DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image and Vision*

Par
Diego DI CARLO

Echo-aware signal processing for audio scene analysis

«The Call of Echo»

Thèse présentée et soutenue à Rennes, le 04 December 2020
Unité de recherche : IRISA / INRIA
Thèse N° : 88666

Rapporteurs avant soutenance :

GIRIN Laurent	Professeur	GIPSA-Lab, Grenoble-INP
Simon DOCLO	Full professor	Carl von Ossietzky Universität, Oldenburg

Composition du Jury :

Président :	Laurent GIRIN	Professeur	GIPSA-Lab, Grenoble-INP
Examinateurs :	Simon DOCLO	Full professor	Carl von Ossietzky Universität, Oldenburg
	Renaud SEGUIER	Professeur	CentraleSupélec, Cesson-Sévigné
	Fabio ANTONACCI	Assistant professor	Politecnico di Milano
Dir. de thèse :	Nancy BERTIN	Chargée de recherche	IRISA, Rennes
Co-dir. de thèse :	Antoine DELEFORGE	Chargée de recherche	Inria Grand Est, Nancy

Abstract

Résumé en français

Acknowledgements

Contents

BIBLIOGRAPHY 113

Glossary:

- A list of terms in a particular domain of knowledge with their definitions.
- From Latin *glossarium* "collection of glosses", diminutive of *glossa* "obsolete or foreign word".

Notations

Linear Algebra

x, X	scalars
$\boldsymbol{x}, \mathbf{x}$	vectors
x_i	i-th entry of $\mathbf{x}$
$\mathbf{0}_I$	$I \times 1$ vector of zeros
$\mathbf{x}^\mathsf{T}$	transpose of the vector $\mathbf{x}$
$\mathbf{x}^\mathsf{H}$	conjugate-transpose (hermitian) of the vector $\mathbf{x}$
$\mathrm{Re}[x]$	real part scalar (vector) x ($\mathbf{x}$)
$\mathrm{Im}[x]$	imaginary part scalar (vector) x ($\mathbf{x}$)
i	imaginary unit
$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of real positive numbers
$\mathbb{C}$	set of complex number

Common indexing

i	microphone or channel index in $\{0, \ldots, I-1\}$
j	source index in $\{0, \ldots, J-1\}$
r	reflection (echo) in $\{0, \ldots, R-1\}$
t	continuous sample index
n	discrete sample index in $0, \ldots, N-1\}$
f	continuous frequency index
k	discrete frequency index in $\{0, \ldots, K-1\}$
l	discrete time-frame index $\{0, \ldots, L-1\}$
τ	tap index in $\{0, \ldots, T-1\}$

Geometry

$\underline{\mathbf{x}}_i$	3D location of microphone i recording $x_i(t)$
$\underline{\mathbf{x}}_i$	3D position of the microphone i recording $x_i(t)$
$\underline{\mathbf{s}}_j$	3D position of the source j emitting $s_j(t)$
$d_{ii'}$	distance between microphone i and i'
q_{ij}	distance between microphone i and source j
$\underline{\mathbf{s}}_j$	3D location of (target) point source j emitting $s_j(t)$
$\underline{\mathbf{q}}_j$	3D location of (interfering) point source j emitting $q_j(t)$
r_j	distance of source j wrt to the array origin
θ_j	azimuth of source j wrt to the array origin
φ_j	elevation of source j wrt to the array origin

SIGNALS

x_i	input signal recorded at microphone i
$\mathbf{x}$	$I \times 1$ multichannel input signal, i.e. $\mathbf{x} = [x_0, \ldots, x_{I-1}]$
$\mathbf{X}$	matrix of multichannel input signals
s_j	(target) point source signal j
q_j	(interfering) point source signal j
c_{ij}	spatial image source j as recorded at microphone i
a_{ij}	acoustic impule response from source j to microphone i
h_{ij}	generic filter from source j to microphone i
n_i	(white **or** distortion) noise signal at microphones i
u_i	generic interfering **and** distrortion noise signal at microphone i
ε_i	generic noise signal due to mis- or under-modeling i

ACOUSTIC

α_r	attenuation coefficient at reflection r
β_r	reflection coefficient at reflection r
τ_r	time location of the reflection r
c_{air}	speed of sound in air
T	temperature
H	relative humidity
p	sound pressure
h_{ij}	Room Impulse Response between source j to microphone i

MATHEMATICAL OPERATION

$\star$	cross-correlation
$\circledast$	generalized cross-correlation
$*$	convolution

EXAMPLES

Acoustic Impulse Response for single source scenario:

$$a_i(t) = \sum_{r=0}^{R_i} \frac{\alpha_{ir}}{4\pi c_{\text{air}} \tau_{ir}} \delta(t - \tau_{ir}) \tag{1}$$

