that are likely to be in the measured signal are estimated, they can be mapped back to wall locations.

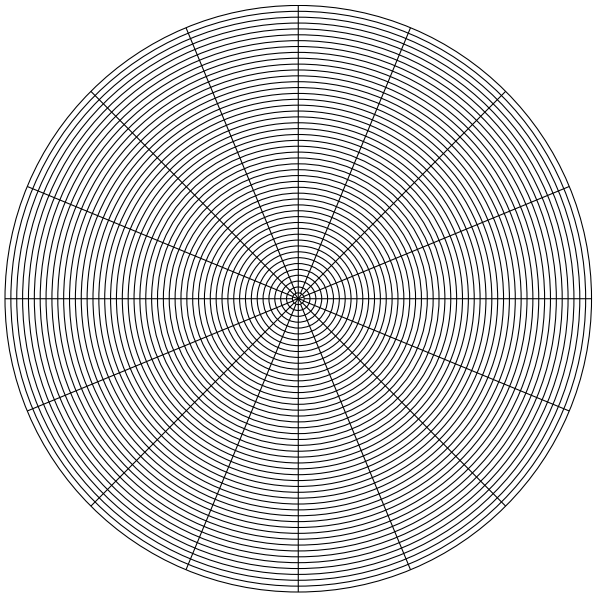


Figure 3.4: Uniform polar grid with $NP = 16$ angle divisions and $T = 50$ radial divisions. The lines intersect at the candidate points.

In this section, it is argued that using Eq. (2.8) the dictionary can be computed efficiently, for candidate locations on a uniform polar grid. By exploiting the symmetry of the uniform circular array, the image source responses can be evaluated for many directions of arrival efficiently. As a consequence, instead of iterating over a set of R_i, θ_i, an input signal h is defined. The length of h is equal to the number of discrete points on the grid. The index of the nonzero values in h will then correspond to image source locations.

Consider the set $\mathcal{H}$ that contains the location of S first and second order virtual sources, that dominate the early part of the room impulse response. A signal model of the room influence that only accounts for these S reflections can be parameterized by S locations (in two dimensions) of the corresponding image sources. Much like in the previous chapter, polar coordinates are convenient to use in this situation, denoted by $\mathbf{r} = \begin{bmatrix} R, \theta \end{bmatrix}^\intercal$ for $R \in [r, R_{\max}]$ and $\theta \in [0, 2\pi)$, where r denotes the radius of the compact microphone array. For now, the S image source locations are denoted by the set $\mathcal{H} = \{\mathbf{r}_i\}_{i=0}^{S-1}$, where each element in $\mathcal{H}$ contains the location of the ith image source in polar coordinates. To make use of matrix-vector operations though, the same information can also be represented in a vector.

If we have that our microphone measurements $y(n, k)$ and pilot signal $x(n)$ have been sampled in time with f_s, we discretize R with steps of $\Delta R = c/f_s$. We denote the total number of discrete steps in the distance by $T = \lceil R_{\max} f_s / c \rceil$ and we denote the discrete angles by NP, where N is the number of microphones and P is a natural number that determines the up-sampling factor. Thus we have a total of NPT *candidate locations* for which we can compute the measurement model. An example of a polar grid is depicted in Fig. 3.4. Next, we define a discrete signal h that contains all the NPT weights for each of these image sources.

The representation of the set $\mathcal{H}$ is mapped to a two dimensional discrete signal $h(n,p)$, where $n = 0,\ldots,T-1$ is proportional to the image source distance (and delay) and $p = 0,\ldots,NP-1$ which is proportional to the direction of arrival (DOA). The index of the nonzero values in $h(n,p)$ correspond to the distance and DOA of the image sources. We have that

$$h(n,p) = \sum_{\mathbf{r}\in\mathcal{H}} \frac{1}{R}\delta\left(n - \left\lceil \frac{R}{R_{\max}}T \right\rfloor\right)\delta\left(p - \left\lceil \frac{\theta}{2\pi}NP \right\rfloor\right) \tag{3.11}$$

where $\mathbf{r} = \big[R,\theta\big]^\intercal$ and where $\delta(n)\delta(p)$ denotes the two dimensional indicator function. Note that ρ_i in Eq. (2.8) is replaced by $1/R$. The reason for scaling the nonzero entries with the inverse of the source distance is because the contribution of a rotated image source impulse response in the measurement model is proportional to the inverse of the source distance. Observe how R and θ from the set $\mathcal{H}$ are rounded to the nearest discrete grid point by using $\lceil . \rfloor$. Throughout this thesis, it is assumed that the error introduced by the discretization in space, is negligible.

Since the measurement model will be expressed using matrix-vector products, it is convenient to now define a vector containing the elements of $h(n,p)$. This is defined as follows

$$\mathbf{h}_p = \big[h(0,p), h(1,p), \ldots, h(T-1,p)\big]^\intercal \in \mathbb{R}^T \tag{3.12}$$

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}_0 \\ \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{NP-1} \end{bmatrix} \in \mathbb{R}^{TNP} \tag{3.13}$$

The choice for stacking the $p = 0$ responses first, rather then the $n = 0$ is arbitrary. However, the Equations defined in Section 3.5 follow this convention.

Observe the relationship between the number of image sources S and the vector $\mathbf{h}$ as $\|\mathbf{h}\|_0 = S$, where $\|.\|_0$ denotes the ℓ_0 norm. For a room model of a shoe-box shape, the number first and second order reflections are $S = 8$. The input vector is sparse, since in general we have that $\|\mathbf{h}\|_0 = S << NPT$.

3.4 Evaluate the forward model for all candidate locations weights

In this section a linear measurement model is proposed to map the input signal $h(n,p)$ to microphone measurements $y(n,k)$. Remember that any two dimensional shoe-box shaped room has 8 first and second order image sources. The locations of these image sources are indicated by the nonzero indices of $h(n,p)$. If we combine the signal model from Chapter 2 given in Eq. (2.8) with the UCA response from Eq. (3.10) we obtain the *forward measurement model* as a function of h. Remember that this is used to compute a prediction on y to solve the inverse problem of locating walls from microphone

measurements.

$$y(i,j) = \sum_{t_1=0}^{L-1} x(i-t_1) \left[a^{\mathrm{dp}}(t_1) + \sum_{\alpha=0}^{NP-1} \sum_{d=0}^{T-1} m\big(t_1 - d, [jP-\alpha]_{\mathrm{mod}\ NP}\big) \sum_{t_2=0}^{K-1} v(d-t_2,\alpha) h(t_2,\alpha) \right] \tag{3.14}$$

Please observe how Eq. (3.14) performs a two dimensional convolution on input signal $h(n,p)$. In particular since the first dimension of h denotes the distance R_s, convolving in the time dimension will alter the delay of the rotated image source response. Secondly, a circular convolution in the second dimension of h, will permuted the microphone channels, such that any plane wave arriving from $\theta_r = 2\pi\frac{i}{NP} \forall i \in \mathbb{N}$ can be modeled.

Please remember that h is the input of the forward model and a^{dp}, v, m and x can be considered as constants.

1. $h(t,p)$ - Two dimensional grid, corresponding to an echo delay and directions of arrival. In practical room acoustic scenarios. we assume this is sparse.

2. $v(u,s)$ - Loudspeaker anechoic far-field impulse response as a function of transmit angle. The loudspeaker impulse response is of length K and is modeled for NP distinct transmit directions. As explained in Section 2.3.

3. $m(t,n)$ - This multichannel filter is explained in Section 3.3. It exploits the symmetry in the uniform circular array response and its relative delays $\Delta d_k(\theta_r)$ for the kth microphone.

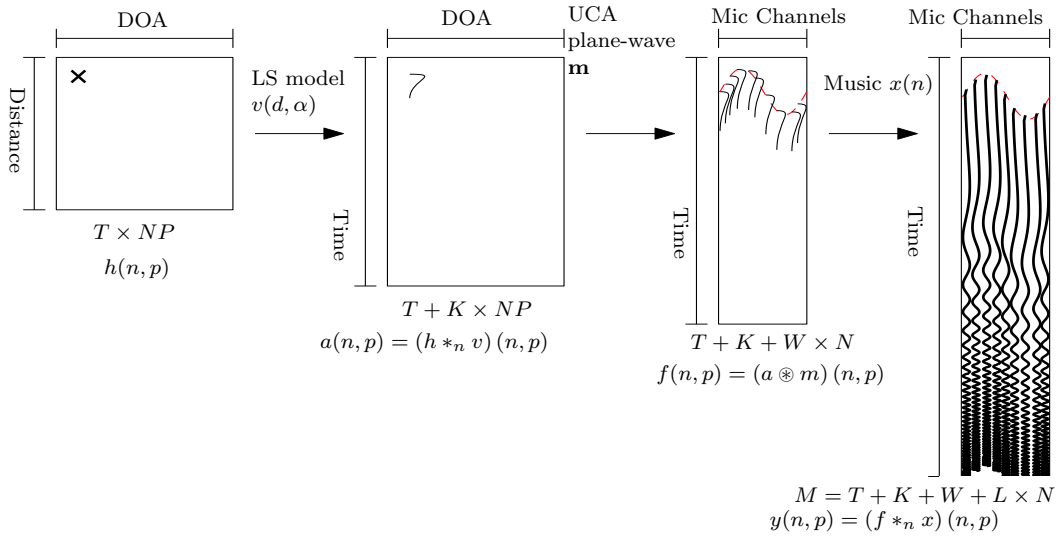
4. $x(t)$ - Acoustic excitation signal, can be a piece of music or a pilot signal. Sampled at f_s

In Fig. 3.5a a single image source is modeled. As one can see, the channel is composed in three sequential steps: The loudspeaker impulse response is added, the microphone plane wave response is added and finally (not shown in the figure) the direct path is added and its convolved with $x(t)$. The key observations are that i) as the candidate location moves further away from the system, the signal is translated in the time dimension and ii) if the source circles around the system then the loudspeaker impulse response changes and the array template mask permutes circularly. In the next section, Eq. (3.14) is reformulated in terms of matrix-vector multiplications.

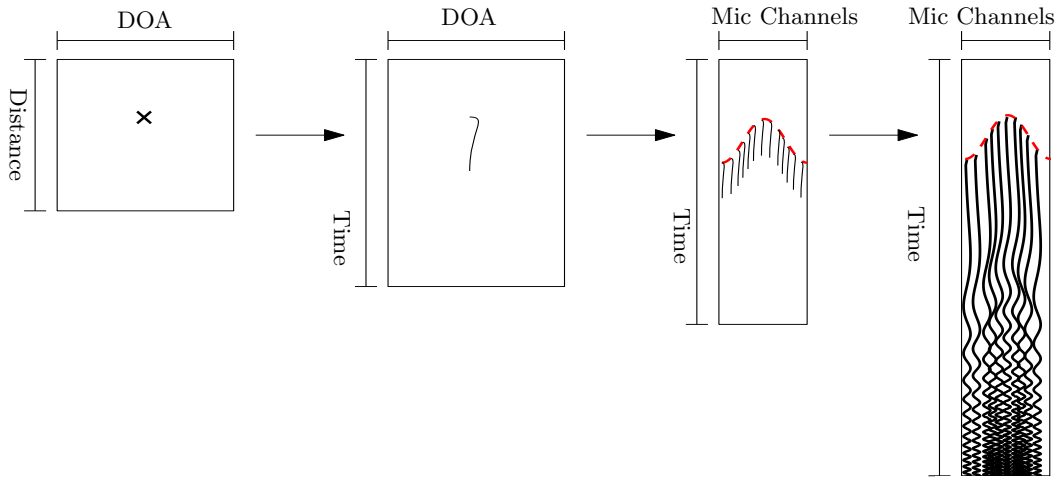
3.5 Constructing the forward model matrix

Let $\mathbf{\Phi}$ denote a matrix of size $MN \times NPT$, $\mathbf{h}$ the *input* and let $\mathbf{y}$ denote the *output* of the forward model. The forward model can then be written as:

$$\mathbf{y} = \mathbf{\Phi h}, \tag{3.15}$$



(a) Diagrams depicting Eq. (3.14) from right to left. It is interesting to assign physically motivated dimensions in each stage. The direct path is disregarded in the last step. Here $*_n$ denotes the convolution in n. When no subscript is provided, all dimensions are convolved.



(b) Second signal digram with a single source at the back of the loudspeaker. A different loudspeaker impulse response is used and the plane wave on the UCA is circularly shifted.

Figure 3.5: The forward signal model: Going through various stages to end up with microphone measurements that take into account the loudspeaker model, the microphone array geometry and the excitation signal.

where $\mathbf{y}$ is

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{N-1} \end{bmatrix} \in \mathbb{R}^{(T+K+L)N} \tag{3.16}$$

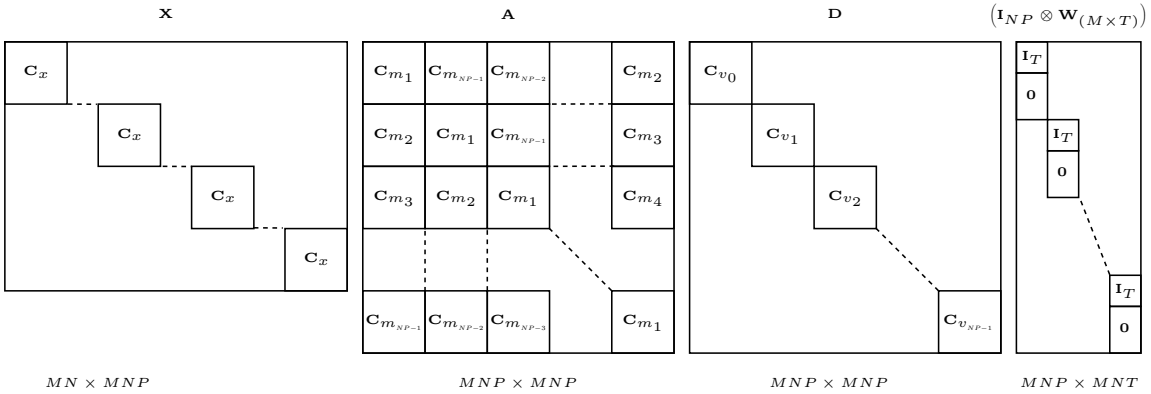


Figure 3.6: Visualization of the matrices that construct $\mathbf{\Phi}$ as defined in Eq. (3.25). The linear convolutions in time are computed by zero padding all impulse responses up to length M and performing a circular convolution instead (Circulant matrices denoted by $\mathbf{C}$, all have size $M \times M$ here).

$$\mathbf{y}_i = \big[y(0,i), y(1,i), \ldots, y(T+K+L-1,i)\big]^\intercal. \tag{3.17}$$

Remember that the size of $y(t,p)$ is determined by the number of microphones and the lengths of x, v and h which are L, K, and T respectively. The minimum samples in time for the microphone measurements is given by:

$$M = L + T + K. \tag{3.18}$$

In this section,the convolution in terms of summations in Eq. (3.14) is reformulated into this matrix-vector operations. The symbol $\otimes$ denotes the Kronecker product. The main idea is to formulate the linear convolutions as appropriately zero padded circular convolutions (see Appendix A about Toeplitz and Circulant matrices). For this, we introduce the generalized discrete windowing/zero padding matrix

$$\mathbf{W}_{(a\times b)} \triangleq \begin{cases} \begin{bmatrix} \mathbf{I}_{a\times a} & \mathbf{0}_{a\times b-a} \end{bmatrix} & \text{for } a < b \textit{ (Windowing)} \\ \\ \begin{bmatrix} \mathbf{I}_{b\times b} \\ \mathbf{0}_{b\times a-b} \end{bmatrix} & \text{for } a \geq b \textit{ (zero padding)} \end{cases}, \tag{3.19}$$

where we have that $\mathbf{W}_{(a\times b)} \in \{0,1\}^{a\times b}$ and $\mathbf{I}$ denotes the identity matrix.

Remember that the discrete signals $h(t,p)$, $x(t)$, $v(t,s)$ and $y(t,p)$ can be interpreted as column vectors. We have that $\mathbf{h}$ is the column vector defined at Eq. (3.13), $\mathbf{y}$ is defined at Eq. (3.16). $\mathbf{x}$, $\mathbf{m}$ and $\mathbf{v}$ are defined similarly. We have that

$$\mathbf{x} = \big[x(0), x(1), \ldots, x(L-1)\big]^\intercal \in \mathbb{R}^L \tag{3.20}$$

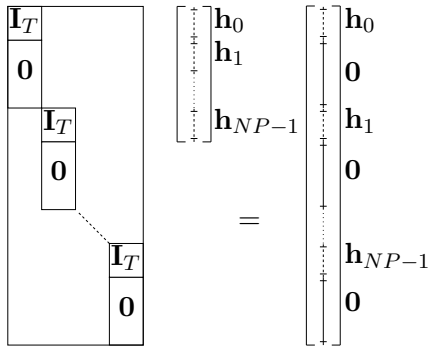


Figure 3.7: The right part of Eq. (3.25) is denoted here. Since $M >> T$ we have that $\mathbf{h}_{\text{zp}} = \left(\mathbf{I}_{NP} \otimes \mathbf{W}_{(M\times T)}\right)\mathbf{h}$ performs zero padding on $\mathbf{h}$ using Kronecker products. The result is a column vector of length MNP

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_{NP-1} \end{bmatrix} \in \mathbb{R}^{KNP}. \tag{3.21}$$

where,

$$\mathbf{v}_i = \left[v(0,i), v(1,i), \ldots, v(K-1,i)\right]^\intercal \in \mathbb{R}^K \tag{3.22}$$

Furthermore, we stack each column of the matrix $\mathbf{M}$ such that we obtain a column vector. The columns are given by

$$\mathbf{m}_i = \left[m_{1,i}, m_{2,i}, \ldots, m_{W,i}\right]^\intercal, \tag{3.23}$$

which are stacked on top of each other to obtain

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_{NP} \end{bmatrix} \in \mathbb{R}^{WNP}. \tag{3.24}$$

Observe how the convention in this thesis is to use zero-based numbering except when denoting entries in a matrix, like $\mathbf{M}$, where 1-based number is used instead.

The construction of $\mathbf{\Phi}$ uses $\mathbf{x}$, $\mathbf{v}$ and $\mathbf{m}$ and is given by

$$\mathbf{y} = \mathbf{\Phi h} = \mathbf{XAD}\left(\mathbf{I}_{NP} \otimes \mathbf{W}_{(M\times T)}\right)\mathbf{h}. \tag{3.25}$$

The size of $\mathbf{\Phi}$ is $MN \times NPT$. The components will be explained in the remainder of this Subsection. A visualization of Eq. (3.25) is given in Fig. 3.6.

First, the rightmost part of Eq. (3.25) is explained, where $\mathbf{h}$ is zero padded in the time/distance dimension up to length M. In Fig. 3.7 it is shown how multiplying $\left(\mathbf{I}_{NP} \otimes \mathbf{W}_{(M\times T)}\right)$ with $\mathbf{h}$ results in a zero padded version of $\mathbf{h}$, denoted by $\mathbf{h}_{\text{zp}}$. The Kronecker product is used here to structure the matrix to work with these *collapsed* vectors that represent two dimensions.

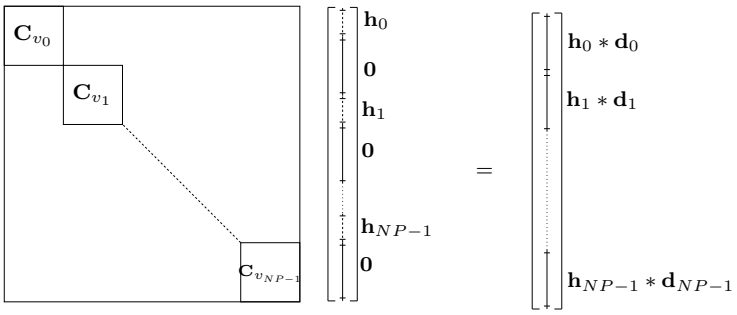


Figure 3.8: Since $\mathbf{h}$ is zero padded to length M, the loudspeaker impulse responses $\mathbf{v}_i$ are also zero padded to length M. Here $\mathbf{C}_{v_i}$ denote the Circulant matrix constructed from $\mathbf{v}_i$

Secondly, we have that $\mathbf{X}$ convolves all microphone channels with the excitation signal. Since we used a high resolution template mask for the microphone array, assuming NP microphones, the downsampling to N is done in this matrix. The convolution is repeated for all channels, which is achieved by the Kronecker multiplication with the identity matrix. Let circ(.) denote the circulant matrix operator, whose first row is given (this uniquely defines the matrix as explained in Appendix A). Alternatively, the eigenvalue decomposition using the Discrete Fourier Transform matrices, denoted by $\mathbf{F}$, can be used to define these circulant matrices. Where $\mathbf{F}^{\dagger} = \mathbf{F}^{-1}$ is a unitary matrix, as explained in Appendix A. We can express $\mathbf{X}$ as

$$\begin{aligned} \mathbf{X} &= \mathbf{I}_N \otimes \left[1, \mathbf{0}_{P-1}\right] \otimes \mathrm{circ}\{\mathbf{W}_{(M\times L)}\mathbf{x}\} \\ &= \left(\mathbf{I}_N \otimes \left[1, \mathbf{0}_{P-1}\right]\right) \otimes \left(\mathbf{F}_M^{-1}\mathbf{\Lambda}_x\mathbf{F}_M\right) \in \mathbb{R}^{MN\times MNP} \end{aligned} \tag{3.26}$$

where we have that

$$\mathbf{\Lambda}_x = \mathrm{diag}\left\{\mathbf{F}_M\mathbf{W}_{(M\times L)}\mathbf{x}\right\}, \tag{3.27}$$

where diag{.} is a square diagonal matrix operator with the input vector as the diagonal elements, if the input is of length M then the output is a diagonal matrix of size $M \times M$.

The matrix $\mathbf{D}$ that includes the loudspeaker model, is constructed similarly. The main difference is that in $\mathbf{v}$ many loudspeaker impulse responses are stacked. The convolution will be dependent on the direction of arrival of the wall echo, as a different loudspeaker impulse response must be used. Each loudspeaker impulse response $\mathbf{v}_i$ is first zero padded to length M. Using Kronecker products, the ith circulant matrix is multiplied with the ith direction of arrival. We have that

$$\mathbf{D} = \left(\mathbf{I}_N \otimes \mathbf{F}_M\right)^{-1} \mathbf{\Lambda}_V \left(\mathbf{I}_N \otimes \mathbf{F}_M\right), \tag{3.28}$$

where $\mathbf{\Lambda}_V$ is a diagonal matrix that is given by

$$\mathbf{\Lambda}_V = \mathrm{diag}\left\{\left(\mathbf{I}_{NP} \otimes \mathbf{F}_M\mathbf{W}_{(M\times K)}\right)\mathbf{v}\right\}. \tag{3.29}$$

The matrix-vector multiplication $\mathbf{D}\mathbf{h}_{\mathrm{zp}}$ is visualized in Fig. 3.8. The result of $\mathbf{D}\mathbf{h}_{\mathrm{zp}}$ is a column vector containing delayed loudspeaker impulse responses, where the direction of arrival in h determines the transmit angle of the loudspeaker model and the wall distance determines the delay.

Lastly the matrix that performs the two dimensional (circular) convolution with the template mask is constructed. Remember that the template mask is of size W in the *time* dimension and NP in the microphone dimension. The template mask is only zero padded in time, since the circular convolution is required in the microphone dimension. The matrix $\mathbf{A}$ is therefore constructed as

$$\mathbf{A} = \left(\mathbf{F}_N \otimes \mathbf{F}_M\right)^{-1} \mathbf{\Lambda_M} \left(\mathbf{F}_N \otimes \mathbf{F}_M\right), \tag{3.30}$$

where $\mathbf{\Lambda_M}$ is a diagonal matrix of size $MNP \times MNP$ and is given by

$$\mathbf{\Lambda_M} = \text{diag}\left\{\left(\mathbf{F}_{NP} \otimes \mathbf{F}_M\right)\left(\mathbf{I}_N \otimes \mathbf{W}_{(M\times T)}\right)\mathbf{m}\right\}. \tag{3.31}$$

Notice how $\mathbf{m}$ is zero padded with $M-T$ zeros in the time dimension, but no zeros are added in the microphone observation dimension.

Now that the measurement model has been reformulated in in matrix-vector notation, the inverse problem can be expressed as a convex optimization problem in standard form. This is done in the next section.

3.6 Solving the inverse problem

Thus far in this chapter, the measurement model is explained. The far field assumptions have let to a linearized measurement model that is expressed as

$$\mathbf{y} = \mathbf{\Phi h} + \mathbf{n}. \tag{3.32}$$

Where we have that $\mathbf{y} \in \mathbb{R}^{MN}$, $\mathbf{\Phi} \in \mathbb{R}^{MN\times NPT}$ and both $\mathbf{n}$ and $\mathbf{h}$ are of length NPT. Here $\mathbf{n}$ denotes the noise. Note how $\mathbf{\Phi}$ is an overdetermined set of linear equations.

In this section, it is explained how to solve the inverse problem, ie. how to estimate $\mathbf{h}$ from noisy observations of $\mathbf{y}$. In literature, this problem is usually referred to as deconvolution. Typical in those problems, is that the linear system is a convolution-type kernel. This means that the kernel only depends on the difference between two independent variables [37]. If an omni-directional loudspeaker model is assumed, i.e. $\mathbf{v}_i\mathbf{v}_j \forall i,j$, this property holds. However, a loudspeaker impulse response that depends on the wall echo direction of arrival breaks this property. Such deconvolution problems can be solved in a robust way by using regularization.

When $\mathbf{\Phi}$ is well conditioned and the noise is independent and identically distributed Gaussian, the solution can be found by minimizing the quadratic loss between the measurements and the model's prediction $\|\mathbf{y} - \mathbf{\Phi h}\|_2^2$. However, $\mathbf{\Phi}$ is ill-conditioned and when dealing with real-world data, the presence of small reflecting surfaces, correlated noise sources and model mismatches must be accounted for.