Acoustic Transfer Function for single source scenario:

$$a_i(f) = \sum_{r=0}^{R_i} \frac{\alpha_{ir}}{4\pi c_{\text{air}} \tau_{ir}} e^{-\jmath 2\pi f \tau_{ir}} \tag{2}$$

Time of Arrival between source and microphone

$$\tau_{ij} = \frac{\left\| \underline{\mathbf{x}}_i - \underline{\mathbf{s}}_j \right\|}{c_{\text{air}}} \tag{3}$$

Part I

PROLOGUE

1

Overture

"*Écho. Citer ceux du Panthéon et du pont de Neuilly.* "
—Gustave Flaubert, Dictionnaire des idées reçues

"*Echoes shows the direction that we're moving in.* "
—David Gilmour, about the making of "The Dark Side Of The Moon"

▶ This Ph. D. thesis is about acoustic Echoooes
We are surrounded by Acoustic Echoes

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

In the everyday context, when a sound reflection is perceived distinctly is referred to as *echo*. While phenomenon can be observed clearly in outdoors environment, such in the mountains or within huge buildings, in closed rooms it is less noticeable. In fact, echoes are usually masked by a general reverberation of the room.

Sound, every sound, propagate in air and in the space before reaching our hears. It travels

1.1 Echo-aware Signal Processing for Audio Scene Analysis

The problems addressed in this thesis are indicated in the thesis title: *Echo-aware signal processing for audio scene analysis*. There are two parts in the sentence that deserve an explanation: *echo-aware signal processing* and *audio scene analysis*. Let us start from the second one which provide the general context.

1.1.1 *Audio scene analysis*

Sounds carry information about sound sources.

Information about the *content* and the *nature* of the sound. But other information as well.

▶ Audio Scene Analysis

- Audio Scene Synthesis

1.1.2 *Echo-aware signal processing*

Signal processing is the process of analyzing and modifying a *signals*, which are mathematical representation of quantities carrying information about a phenomenon. When this signals represents sound, such as music or speech, then we speak about *sound* or *audio signal processing*. *Audio signal processing* involves applying various mathematical and computational techniques to analog and digital signals. There are multiple reasons to do this, such as produce new signals with higher quality than the original signal and extract high-level information the signal carries. In order to achieve this, complex system are built which can be represented as collection of simpler subsystems, with well-defined tasks, interacting with each other. In (audio) signal processing, these subsystems roughly fall into four categories: *representation*, *enhancement*, *estimation*, and *adaptive processing*. Many related problems can be then decomposed into blocks that belong to one of more of these categories.

Audio is a more technical term, referring to sound coming from a recording, transmission or electronic device. Sound is a more generic word and can be caused by any source.

Representation Signal can be represented and described in many different way. Through different representations, the *information* contained in the signals becomes more relevant and suitable for certain tasks than other.
Representation can be lossy or lossless, and are generally implemented through change of *domain* or *feature*. The most famous representation in case of audio is the Fourier basis which change the signal domain from from time to frequencies. The process of changing representation is often called: *analysis* and *synthesis*.

Enhancement Measurement are affected by noise and interferences which corrupt and hide relevant information, making its retrieval harder and sometimes impossible. Therefore, signal enhancement, that is removing noise, is typically a necessary step.
Enhancement constitute a huge dome of methods: form simple denoising by averaging of repeated measurement to huge system based on neural network.

Estimation Often we wish to estimate some key properties of the target signal which may be used as inputs to a different algorithm.

Adaptive processing deals with adaptive algorithms and filters controlled by variable parameters. A common means to adjust those parameters according to an optimization algorithm which rely on statistical properties of the signal of interest. They often implement a kind of online optimization where an objective function is being minimized. When new data is observed, its discrepancy with the current estimate is used to produce a new estimate in a way that reduces the objective.

That is being said, the goal of this thesis is to improve the above state of in indoor audio signal processing along two axes: First, by deepening our understanding of acoustic echoes, provide new methodologies to estimate them surpassing the limits of current approaches. Second, by extending previous

echo-aware methods, show how typical audio application can benefits of prior knowledge of these elements of acoustic propagation.

To that end, the dissertation demonstrates two claims:

1. Acoustic echoes can be estimated blindly from microphone recordings;
2. Typical audio scene analysis and audio processing methods can take advantage of acoustic echoes, by easily integrating their knowledge in standard algorithms.