One approach to deconvolve the loudspeaker impulse response from the measured signal in literature is the use of the matched filter (MF). The measurements are convolved with a conjugated time reversed version of the loudspeaker impulse response. As an example, an omni-directional assuming loudspeaker model with matched filter loudspeaker decorrelation is proposed in [31]. The decorrelation with the matched filter is achieved by taking the Hermitian of the Circulant or Toeplitz matrix. Multiplying

with $\mathbf{D}^\dagger$ and $\mathbf{X}^\dagger$ therefore correlates the measurements with these templates. The inverse does not need to be computed for these large matrices, as it is approximated by its transpose. Similarly, the delay-and-sum steered response power can be evaluated by multiplying with the Hermitian of $\mathbf{A}$. The gain on candidate locations can therefor be estimated by assuming $\mathbf{\Phi}^\dagger\mathbf{\Phi} \approx \mathbf{I}$, which is equivalent to matched filtered deconvolution and delay-and-sum (DAS) steered response power (SRP) beamforming. Therefore the first proposed method to solve the inverse problem is the matched filter with a delay-and-sum (MF-DAS) response given by

$$\hat{\mathbf{h}}_{\text{MF-DAS}} = \mathbf{\Phi}^\dagger\mathbf{y}. \tag{3.33}$$

The second approach is computationally more intensive. It is inspired by high resolution techniques in acoustic source localization. The main idea is to solve a constrained optimization problem. As explained in Section 2.1, in room acoustics the idea to exploit the sparsity of the arriving echoes is not new. The main idea is to fit measurements, whilst constraining the necessary image source locations to be sparse. To measure the sparsity of the solution, the ℓ_0 norm can be used. The optimization problem can be constrained to have S contributing image sources, by solving the following constrained optimization problem

$$\begin{aligned} \underset{\mathbf{h}}{\text{minimize}} \quad & \|\mathbf{y} - \mathbf{\Phi}\mathbf{h}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{h}\|_0 \leq S. \end{aligned} \tag{3.34}$$

However, this leads to a non-convex optimization problem, that is NP-hard [38], so what is often done in literature is to relax this to the ℓ_1 norm. The second estimator $\hat{\mathbf{h}}_{\text{sparse}}$ can be found by solving the following optimization problem

$$\begin{aligned} \underset{\mathbf{h}}{\text{minimize}} \quad & \|\mathbf{y} - \mathbf{\Phi}\mathbf{h}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{h}\|_1 \leq \beta. \end{aligned} \tag{3.35}$$

In statistics and machine learning, this optimization problem is called *least absolute shrinkage and selection operator* (LASSO). β is the regularization term. Increasing this value, yields a less sparse solution, that may fit better with measurements. A smaller β will force the solution to be more sparse, at the cost of a worse fit.

It must be noted that Eq. (3.35) is the standard Lasso formulation, however most solvers consider the so-called Lagrangian form

$$\hat{\mathbf{h}}_{\text{sparse}} = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{h}\|_2^2 + \lambda \|\mathbf{h}\|_1 , \tag{3.36}$$

where the exact relationship between β and λ is data dependent. The standard Lasso optimization problem works well, when the received echo power from each wall of interest is approximately equal. However, if the loudspeaker is placed in the corner of a room, it is expected that the first echoes have higher power, compared to the later echoes from further walls. This is accounted for in the forward measurement model by the gains of the active image sources in $\mathbf{h}$. A second problem is that, for many loudspeakers, the loudspeaker impulse response total energy varies with the angle of

transmission. This influences the signal to noise ratio of the detection problem. It is expected that an echo from a nearby wall, in the on-axis direction of the loudspeaker has a higher influence on the microphone measurements compared to a wall facing the back of the loudspeaker that is placed further away.

We have that the expected value for $\mathbf{h}$ is dependent on the distance of the wall and the DOA and that the energy in $\mathbf{y}$ decays over time. Both influences can be compensated for by having a weighted least squared and a weighted ℓ_1 norm. Let $\mathbf{\Lambda}_{\text{ls}} \in \mathbb{R}^{MN \times MN}$ denote a diagonal weighting matrix for the total least squares and let $\mathbf{\Lambda}_h$ denote a diagonal weighting matrix for the gain on the candidate locations. The general optimization problem is then given by

$$\hat{\mathbf{h}}_{\text{sparse}} = \arg\min_{\mathbf{h}} \left(\mathbf{y} - \mathbf{\Phi h}\right)^{\dagger} \mathbf{\Lambda}_{\text{ls}} \left(\mathbf{y} - \mathbf{\Phi h}\right) + \left\|\mathbf{\Lambda}_h \mathbf{h}\right\|_1 . \tag{3.37}$$

In the next section, the proximal gradient method is explained. This method can be used, to find a solution to the optimization problem.

3.6.1 Proximal Gradient Methods

Our general optimization problem stated in Eq. (3.37) does not have a differentiable objective function due to the ℓ_l norm. One solution to overcome this problem is to use subgradients instead. However, the convergence rate of subgradient based descent methods are slow. In this section the Proximal Gradient Descent Method and the Accelerated Proximal Gradient method are explained [39]. These methods have a faster convergence rate than subgradient descent.

Observe that our objective function $f(\mathbf{h})$ can be decomposed into two functions

$$f(\mathbf{h}) = e(\mathbf{h}) + g(\mathbf{h}), \tag{3.38}$$

where $e(\mathbf{h}) = \left(\mathbf{y} - \mathbf{\Phi h}\right)^{\dagger} \mathbf{\Lambda}_{\text{ls}} \left(\mathbf{y} - \mathbf{\Phi h}\right)$ is the total least squares function and $g(\mathbf{h}) = \left\|\mathbf{\Lambda}_h \mathbf{h}\right\|_1$ is the sparsity enforcing penalization function. The first function is convex and differentiable, whereas the second is convex and not differentiable. Let $\mathbf{h}_{k+1}$ denote the updated estimate for $\mathbf{h}$ at iteration $k+1$, the proximal gradient method is given by

$$\begin{aligned} \mathbf{h}_{k+1} =& \arg\min_{\mathbf{u}} \left(g(\mathbf{u}) + e(\mathbf{h}_k) + \nabla e(\mathbf{h}_k)^{\dagger} \left(\mathbf{u} - \mathbf{h}_k\right) + \frac{1}{2t} \left\|\mathbf{u} - \mathbf{h}_k\right\|_2^2 \right) \\ =& \arg\min_{\mathbf{u}} \left(g(\mathbf{u}) + \frac{1}{2t} \left\|\mathbf{u} - \mathbf{h}_k + t\nabla e(\mathbf{h}_k)\right\|_2^2 + \right), \end{aligned} \tag{3.39}$$

where $t > 0$ is the step size. At this point it may seem that one optimization problem has been replaced by another. Because as states in Eq. (3.39), for each iteration a different optimization problem must be solved to compute the update step. Fortunately, the minimization in Eq. (3.39) can be computed analytically for some simple non-differentiable functions $g(\mathbf{h})$. In general the proximal mapping is defined as

$$\text{prox}_{t,g}(\mathbf{h}) = \arg\min_{\mathbf{u}} \left(g(\mathbf{u}) + \frac{1}{2t} \|\mathbf{u} - \mathbf{h}\|^2 \right). \tag{3.40}$$

Note how the proximal mapping doesn't depend on $\mathbf{u}$, but is determined by the non-differentiable function g and the stepsize.

For the weighted ℓ_1-norm, i.e. $g(\mathbf{h}) = \|\mathbf{\Lambda}_h \mathbf{h}\|_1$, the proximal mapping can be computed analytically. The solution is given by the *Soft-thresholding operator*. This function is computed easily. It is given as:

$$[S_{\mathbf{\Lambda}}(\mathbf{h})]_i = \begin{cases} h_i - \Lambda_i & \text{if } h_i > \Lambda_i \\ 0 & \text{if } -\Lambda_i \leq y \leq \Lambda_i \\ h_i + \Lambda_i & \text{if } i \leq -\Lambda_i \end{cases}, \tag{3.41}$$

where Λ_i denotes the ith diagonal entry of $\mathbf{\Lambda}_h$.

The *Iterative soft-thresholding algorithm* (ISTA) is a simple algorithm that can calculate the lasso solution. The update function is by combining Eq. (3.39) with Eq. (3.41):

$$\begin{aligned} \mathbf{h}_{k+1} =& S_{\mathbf{\Lambda}} \left(\mathbf{h}_k - t \nabla e(\mathbf{h}) \right) \\ =& S_{\mathbf{\Lambda}} \left(\mathbf{h}_k - t \mathbf{\Phi}^\dagger \mathbf{\Lambda}_{\text{ls}} \left(\mathbf{y} - \mathbf{\Phi} \mathbf{h} \right) \right). \end{aligned} \tag{3.42}$$

This method can be accelerated with Nestrov's idea to use *momentum* [40]. Fast ISTA (FISTA) does not evaluate the proximal map at $\mathbf{h}_{k-1}$ but adds a momentum term. This allows for some of the history to be exploited for a faster convergence. However, FISTA is not guaranteed to decrease the objective function at each step.

The update step for FISTA are given as

$$\mathbf{v} = \mathbf{h}_{(k)} + \frac{k-2}{k+1} \left(\mathbf{h}^{(k)} - \mathbf{h}^{(k-1)} \right) \tag{3.43}$$

$$\mathbf{h}_{k+1} = S_{\lambda t_k} \left(\mathbf{v} + t_k \mathbf{\Phi}^\dagger (\mathbf{y} - \mathbf{\Phi} \mathbf{v}) \right). \tag{3.44}$$

Observe how the update at step $k+1$ is dependent on the previous solution k and the one before that $k-1$.

4 Results

This chapter presents four experiments to show the performance of the proposed method and to compare it to state of the art techniques. The first experiment assumes that an omnidirectional loudspeaker is placed near a single wall. The experiment compares the time-of-arrival estimates of the reflections for methods that i) have no microphone geometry prior with ii) methods that exploit the array geometry to only estimate a consistent set of TOA's for all microphones. In experiment B an omnidirectional loudspeaker is placed inside a rectangular room. It is shown that methods based on steered response power maximization can not resolve closely sepatated reflections, especially for mid-range loudspeakers with limited bandwidth. The third experiment is based on a typical stereo loudspeaker model, that has the inherent directivity. Here it is shown that methods that assume an omnidirectional loudspeaker will suffer from model miss matches. The chapter concludes with a preliminary investigation using real world measurements where the directivity model proposed is compared to measurements.

4.1 Experiment A: Single wall with omnidirectional loudspeaker

The purpose of this isolated simulation experiment is to show that single channel time of arrival estimation is less robust compared to TOA methods that use the microphone array geometry as prior. For this, a simple single wall setup is used. The static measurement $y(n, k)$ is generated using the signal model presented in the previous chapter. The signal contains a single Rotated Image Source Impulse Response (RISIR) at some arbitrary position. The goal is to detect this source at increasingly challenging Single to-Noise Ratio (SNR). To save time in computations, the signal model is reduced. The exponential sine sweep is disregarded in the signal model. It is assumes that the excitation signal is deconvolved perfectly. The microphone measurements y consist of a delayed version of the loudspeaker impulse response.

in Fig. 4.2 the static microphone signal is depicted. Here one can observe the loudspeaker impulse response, delayed by the microphone array response, for each microphone channel. It is assumed that the direct path has been removed. As the noise increases, the probability of detecting the image source decreases.

Two groups of methods are compared, as summarized in Fig. 4.1 we have:

- Two steps methods: First estimate the Time Of Arrival (TOA) using the loudspeaker impulse response, then use the known microphone locations to geometrically infer the source location. One such example of a single channel TOA estimation methods is using the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [41]. The times-of-arrival (TOA) are then combined with the

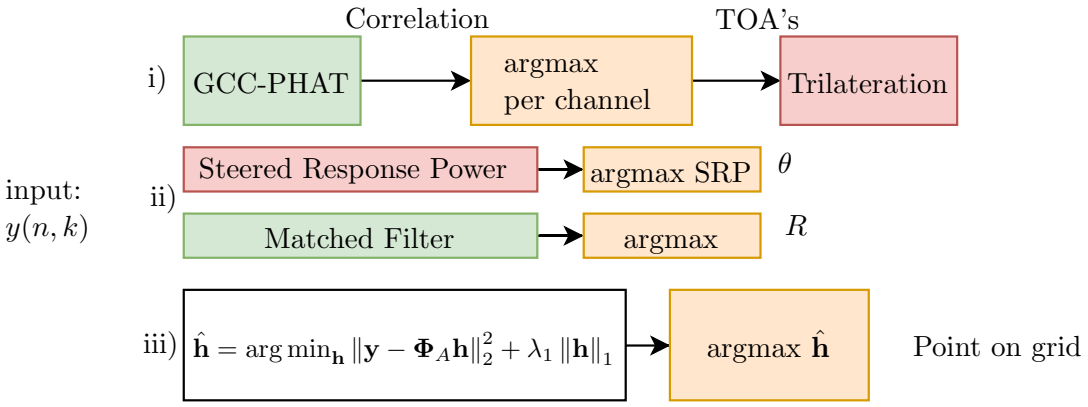
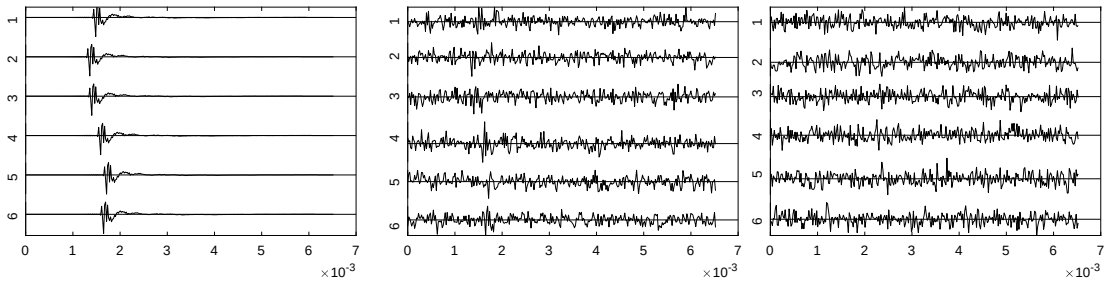


Figure 4.1: Experiment A: Three methods to compare. In green are the steps that make use of the loudspeaker impulse response. Red are the steps that use the microphone geometry and orange uses the single source assumption. The optimization problem in iii makes use of both the loudspeaker impulse response as well as the microphone array in a single step.



(a) Noiseless. (b) SNR=-9 (c) SNR=-20

Figure 4.2: Experiment A: $y(n,k)$ used, where a single wall reflection is present. Each figure has six microphone channels.

known array geometry to compute the image source location using least squares trilateration [42]. It must be noted that the microphone array geometry is not used until after the TOA's have been estimated.

- Single step method: Combine the Time Of Arrival estimation for distance, with the Time Difference Of Arrival for Direction Of Arrival (DOA) estimation. There are numerous multichannel methods, one such class relies on beamforming techniques. The idea is to use a beamformer to steer the beam into a particular direction. The *Steered Response Power* (SRP) is then evaluated for NP uniformly spaced directions of arrival. The direction of arrival is estimated as the angle for which the SRP is highest.

In each Monte Carlo loop, a new realization of noise is added to $\mathbf{y}$, after which the noisy observation is run through the three methods. The resulting three estimated locations are compared to the true wall location. The performance is measured with a binary feature, the hitrate, where a hit is considered if the estimated location is within the Voronoi region of the true location on the already defined polar grid. It must be noted that all three methods assume that there is a single source to detect, therefore

Table 4.1: Experiment A: The signal model assumes the excitation signal $x(t)$ to be perfectly removed and assumes the loudspeaker is omni-directional. Method i is a single channel method whereas ii and iii use the array geometry.

Signal Model	$\mathbf{\Phi}_A = \mathbf{AD}\left(\mathbf{I}_{NP} \otimes \mathbf{W}_{(M\times T)}\right)\mathbf{h}$ with loudspeaker response $v(t)$ being the same for all transmit directions
Room geometry	Single wall: h is zero everywhere except at $R = 30c/f_s$ and $\theta = \frac{6}{NP}2\pi$, i.e. $h((6-1)T+30) = 1$

Monte Carlo setup	$\mathbf{h} \leftarrow$ one source at $R = 0.2125$m and $\theta = \pi/2$
	$\mathbf{y} \leftarrow \mathbf{\Phi}_A\mathbf{h}$ ▷ Using omnidirectional model
	for SNR **do** ▷ Vary the noise power
	  for repeats **do** ▷ Repeat for different realizations of noise
	  $\mathbf{y}_{\text{noise}} \leftarrow \mathbf{y} + \mathbf{n}$ ▷ New realizatation of noise at SNR
	  $\tau_i \leftarrow \text{GCCphat}(\mathbf{y}_{\text{noisy},i})$ ▷ Repeated for all channels i
	  $\hat{R}, \hat{\theta} \leftarrow$trilateration$(\boldsymbol{\tau})$ ▷ Trilateration returns polar coordinate
	  $\hat{\mathbf{h}}_{\text{MF-DAS}} \leftarrow \mathbf{\Phi}^\dagger\mathbf{y}_{\text{noise}}$
	  $\hat{\mathbf{h}}_{\text{sparse}} \leftarrow \arg\min_{\mathbf{h}} \|\mathbf{y}_{\text{noise}} - \mathbf{\Phi h}\|_2^2 + \lambda\|\mathbf{h}\|_1$ ▷ Solved using FISTA
	  $\hat{R}, \hat{\theta} \leftarrow \arg\max_{R,\theta} h_*(\hat{R}, \theta)$ ▷ single source prior. Repeated for both $\hat{\mathbf{h}}$
	  hitrate $\leftarrow$ compare$(\hat{R}, \hat{\theta}, R, \theta)$ ▷ Rounded to nearest polar grid point
	  end for
	end for

Performance metric	All estimated locations are rounded to the nearest grid point. Hitrate: If estimated point corresponds to source location then 1, otherwise 0
Hypothesis	Time-of-Arrival estimation that uses array geometry is more robust
Variables and size	$N = 6$, $P = 1$, $T = 64$, $K = 250$, $M = 313$, $r = 0.06$m, $f_s = 48$kHz, $S = 1$, SNR$= [0, -30]$db

the detection problem in the peak-picking is reduced to extracting the highest value. A summary of experiment A is given in Table 4.1.

The results of the Monte Carlo simulation are shown in Fig. 4.3. Here it is shown that for a SNR larger than -7, all methods perfectly resolve the image source. The single channel method starts to miss the estimated location for some noise realizations at -7 db. The second and third method are more robust and succesfully resolve the image source reliably until -10 db. In experiment A however, the difference between maximizing the beamformers response and the proposed sparse reconstruction is not shown. The difference only becomes clear once the loudspeaker system is placed in an environment with multiple walls. This is shown in experiment B.

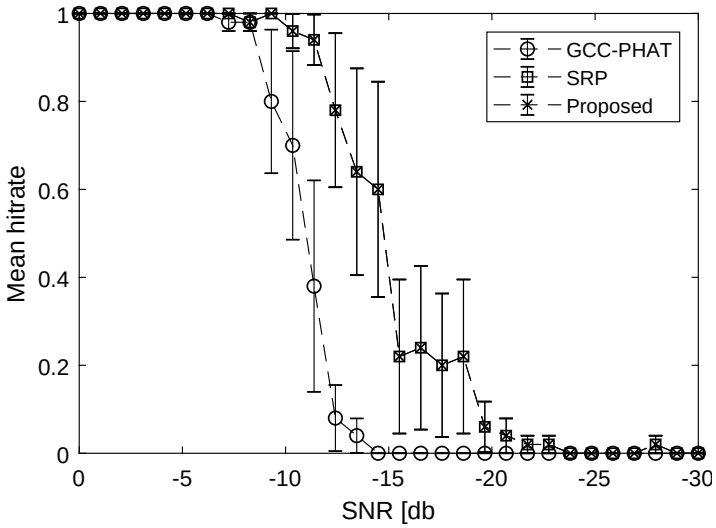
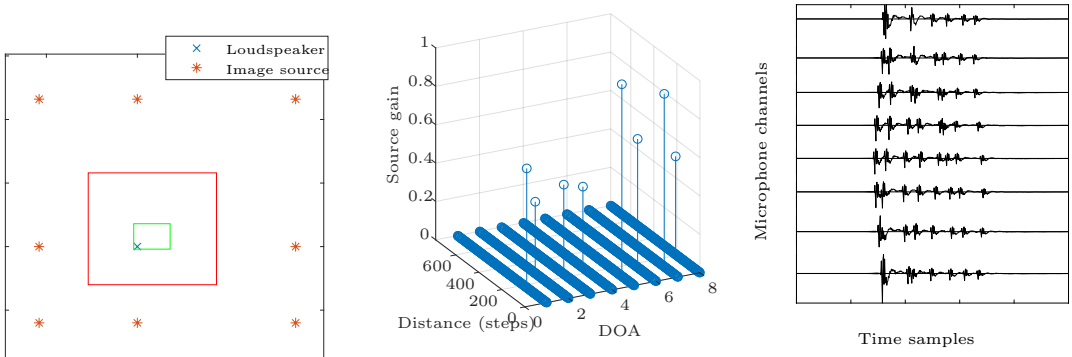


Figure 4.3: Experiment A: Mean hitrate (with standard deviation errorbar) depicted for the three methods for decreasing SNR. The mean is computed over 100 realizations of the noise. The second and third method have equal mean hitrate.



(a) Top view of a room. The red rectangle represents the walls of the room. The loudspeaker is placed within the green rectangle.

(b) The image source locations are snapped to the grid. $h(k,l)$ is generated for this room. There are 8 nonzero values, where the gain is determined by the distance.

(c) $y(n,k)$ that is generated for this room. One can see that all loudspeaker impulse responses are the same, i.e. an omnidirectional loud speaker model is used.

Figure 4.4: Experiment B: The loudspeaker is placed in a rectangular room, the locations of the image sources are embedded in $h(k,l)$ and are used to generate microphone measurements $y(n,k)$. The convolution with the excitation is disregarded.

4.2 Experiment B: Omnidirectional loudspeaker in rectangular room

In this experiment the same signal model is used as in experiment A. The difference is that instead of a single wall, a shoebox shaped room is modeled. All first and second order reflections (in two dimensions) are taken into account when generating $\mathbf{h}$,

resulting in 8 distinct reflections from image sources. A summary of this experiment can be found in Table 4.2.

It is expected that method ii and iii will yield similar results under i.i.d. white Gaussian noise, as long as the echoes do not overlap in time. To show the difference between the two methods, one of three things can be done: i) realistic impulsive/correlated noise that is to be expected as interference in the real world can be added, ii) the most challenging room geometries can be tested, or iii) the loudspeaker bandwidth can be altered.

In this Monte Carlo results, the noise has been limited to additive white Gaussian noise, however by placing the loudspeaker close to the center of the room, it is expected that the echoes will overlap and be challenging to separate. The experiment is first performed using a full range loudspeaker. In a second experiment, the bandwidth is reduced to have a second order low pass cut-off frequency at 5kHz.

A small shoebox shaped room is generated. The size of this room is 1m by 1.25m. In each loop, the loudspeaker system is moved around the room center, to a different location. This region is depicted in green in Fig. 4.4a. The image source location are computed in simulation and snapped to the nearest point on the polar grid (Fig. 4.4b). If the distance to these image sources is approximately equal, then the echoes should overlap in time (Fig. 4.4c). To best show the limitations of MF-DAS, the loudspeaker is moved around the room, as some wall geometries are more challenging than others. The experiment is also repeated for different realizations of noise for each position in the green rectangle. The noise statistics are constant, a SNR of 0db is used throughout the experiments.

To compensate for the decreasing power of the echoes, it is proposed to solve the weighted least squares problem, as shown in Table 4.2. Here $\mathbf{\Lambda}_{\text{ls}}$ scales the optimization problem such that far away faint image sources are normalized when compared to a close wall that echoes back higher power. Since the pressure is inversely proportional to the distance (which is related to how $h(n,p)$ is defined, see Eq. (3.11)), the proposed weighting is given by:

$$\mathbf{\Lambda}_{\text{ls}} = \text{diag}\left\{\left[1, 2, \ldots, M-1\right]\right\} \otimes \mathbf{I}_N \in \mathbb{R}^{MN \times MN}. \tag{4.1}$$

The results of the first experiment are shown in Fig. 4.5 and the repeated results using a band limited loudspeaker are depicted in Fig. 4.6. In the first experiment, the average hit rate over all simulations is 88% and 92% for the MF-DAS and sparse method respectively. When using the mid range loudspeaker the difference is larger, as MF-DAS has an average hit rate of 36% and the sparse solution 59%. This shows that at limited bandwidth the time-smearing caused by the loudspeaker significantly reduces the ability to resolve distinct reflections. Nevertheless, solving the inverse using a sparse prior improves the robustness significantly.

One can also observe that the performance is dependent on the loudspeaker positioning in the room. In particular, one can see in Fig. 4.6a that if the loudspeaker is placed equidistantly to two walls, that the hitrate is lower. This is to be expected, as two reflections overlapping in time are harder to resolve accurately.

The MF-DAS method takes a few milliseconds to run whilst the sparse optimization convergence takes anywhere from one up to four seconds. The same sparsity inducing

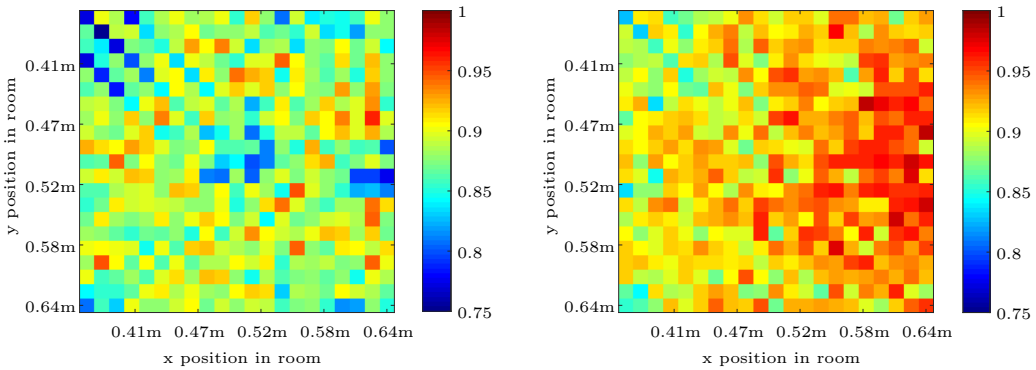
Table 4.2: Experiment B: The signal model assumes a known sinesweep $x(t)$ and assumes the loudspeaker is omni-directional. Method i) has no sparsity prior whereas ii) is a high resolution technique that seeks a sparse solution

Signal Model	$\mathbf{\Phi}_B = \mathbf{AD}\left(\mathbf{I}_{NP} \otimes \mathbf{W}_{(M\times T)}\right)\mathbf{h}$ with loudspeaker response $v(t,n)$ being the same for all transmit directions
Room geometry	A static room of 1m by 1.25m is used. The loudspeaker is placed in the center of the room and is moved around (Fig. 4.4a)

Monte Carlo setup	**for** l **do**oudspeakermodel ▷ Two loudspeakers: Full range, mid range
	for x in grid **do**
	for y in grid **do** ▷ Move loudspeaker on cartesian grid
	for repeats **do** ▷ Repeat for different realizations of noise
	$\{R,\theta\}_i \leftarrow$ getImageSourceLocations(x,y) ▷ Get 8 image sources
	$\mathbf{h} \leftarrow$ constructGainVector$(\mathbf{R},\boldsymbol{\theta})$ ▷ using Eq. (3.11)
	$\mathbf{y} \leftarrow \mathbf{\Phi h} + \mathbf{n}$ ▷ Using Eq. (3.25). New noise realization
	$\hat{\mathbf{h}}_{\text{MF-DAS}} \leftarrow \mathbf{\Phi}^{\dagger}\mathbf{\Lambda}_{\text{ls}}\mathbf{y}$
	$\hat{\mathbf{h}}_{\text{sparse}} \leftarrow \arg\min_{\mathbf{h}} \left(\mathbf{y}-\mathbf{\Phi h}\right)^{\dagger}\mathbf{\Lambda}_{\text{ls}}\left(\mathbf{y}-\mathbf{\Phi h}\right) + \lambda\left\|\mathbf{h}\right\|_1$ ▷ Solved using FISTA
	$\hat{R},\hat{\theta} \leftarrow \text{maxk}\left(\hat{h}(R,\theta),8\right)$ ▷ Knowledge on 8 sources used here
	hitrate $\leftarrow$ compare$(\hat{R},\hat{\theta},R,\theta)$ ▷ Rounded to nearest polar grid point
	end for
	end for
	end for
	end for

Performance metric	All estimated locations are rounded to the nearest grid point. Hitrate: If estimated point corresponds to source location then 1, otherwise 0
Hypothesis	Matched Filter steered response maximization will suffer from overlapping reflections and reduced loudspeaker bandwidth. Using a sparse prior can increase robustness to solve the inverse problem.
Variables and size	$N=8$, $P=1$, $T=774$, $M=1025$ $K=250$. Room size $[1, 1.25m]$. Loudspeaker is moved around in a grid. $\lambda=68$, $x=y=[0.35, 0.64]m$ in 21 steps. Repeats = 35. Static SNR = 0db.

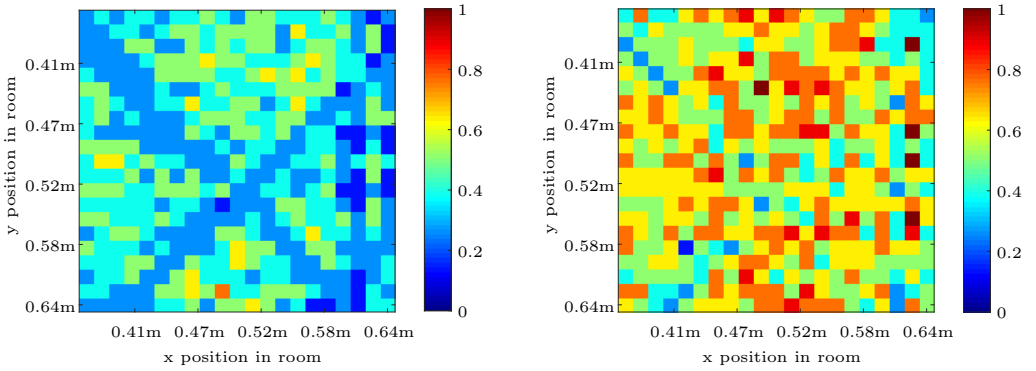
parameter λ is used for all locations. The runtime for each experiment was about 12 hours on a high end consumer laptop using Matlab. However it must be noted that the performance of the sparse optimization is heavily dependent on the correct estimation of this parameter. A suboptimal λ will generally result in much worse performance compared to MF-DAS that has no parameter to set. Conversely, the results on the sparse optimization presented here are likely to be sub optimal because the λ was picked heuristically to work well on average for all locations.



(a) Mean hitrate for MF-DAS. Average over all locations: 88%

(b) Mean hitrate using sparse optimization. Average over all locations: 92%

Figure 4.5: Experiment B results: Mean hitrate for locations within the green rectangle of Fig. 4.4a using full range loudspeaker



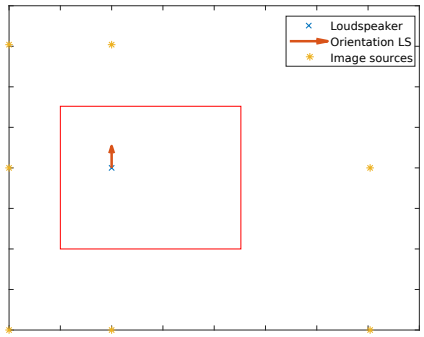
(a) Mean hitrate for MF-DAS. Average over all locations: 36%

(b) Mean hitrate using sparse optimization. Average over all locations: 59%

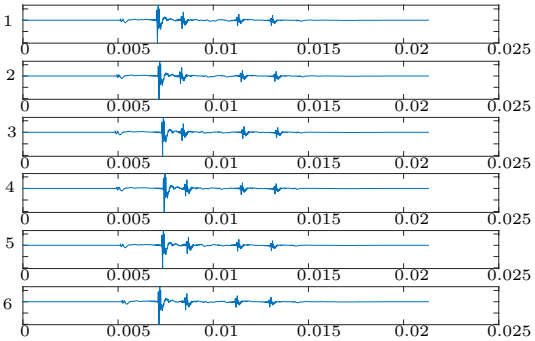
Figure 4.6: Experiment B repeated for an omnidirectional loudspeaker limited to 5kHz.

4.3 Experiment C: Omni assuming versus directivity aware models

At this point, it has been shown that solving multiple steps jointly can improve robustness (experiment A) and that solving the inverse problem can benefit from a sparse prior (experiment B). In this simulation experiment, we model a Genelec 1029A studio monitor that has been placed inside a shoebox shaped room. The goal here is to compare the omni assuming loudspeaker measurement model with the directivity aware model. The omni assuming approach can be considered the current state of the art, as all literature assumes an omnidirectional loudspeaker when localizing reflections from measured room impulse responses. The sole exception is the approach presented in [28] as was mentioned in Section 2.1. However that method relies on hundreds of special anechoic measurements. Due to a lack of resources and time, those results are not



(a) The Genelec loudspeaker is placed in a room. The 8 image sources are generated.



(b) Simulated channel response for each microphone in the array. The 8 reflections are not all distinguishable. The shape of the reflected signal is different and based on the loudspeaker directivity model.

Figure 4.7: Experiment C: An example of a room that is generated in which a loudspeaker is placed with known directivity model.

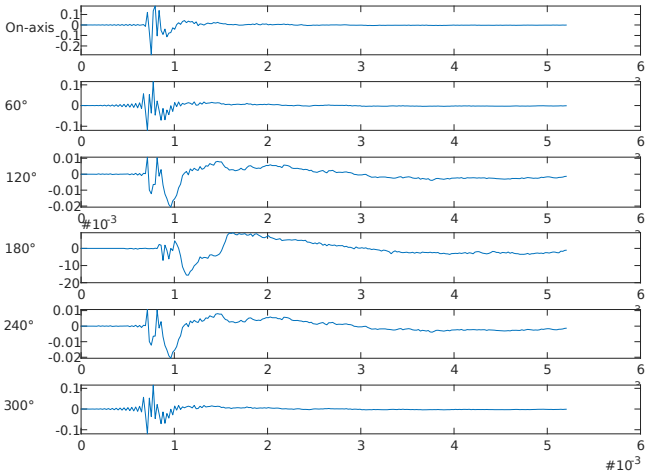


Figure 4.8: The loudspeaker impulse response of the Genelec 1029A, measured at 3 meter distance. The response is given for six uniformly spaced angles.

reproduced.

The loudspeaker was measured from many different angles in anechoic conditions at the Bang & Olufsen facilities in Struer, Denmark (Fig. 4.10a). The loudspeaker was placed on a crane in the center of a large empty space. By rotating the loudspeaker, the impulse response is measured at 3 meters away for many transmit directions (in the horizontal plane). The resulting directivity is visualized in Fig. 4.10c. One can observe here that this loudspeaker is highly directional and that the bandwidth is maximum in the on-axis direction. This is typical for a single (or dual) loudspeaker system as the high frequencies tend to leave the speaker in narrow beams.

The directivity of the loudspeaker impulse response can also be interpreted in the time domain. In Fig. 4.8 the loudspeaker impulse response is given for six uniform angular steps. It shows that maximum pressure difference is a function of this angle. Due to the reduced bandwidth in the 180° loudspeaker impulse response, it is expected that resolving reflections from that direction will suffer from time-smearing.

The Monte Carlo simulation is setup similarly as with the previous experiment. A loudspeaker is placed in a shoebox shaped room of size 2.5m by 2.5m and is moved around the center. The image source locations are generated the same way. The difference is the signal model matrix $\mathbf{\Phi}_{\text{Genelec}}$ that is used to generate the measurements. This has used the directivity measurements from Fig. 4.10c to generate $\mathbf{D}$. The upsampling factor P is set to one. Therefore the measurement model is using the six loudspeaker impulse responses from Fig. 4.8. In contrast with experiment B, the loudspeaker is not only moved but is also rotated in 60° degree intervals. An example of a channel response is shown in Fig. 4.7b. Here one can see that the front reflections are dominating the channel response. The much fainter reflections from the back of the loudspeaker can easily be obscured by these stronger reflections. A summary of the experiment is provided in Table 4.3.

The results show much lower hit rates compared to experiment B. This is due to the reduced power in some of the reflections. The hit rate averaged over all room positions and orientations is lowest for the omnidirectional assuming methods. Both MF-DAS and the sparse optimization using $\mathbf{\Phi}_{\text{omni}}$ have 15% mean hit rate. Essentially only reliably resolving the reflection from the wall facing the front of the loudspeaker. The models that use the directivity model have a mean hitrate of 41% and 53% for the MF-DAS and sparse estimation respectively. In Fig. 4.9 the mean hitrates are presented for each direction of arrival of the reflector. The results show that resolving the reflections arriving from 180° are most challenging.

4.4 Experiment D: Real world measurements

In this experiment, real world measurements are compared with the proposed signal model. They are compared to the conventional model that assumes an omnidirectional loudspeaker. The measurements are performed using the same Genelec studio monitor.

An uniform circular microphone array is placed on top of this loudspeaker, as seen in Fig. 4.10b. The direct path is measured in anechoic conditions. The loudspeaker was placed in front of a single large wall at 2 meter distance. In this experiment, the loudspeaker is rotated with respect to the wall to show that the proposed forward model that takes into consideration the directivity model of the loudspeaker can better predict microphone measurements compared to omnidirectional assuming models.

Until thus far, it is assumed that the loudspeaker is placed in a two dimensional room, i.e. a room limited to vertically reflecting surfaces. In practical scenarios, it is expected to also receive reflections from the floor and ceiling. Rather than solving the inverse problem directly using raw measurements, in this experiment the focus is laid on comparing the two measurement models.

Since we have six microphones in the array, it was chosen to rotate the system 60° for each measurement. The first microphone is the microphone that is aligned for

Table 4.3: Experiment C: The Genelec 1029A loudspeaker is placed in a shoebox room. The signal model utilizes the measured directivity. Two classes of methods are compared: Those that assume an omnidirectional loudspeaker and those aware of the directivity.

Signal Model	$\mathbf{\Phi}_{\text{genelec}} = \mathbf{AD}\left(\mathbf{I}_{NP} \otimes \mathbf{W}_{(M\times T)}\right)\mathbf{h}$ with loudspeaker response $v(t,n)$ from measurements of Genelec 1029A
Room geometry	A static room of 2.5m by 2.5m is used. The loudspeaker is placed in the center of the room and is moved around

Monte Carlo setup	**for** x in grid **do**
	  for y in grid **do** ▷ Move loudspeaker on cartesian grid
	  for orientation **do** ▷ rotate 60° each
	  $\{R,\theta\}_i \leftarrow$ getImageSourceLocations$(x, y,$ orientation) ▷ Get 8 image sources
	  $\mathbf{h} \leftarrow$ constructGainVector$(\mathbf{R}, \boldsymbol{\theta})$ ▷ Using Eq. (3.11)
	  $\mathbf{y} \leftarrow \mathbf{\Phi}_{\text{genelec}}\mathbf{h}+$ ▷ Using Eq. (3.25). Noiseless
	  $\hat{\mathbf{h}}_{\text{MF-DAS-omni}} \leftarrow \mathbf{\Phi}^{\dagger}_{\text{omni}}\mathbf{\Lambda}_{\text{ls}}\mathbf{y}$
	  $\hat{\mathbf{h}}_{\text{MF-DAS-dir}} \leftarrow \mathbf{\Phi}^{\dagger}_{\text{genelec}}\mathbf{\Lambda}_{\text{ls}}\mathbf{y}$
	  $\hat{\mathbf{h}}_{\text{sparse-omni}} \leftarrow \arg\min_{\mathbf{h}} \left(\mathbf{y}-\mathbf{\Phi}_{\text{omni}}\mathbf{h}\right)^{\dagger}\mathbf{\Lambda}_{\text{ls}}\left(\mathbf{y}-\mathbf{\Phi}_{\text{omni}}\mathbf{h}\right)+\lambda\left\|\mathbf{h}\right\|_1$
	  $\hat{\mathbf{h}}_{\text{sparse-dir}} \leftarrow \arg\min_{\mathbf{h}} \left(\mathbf{y}-\mathbf{\Phi}_{\text{genelec}}\mathbf{h}\right)^{\dagger}\mathbf{\Lambda}_{\text{ls}}\left(\mathbf{y}-\mathbf{\Phi}_{\text{genelec}}\mathbf{h}\right)+\lambda\left\|\mathbf{h}\right\|_1$
	  $\hat{R},\hat{\theta} \leftarrow \text{maxk}\left(\hat{h}(R,\theta), 8\right)$ ▷ for all 4 estimates
	  {hitrate, hitratePerDOA } $\leftarrow$ compare$(\hat{R},\hat{\theta},R,\theta)$
	
	  end for
	  end for
	end for

Performance metric	Mean hitrate (averaged over 8 sources) and the hit rate as a function of direction of arrival
Hypothesis	If a typical loudspeaker is used, one cannot assume that the loudspeaker is omnidirectional as the true directivity will introduce model missmatches and will decrease performance.
Variables and size	$N=6$, $P=1$, $T=774$, $M=1025$ $K=250$. Room size $[2.5, 2.5m]$. Loudspeaker is moved around in a grid. $\lambda=55$, $x=y=[0.71, 1.77]m$ in 21 steps. Repeated for 6 orientations. No Noise.

the on-axis loudspeaker direction. The exponential sine sweep was used as excitation signal. In Fig. 4.11 one can see that the loudspeaker with uniform circular array is placed near a large single wall. The loudspeaker is not placed on the floor, rather, it is placed about 1 meter above the floor.

In Fig. 4.12 one can find the measurements performed in this setup. To better interpret the results, the excitation signal has been removed and the direct path is subtracted. One can observe however, that the direct path has not been fully removed, as all channels have residuals around $t = 1.5$ms. At $t = 6$ms, in all orientations and

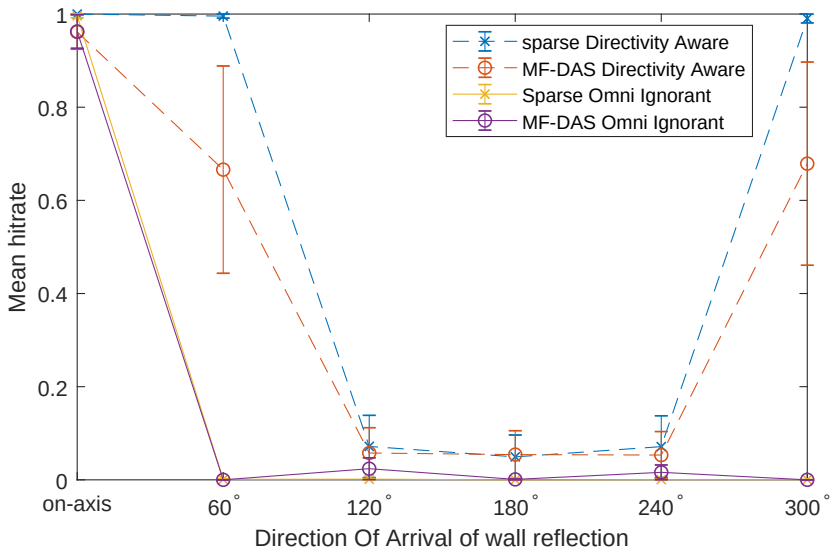
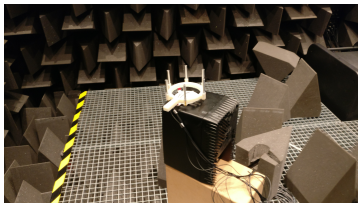
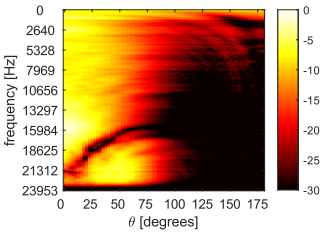


Figure 4.9: Experiment C: Hitrate averaged over all room positions and orientations. The reflections are grouped in direction of arrivals.

(a) Genelec 1029A positioned on the crane at the *cube* facility in Struer, Denmark. The loudspeaker is rotated to perform directivity measurements.



(b) An uniform circular microphone array with six microphones is placed on top of the loudspeaker. The direct path is measured in anechoic conditions.



(c) Magnitude frequency response of the loudspeaker impulse response for various transmit angles.

Figure 4.10: Experiment C: The Genelec 1029A is measured to construct a directivity model. In a separate measurements the direct path is measured for all microphones in the array.

in each channel, an event is picked up. The hypothesis is that this is the first order reflection with the floor. One can observe that the reflection arrives roughly at the same time for each channel, indicating that it may have arrived from a horizontal wall parallel to the array's plane. Most interesting, however in Fig. 4.12 is to compare the

reflection that arrives at $t = 12$ms as this corresponds the the large wall at 2 meters. Furthermore at $t = 13$ms, a second reflection is detected, with a similar shape as the first. The hypothesis is that this is the second order reflection from the floor to the wall, back to the microphone.

Most notably, the power and shape of the reflection at $t = 12$ms changes with the orientation of the system. Upon closer look (top plot in Fig. 4.13), one can detect the vertical sine shape over all microphone measurements, where the orientation of the loudspeaker determines the first microphone to pick up the plane wave. One can also observe that the amplitude of this reflection decreases as the loudspeaker is rotated. This suggests that a forward measurement model benefits from loudspeaker directivity awareness. In the experimental study that follows, the fundamental assumption that a single Rotated Image Source Impulse Response can model the reflections in a practical scenario is challenged.

The experiment starts by pre-processing the measured data. The exponential sine sweep is again deconvolved and the channel is manually cut around 12ms to only capture the reflection from the wall (as seen in Fig. 4.13). The inverse problem is then solved and the single highest candidate location is picked and used as prediction reference. The predictions $\hat{\mathbf{y}}$ can be interpreted as a single row of the respective $\mathbf{\Phi}$. This is repeated for both the omnidirectional assuming as well as the directivity aware model. The complete procedure is explained in pseudocode in Algorithm 1. The hypothesis here is that the directivity aware model provides a better prediction compared to the omnidirectional loudspeaker. The second objective is to challenge the sparseness that we wish to enforce on our measurements.

In Fig. 4.13 one can see the sinus shape over the six microphones as the plane wave propagates over the uniform circular array. The sparse optimization problem is solved and the highest contributing row of $\mathbf{\Phi}$ is selected as the prediction. This candidate location correctly estimates the direction of arrival and to a high degree the distance of the reflecting surface, for all loudspeaker orientation. This is to be expected, as a single reflection is present in the measurements. Not surprisingly, when the loudspeaker is placed in the front, as seen in Fig. 4.13a, the two predictions are identical. In this case the loudspeaker impulse response for both models is equal for walls in the front direction. The other predictions are different. To quantify this, let us denote the cut measurements by $\mathbf{y}_c$ and the predictions based on a single RISIR by $\hat{\mathbf{y}}$. The normalized model miss fit is defined as

$$\epsilon = \|\mathbf{y}_c - \hat{\mathbf{y}}\|_2^2 / \|\mathbf{y}_c\|_2^2. \tag{4.2}$$

This error is computed for both the omnidirectionally assuming model as well as the directivity aware model. The results are given in Table 4.4. There are two interesting observations. First, the directivity aware model seems to reduce the least square error the most, indicating that the model miss match is lower compared to the omni directional assumption. Secondly, none of the predictions seem to match well. This challenges the assumption that the inverse problem can be solved by constraining the solution with $\|\mathbf{h}\|_0 \leq S$. One possible explanation is that the acoustic reflection at the wall surface has altered the signal. If this influence on the signal could be modeled by an LTI system that is identified by a finite impulse response filter, then the number of non-zero entries to expect would be equal to that filter order. In this case, the inverse

Table 4.4: Experiment D: The measured channel responses from Fig. 4.12 are compared with the single best Rotated Image Source Impulse Response.

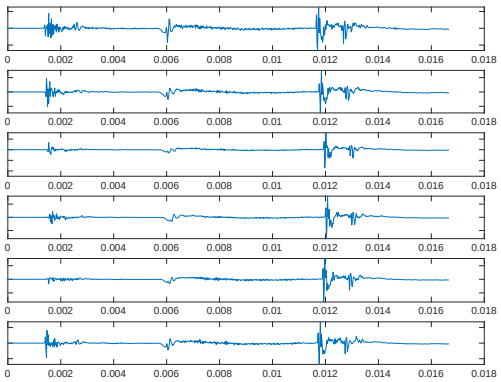
Wall orientation	Omni assuming	Directivity aware
0°	0.93	0.93
60°	0.92	0.73
120°	0.95	0.83
180°	0.95	0.74

Algorithm 1 Experiment C: Comparing the omnidirectional model with the directivity aware model

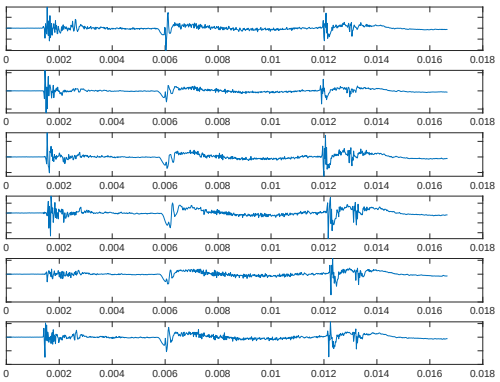
for Orientation **do**
  $\mathbf{y}_c \leftarrow \mathbf{y}(\text{orientation}, 11.5ms : 12.5ms, :)$ ▷ Cut single reflection for all microphones
  $\hat{\mathbf{h}}_{\text{dir}} \leftarrow \arg\min_{\mathbf{h}} \|\mathbf{y}_c - \mathbf{\Phi}_{\text{dir}}\mathbf{h}\|_2^2 + \lambda \|\mathbf{h}\|_1$
  $\hat{\mathbf{h}}_{\text{omni}} \leftarrow \arg\min_{\mathbf{h}} \|\mathbf{y}_c - \mathbf{\Phi}_{\text{omni}}\mathbf{h}\|_2^2 + \lambda \|\mathbf{h}\|_1$
  Ind1, Ind2 $\leftarrow \arg\max_i h_{\text{omni}}(i)$ and $\arg\max_i h_{\text{dir}}(i)$ respectively
  $\hat{\mathbf{y}}_{\text{omni}} \leftarrow h(Ind1)\mathbf{\Phi}_{\text{omni}}(Ind1)$ ▷ Take the dominating row as your prediction
  $\hat{\mathbf{y}}_{\text{dir}} \leftarrow h(Ind2)\mathbf{\Phi}_{\text{dir}}(Ind2)$ ▷ Normalized by the gain found
  modelMissfitOmni $\leftarrow \|\mathbf{y}_c - \hat{\mathbf{y}}_{\text{omni}}\|_2^2 / \|\mathbf{y}_c\|_2^2$
  modelMissfitDirectivity $\leftarrow \|\mathbf{y}_c - \hat{\mathbf{y}}_{\text{dir}}\|_2^2 / \|\mathbf{y}_c\|_2^2$
end for

problem may benefit from a group sparsity constrain [43].

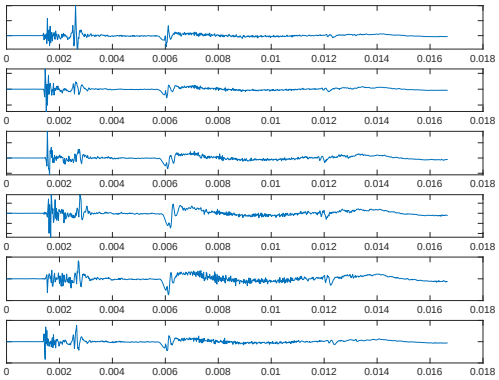
Figure 4.11: Genelec 1029A positioned in front of a large single wall. The microphone array is approximately one meter from the floor.



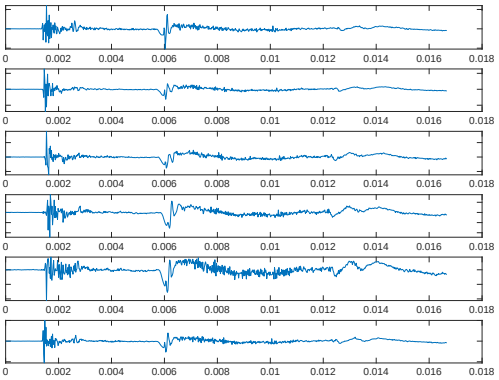
(a) Wall at 0°



(b) Wall at 60°

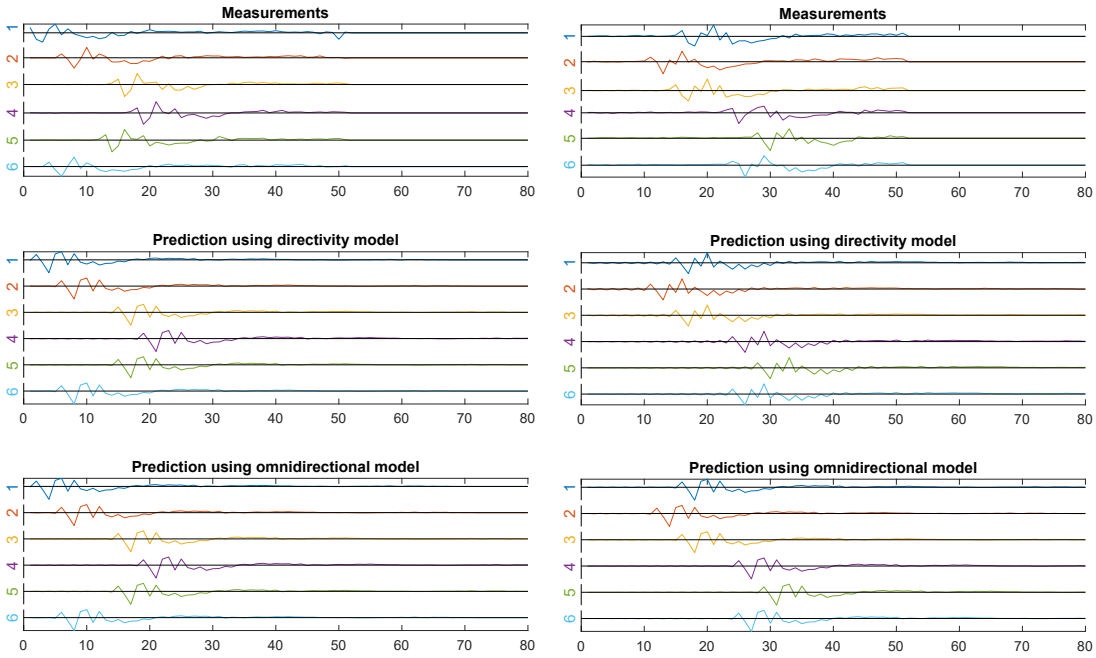


(c) Wall at 120°



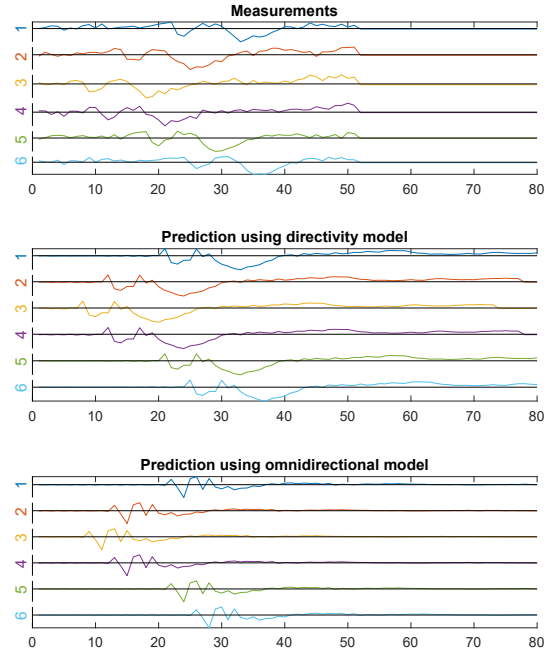
(d) Wall at 180°.

Figure 4.12: Experiment D: The Genelec 1029A is placed in front of a single wall at 2 meters. The excitation signal is deconvolved. The direct path from anechoic conditions is subtracted. The microphone channels are ordered such that for a) the first , b) the second, c) the third and d) the fourth microphone is closest to the wall.

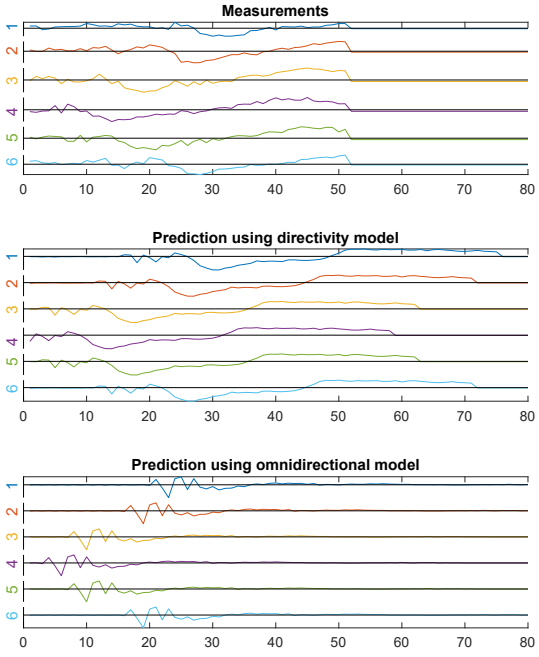


(a) Wall at 0°

(b) Wall at 60°



(c) Wall at 120°



(d) Wall at 180°.

Figure 4.13: Experiment D: The measurements (top) are manually cut to only include the reflection from the large wall. Two different model predictions are made. One is aware of the directivity of the loudspeaker (center) and the other assumes the front loudspeaker impulse response uniformly for all angles (bottom).

5 Conclusions

Inferring the location of reflecting surfaces is crucial in understanding the influence that the room has on the listening experience. Sound field estimation is extremely challenging, when a limited number of microphones are placed in a room. Knowing where the room boundaries are is helpful for practical acoustic signal processing scenarios. A smart loudspeaker system can exploit the reflections in a room to create a more rich sound stage by adjusting the mixing over the drivers. As a first step towards sound field control, we have presented a general framework to better resolve the times of arrival's of these reflections using a practical loudspeaker with co-located compact microphone array.

Contrary to prior art, the framework proposed in Chapter 3 is end-to-end, where all prior knowledge is exploited in a single step. Furthermore the proposed method does not have to assume an idealized acoustic point source nor an omnidirectional loudspeaker model. The Matched Filter Delay-And-Sum has been extended with the directivity aware loudspeaker model and a second estimation procedure that exploits a sparse prior is presented. The methods are shown to outperform the prior art when a typical loudspeaker is used. An experimental study with real world measurements shows that the extended model reduces the model miss fit.

5.1 Future work

To conclude, in this section a list of ideas is presented for possible extensions and future research topics:

- **Extending the method for horizontal surfaces** In my view, there are two plausible ways to extend the presented method to account for horizontal surfaces.
  - In the scenario where the microphone array geometry is limited to lie in the horizontal plane, an extension of the model $\mathbf{\Phi}$ can impact the conditioning such that the inverse problem cannot be solved. However, most high end loudspeaker systems are aesthetically designed to be placed in a living room in a rather specific way. In particular, larger systems such as the Bang & Olufsen Beolab 90's are likely to be placed directly on the floor. Other smaller products can be placed on a shelf or as table centerpiece. This prior knowledge can be used in conjunction with the standardized ceiling height used in construction. The distance to the floor and ceiling determines the location of the first order image sources in the horizontal axis. This structure can be exploited when solving the inverse problem. As was seen in the real world measurements, it is expected that a floor reflection follows in quick succession from a first order wall reflection. The search space could be reduced

drastically by including these priors.

- If one allows for a compact microphone array that can span in 3D, then the preferred geometry would be that of the uniform spherical array. In this situation the shift invariant property can be extended to three dimensions. It is imagined that $\mathbf{\Phi}$ can be extended to be a product of thee convolutions without reducing the conditioning.

- **Extending the signal model to multiple loudspeaker drivers**
  In the case where the loudspeaker system includes several drivers, the directivity model may be a combined model, measured during simultaneous excitation of all drivers. Alternatively, an individual directivity submodel may be determined for each driver, and then superimposed.

  A directivity model modeling each driver individually has the advantage that the estimation process may involve selectively exciting one or several drivers, and identifying walls for each such measurement. Significantly reducing the interference as was also shown using a single highly direction loudspeaker that was rotated in [32]. In the application of sound staging of object based audio, the presented forward model can also be used as a prediction. This could improve the mixing process.

- **Solving the inverse problem differently**

  - This thesis assumes that the candidate locations lie on the grid. Future work could study the errors introduces by reflectors not lying on the grid. One possible direction would be to reformulate the problem in a continuous domain and to use total variation for sparse recovery.
  - The proposed ℓ_1 regularization assumes that $h(l, k)$ is sparse. As shown in results, it is likely that a model miss match violates this assumption. An investigation into the structure of the non-zero elements could improve solving the inverse problem. One direction could be to assume group sparsity.
  - The location of second order image sources are determined if the two first order image sources are already located. This structure is currently not being exploited when solving for $\mathbf{h}$.

- **Using a predermined but observable excitation signal**
  We wish to have an inference procedure that requires least effort from the user of the product. It would be desirable to have that instead of an exponential sine sweep, customer music is used as excitation. However two main problems arise: The inverse of a large matrix must be computed (what was previously a closed analytical solution) and secondly, the music will likely not be wide band and contain periodicity. The research direction provided in this thesis may regularize the channel identification that could potentially overcome some of these issues.

Circulant Matrix

A

A circulant matrix has the form $\mathbf{A} \in \mathbb{C}^{n \times n}$

$$\begin{bmatrix} a_1 & a_2 & & \dots & a_n \\ a_n & a_1 & a_2 & \dots & a_{n-1} \\ a_{n-1} & a_n & a_1 & \dots & a_{n-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_2 & a_3 & \dots & a_n & a_1 \end{bmatrix} \tag{A.1}$$

Each row is the previous row cycled forward one step. A circulant matrix can thus be uniqly defined by its first row, often denotes as $\mathbf{A} = \text{circ}(\mathbf{a})$ such that if we use the cyclic permutation matrix $\mathbf{C}_n$ we have

$$\mathbf{C}_n = \begin{bmatrix} 0 & 1 & & \dots & 0 \\ \vdots & 0 & 1 & & \vdots \\ & & \ddots & \ddots & 0 \\ 0 & & & & 1 \\ 1 & 0 & & \dots & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{n-1\times 1} & \mathbf{I}_{n-1\times n-1} \\ 1 & \mathbf{0}_{1\times n-1} \end{bmatrix} \tag{A.2}$$

A circulant matrix $\mathbf{A}$ can be written in the form

$$\mathbf{A} = \sum_{k=0}^{n-1} a_{k+1} \mathbf{C}_n^k \tag{A.3}$$

Observe how we have $\mathbf{C}_n^0 = \mathbf{I} = \mathbf{C}_n^n$ and the vector $\mathbf{a}$ is the first row of the matrix. The polynomial representation reveals the commutative algebraic property of circulant matrices: Linear combinations and products of circulants are also circulant. The inverse of a nonsingular circulant is a circulant, any two circulants of the same size commute.

A.1 Eigenvectors and eigenvalues

The normalized eigenvectors of a circulant matrix are the Fourier modes. Thus any circulant matrix can be diagonalized with the same unitary transformation

$$\mathbf{A} = \mathbf{F}_n^{-1} \mathbf{\Lambda} \mathbf{F}_n = \mathbf{F}^{-1} \text{diag}(\mathbf{F}_n \mathbf{a}) \mathbf{F}_n \tag{A.4}$$

Where $\mathbf{F}_n \mathbf{a}$ is the Discrete Fourier Transform of the first row of $\mathbf{A}$. The N point discrete Fourier transform matrix is given as

$$F_N(k,l) = \frac{1}{\sqrt{N}} e^{-2\pi j k l / N} \tag{A.5}$$

Performing measurements using the exponential sine sweep

B

Throughout this thesis the exponential sine sweep is used as the excitation signal for the loudspeaker. The exponential sine sweep is an appropriate pilot signal for the loudspeaker response, as it is easily invertible and separates the harmonics due to non-linearities in time. As a consequence, the non-linearity can be removed easily. The exponential sine sweep is defined by the start frequency f_1, the end frequency f_2 and the total duration T as follows

$$x(t) = \sin\left(\frac{2\pi f_1 T}{\ln\left(\frac{f_2}{f_1}\right)}\left(\mathrm{e}^{\frac{t}{T}\ln\left(\frac{f_2}{f_1}\right)} - 1\right)\right) \tag{B.1}$$

To measure the loudspeaker impulse response, it is assumed that the loudspeaker is a linear time invariant system. Thus the signal model is

$$y(n) = x(n) * v(n) \tag{B.2}$$

where $v(n)$ is the loudspeaker impulse response. In matrix vector notation we have that

$$\mathbf{y} = \mathbf{X}\mathbf{v} \tag{B.3}$$

With

$$\mathbf{y} = \big[y(0), y(1), \ldots, y(T+L-2)\big]^\intercal \in \mathbb{R}^{T+L-1} \tag{B.4}$$

$$\mathbf{v} = \big[v(0), v(1), \ldots, v(L-1)\big]^\intercal \in \mathbb{R}^{L} \tag{B.5}$$

$$\mathbf{x} = \big[x(0), x(1), \ldots, x(T-1)\big]^\intercal \in \mathbb{R}^{T} \tag{B.6}$$

The matrix $\mathbf{X}$ is Toeplitz and of size $T + L - 2 \times L$ and has the following structure

$$\mathbf{X} = \begin{bmatrix} x(0) & 0 & \ldots & \ldots & \ldots & 0 \\ x(1) & x(0) & 0 & \ddots & & \vdots \\ \vdots & x(1) & x(0) & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ x(T-1) & x(T-2) & x(T-3) & \ldots & \ldots & x(0) \\ 0 & x(T-1) & x(T-2) & \ddots & \ddots & x(1) \\ \vdots & 0 & x(T-1) & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & x(T-1) \end{bmatrix} \in \mathbb{R}^{T+L-1 \times L} \tag{B.7}$$

Another way to express this matrix is by using the Circulant matrix. In the thesis we use $\mathbf{I}$ for the identity matrix and $\mathbf{W}_{a\times b}$ for the zero-padding/windowing matrix. The excitation signal column vector can be zero padded

$$\mathbf{x}_{\text{zp}} = \mathbf{W}_{T+L-1\times T}\mathbf{x} = \begin{bmatrix} \mathbf{x} \\ \mathbf{0}_{L-1} \end{bmatrix} \in \mathbb{R}^{T+L-1} \tag{B.8}$$

We construct a circulant matrix whose first column is $\mathbf{x}_{\text{zp}}$. The eigenvalue decomposition of the circulant matrix is known and uniquely defines the circulant matrix, as shown in Appendix A.

$$\mathbf{C}_x = \text{circ}\left(\mathbf{x}_{\text{zp}}\right) = \mathbf{F}^{-1}_{T+L-1}\mathbf{\Lambda}_x\mathbf{F}_{T+L-1} \tag{B.9}$$

Where we have that $\mathbf{\Lambda}_x$ is a diagonal matrix, constructed from the (complex) Fourier coefficients of the zero-padded excitation signal

$$\mathbf{\Lambda}_x = \text{diag}\left(\mathbf{F}_{T+L-1}\mathbf{x}_{\text{zp}}\right) \tag{B.10}$$

The loudspeaker impulse response can be estimated as

$$\hat{\mathbf{v}} = \mathbf{X}^{\dagger}\mathbf{y} \tag{B.11}$$

For the exponential sine sweep there is a closed form expression of the inverse filter (psuedo inverse). The inverse filter is a frequency-equalized time-reversed version of $x(n)$. If we define the inverse filter to be $z(n)$, we have that

$$z(n) * x(n) = \delta(n) \tag{B.12}$$

$$z(n) = x(-n)\log\left(\frac{f_2}{f_1}\right)\frac{n}{T}\forall n = 0,\ldots,T \tag{B.13}$$

$$\hat{v}(n) = z(n) * y(n) \tag{B.14}$$

This technique can also be used to estimate the room impulse response. Remember that the room impulse response is a function of loudspeaker position and listening position. So, the loudspeaker is placed somwhere in the room and the microphone is placed at the listening position. Or signal model now is

$$y(n) = x(n) * h(n) * v(n) \tag{B.15}$$

However, often the loudspeaker impulse response $v(n)$ is neglected. Thus to room impulse response calculations are similar to that of the loudspeaker impulse response. The difference being that the measurements are not taken in anechoic conditions but rather a room under test.

C Proof of farfield limit

Here we evaluate the following limit:

$$\lim_{R_s\to\infty}\sqrt{R_s^2+r^2-2R_sr\cos\left(\theta_s-\frac{2\pi i}{N}\right)}-R_s \tag{C.1}$$

For ease of notation we observe that $\cos\left(\theta_s-\frac{2\pi i}{N}\right)$ is not a function of R_s and replace it with the constant a. We rationalize the expression by multiplying it with 1.

$$\begin{aligned}\lim_{R_s\to\infty}\left(\left(\sqrt{R_s^2+r^2-2arR_s}-R_s\right)\frac{\sqrt{R_s^2+r^2-2arR_s}+R_s}{\sqrt{R_s^2+r^2-2arR_s}+R_s}\right)=\\ \lim_{R_s\to\infty}\frac{r^2-2arR_s}{R_s+\sqrt{r^2-arR_s+R_s^2}}.\end{aligned} \tag{C.2}$$

Using the product rule for limits we separate the equation and evaluate the limit for one of them

$$\begin{aligned}\lim_{R_s\to\infty}\frac{r^2-2arR_s}{R_s+\sqrt{r^2-arR_s+R_s^2}}&=\lim_{R_s\to\infty}\left(\frac{r^2-2arR_s}{R_s}\right)\lim_{R_s\to\infty}\left(\frac{1}{1+\frac{\sqrt{r^2-2arR_s+R_s^2}}{R_s}}\right)\\ &=-2ar\lim_{R_s\to\infty}\left(\frac{1}{1+\frac{\sqrt{r^2-2arR_s+R_s^2}}{R_s}}\right).\end{aligned} \tag{C.3}$$

Next, R is inserted in the root to obtain:

$$-2ar\lim_{R_s\to\infty}\left(\frac{1}{1+\frac{\sqrt{r^2-2arR_s+R_s^2}}{R_s}}\right)=-2ar\frac{1}{\lim_{R_s\to\infty}\left(\sqrt{\frac{r^2-2arR_s+R_s^2}{R_s^2}}\right)+1}. \tag{C.4}$$

The limit is taken inside the square root and the fraction over R^2 is evaluated:

$$-2ar\frac{1}{\sqrt{\lim_{R_s\to\infty}\frac{r^2}{R_s^2}-\frac{2ar}{R_s}+1}+1}. \tag{C.5}$$

Now the limit of $R_s\to\infty$ can be evaluated to finally obtain:

$$-2ar\frac{1}{\sqrt{\lim_{R_s\to\infty}\frac{r^2}{R_s^2}-\frac{2ar}{R_s}+1}+1}=-2ar\frac{1}{\sqrt{1}+1}=-ar \tag{C.6}$$

Substituting the original expression for a we establish the relationship:

$$\lim_{R_s \to \infty} \sqrt{R_s^2 + r^2 - 2R_s r \cos\left(\theta_s - \frac{2\pi i}{N}\right)} - R_s = -r \cos\left(\theta_s - \frac{2\pi i}{N}\right). \tag{C.7}$$

Or as it is presented in the body of the thesis, we can add r to both sides of the equation to establish

$$\begin{aligned} \lim_{R_s \to \infty} \Delta d(R_s, \theta_s, i) = &\lim_{R_s \to \infty} \sqrt{R_s^2 + r^2 - 2R_s r \cos\left(\theta_s - \frac{2\pi i}{N}\right)} - R_s + r \\ =&r\left(1 - \cos\left(\theta_s - \frac{2\pi i}{N}\right)\right). \end{aligned} \tag{C.8}$$

Bibliography

[1] dr. Jorge Martinez, "Low-complexity computer simulation of multichannel room impulse responses," Ph.D. dissertation, Delft University of Technology, Delft, The Netherlands, Nov. 2013. [Online]. Available: http://homepage.tudelft.nl/c7c8y/Theses/PhDThesisMartinez.pdf

[2] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 81–91, 2015.

[3] F. Jacobsen, M. Olsen, M. Møller, and F. T. Agerkvist, "A comparison of two strategies for generating sound zones in a room." in *18th International Congress on Sound and Vibration*. International Institute of Acoustics and Vibration, 2011.

[4] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, 2005.

[5] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers *et al.*, "Spatial audio object coding (saoc)-the upcoming mpeg standard on parametric object based audio coding," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[6] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection localization from room impulse responses," *ACTA Acustica united with Acustica*, vol. 98, no. 3, pp. 418–440, 2012.

[7] S. Cecchi, A. Carini, and S. Spors, "Room response equalization—a review," *Applied Sciences*, vol. 8, no. 1, p. 16, 2018.

[8] J. A. Pedersen, "Adaptive bass control-the abc room adaptation system," in *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society, 2003.

[9] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1929–1941, Oct, 2017.

[10] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[11] A. Torras-Rosell and F. Jacobsen, "A new interpretation of distortion artifacts in sweep measurements," *Journal of the Audio Engineering Society*, vol. 59, no. 5, pp. 283–289, 2011.

[12] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12 186–12 191, 2013.

[13] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *Signal Processing*, vol. 85, no. 3, pp. 549–570, 2005.

[14] T. van Waterschoot, G. Rombouts, and M. Moonen, "Optimally regularized adaptive filtering algorithms for room acoustic signal enhancement," *Signal Processing*, vol. 88, no. 3, pp. 594–611, 2008.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[16] P. N. Samarasinghe, T. D. Abhayapala, Y. Lu, H. Chen, and G. Dickins, "Spherical harmonics based generalized image source method for simulating room acoustics," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1381–1391, 2018.

[17] B. Bu, C. Bao, and M. Jia, "Simulating the three-dimensional room transfer function for a rotatable complex source," *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 100, no. 11, pp. 2487–2492, 2017.

[18] S. Hafezi, A. H. Moore, and P. A. Naylor, "Modelling source directivity in room impulse response simulation for spherical microphone arrays," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 574–578.

[19] M. Coutino, M. B. Møller, J. K. Nielsen, and R. Heusdens, "Greedy alternative for room geometry estimation from acoustic echoes: A subspace-based method," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 366–370.

[20] A. M. Torres, J. J. Lopez, B. Pueo, and M. Cobos, "Room acoustics analysis using circular arrays: An experimental study based on sound field plane-wave decomposition," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2146–2156, 2013.

[21] T. F. Brooks and W. M. Humphreys, "A deconvolution approach for the mapping of acoustic sources (damas) determined from phased microphone arrays," *Journal of Sound and Vibration*, vol. 294, no. 4-5, pp. 856–879, 2006.

[22] E. Tiana-Roig and F. Jacobsen, "Deconvolution for the localization of sound sources using a circular microphone array," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2078–2089, 2013.

[23] O. Lylloff, E. Fernández-Grande, F. Agerkvist, J. Hald, E. Tiana Roig, and M. S. Andersen, "Improving the efficiency of deconvolution algorithms for sound source localization," *The Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. 172–180, 2015.

[24] Y. Lin and D. D. Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 839–847, 2006.

[25] J.-H. Pan, C.-c. Bao, B. Bu, and M.-s. Jia, "Measurement of the acoustic transfer function using compressed sensing techniques," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[26] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1929–1941, 2017.

[27] E. Zea, "Compressed sensing of impulse responses in rooms of unknown properties and contents," *Journal of Sound and Vibration*, p. 114871, 2019.

[28] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1449–1460, 2011.

[29] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[30] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, p. 026503, 2006.

[31] F. Antonacci, J. Filos, M. R. Thomas, E. A. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2683–2695, 2012.

[32] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection path tracing using a highly directional loudspeaker," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 245–248.

[33] L. Remaggi, P. J. B. Jackson, P. Coleman, W. Wang, L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 2, pp. 296–309, Feb. 2017. [Online]. Available: https://doi.org/10.1109/TASLP.2016.2633802

[34] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[35] F. Jacobsen, T. Poulsen, J. H. Rindel, A. C. Gade, and M. Ohlrich, "Fundamentals of acoustics and noise control," *Department of Electrical Engineering, Technical University of Denmark*, 2011.

[36] W. M. Leach, *Introduction to electroacoustics and audio amplifier design.* Kendall/Hunt Publishing Company, 2003.

[37] P. C. Hansen, "Deconvolution and regularization with toeplitz matrices," *Numerical Algorithms*, vol. 29, no. 4, pp. 323–378, 2002.

[38] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM journal on computing*, vol. 24, no. 2, pp. 227–234, 1995.

[39] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[40] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate o (1/kˆ 2)," in *Dokl. akad. nauk Sssr*, vol. 269, 1983, pp. 543–547.

[41] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[42] D. E. Manolakis, "Efficient solution and performance analysis of 3-d position estimation by trilateration," *IEEE Transactions on Aerospace and Electronic systems*, vol. 32, no. 4, pp. 1239–1248, 1996.

[43] J. Huang and T. Zhang, "The benefit of group sparsity," 2009.