Based on this idea, so-called *echo-aware* methods have been introduced few decades ago, where matched filters (or rake receivers) are used to constructively sum the sound reflections [Jan et al. 1995; Affes and Grenier 1997] and build beamformers achieving much better sound qualities [Gannot et al. 2001]. This methods have recently regained interested as manifested by the European project SCENIC [Annibale et al. 2011] and the UK research S^3A project. They show that knowing the properties of a few early echoes can boosts performances of typical indoor audio inverse problems such as speech enhancement (SE) [Dokmanić et al. 2015; Kowalczyk 2019], sound source localization [Ribeiro et al. 2010b; Di Carlo et al. 2019a], and separation [Scheibler et al. 2018c; Leglaive et al. 2016]. Another fervent area of research spanning transversely the audio and acoustic signal processing fields is estimating the room geometry blindly from acoustic signals. As presented by Crocco *et al.* in [Crocco et al. 2017], the end-to-end room geometry estimation (RooGE) involves many subsequent subtasks: RIR estimation, peak picking, microphones calibration, echo labeling, reflectors estimation. Acoustic echo retrieval (AER) is common to many of these topics. It consists in estimating the properties of echoes such as their TOAs and energies. The former problem is referred to as TOA estimation, or time-delay estimation when the direct-path is taken as reference. Furthermore, as interesting applications, these methods have been recently used in active scenarios, namely knowing the transmitted signals, using unmanned aerial vehicle (UAV, a.k.a. drones) [Jensen et al. 2019; Boutin and Kemper 2020] and mobile-phones [Shih and Rowe 2019].

1.2 Audio Inverse Problems

[Kitic 2015] In § 1.1 we have informally defined *inverse problems*, with an emphasis on inverse problems in signal processing. An inverse problem is a type of a mathematical problem where we start with the observations and we want to estimate model parameters that produced them.
Inverse problems pervades all the field of science and engineering: source localization [], image processing [], acoustic imaging and tomography [],

A inverse problems is defined as the counterpart of a *forward*[1] problem. Without falling in and deep mathematical formalism and taxonomies which can be found in [Bal 2012], we will simply consider the following informal definition:

Forward problem *starts from known input, while* inverse problem *starts from known output [Santamarina and Fratta 2005].*

Kitic, "Cosparse regularization of physics-driven inverse problems"

"*Their generality is of such a wide scope that onemayeven argue that solving inverse problems is what signal processing is all about*"
—Srdan Kitić, *Cosparse regularization of physics-driven inverse problems*

"*everything is an optimization problem*"
—[Watson et al. 2001]

One can see the paralelism with the engineering concepts: analysis and sythesis.

A historical example are the calculation of the Earth circunference by Eratosthenes in III century b.c.
and the calculations of Adams and Le Verrier which led to the discovery of Neptune from the perturbed trajectory of Uranus.

[1] often referred to as *direct*

Santamarina and Fratta, "Discrete signals and inverse problems"

Both these problems focus on an operation relating maps objects of interest, called *parameters* or *variables*, to information collected about these objects, called *measurements*, *data* or *observation*.

For instance, in our context, the direct problem may be the estimation of the Room Impulse Response (RIR)(s) starting from the known room parameters, and, the related inverse problem would be the estimation of such room properties from the observation of the RIR(s).

Formally, a forward problem is defined through a mathematical model, described by a *operation* $\mathcal{M}(\cdot)$ mapping *parameters* $x \in \mathcal{X}$ to the *observation* (or measurement) $y \in \mathcal{Y}$:

$$y = \mathcal{M}(x). \tag{1.1}$$

Then, the inverse problem defines a method $\mathcal{M}^{-1}$ that "reverts" $\mathcal{M}$ in order to recover (estimate) x form the observation of y.

As discussed in [Bal 2012], *solving* the inverse problem consists in finding point(s) $x \in \mathcal{X}$ from (knowledge of) data $y \in \mathcal{Y}$ such that Eq. (1.1) or an approximation of Eq. (1.1) holds. Under this light, the operator $\mathcal{M}$ ant the choice of $\mathcal{X}$ describes our best effort to construct a *model* for the data y and the space where the parameters x belong, respectively.

Bal, "Introduction to inverse problems"

one can already see the paralelism the the definition of the mixing process defined in § 1.1

For instance, in Case of *linear* inverse problem, and for $\mathcal{Y}$ and $\mathcal{X}$ being vector spaces of dimensions M and N respectively, then the forward map can be written as a linear system:

$$\mathbf{y} = \mathbf{M}\mathbf{x} \tag{1.2}$$

where $\mathbf{M}$ being a matrix, namely the operator $\mathcal{M}$ becomes a matrix multiplication by M. It follows that the inverse map associated to Eq. (1.2) is the application of the inverse matrix M^{-1}.

Typically, forward problems are considered somehow the "easier". In fact, even in the observation model $\mathcal{M}$ is known perfectly, it is not always possible to find its counterpart. This because of

- presence of *noise* in the measurement which are not always additive and statistically independent w. r. t. x.
- the problem is *well-posed* and *well-conditioned*, namely $\mathcal{M}$ needs be injective and stable. In other words, some information is recoverable, other is completely lost, other highly sensible to noise [2].

[2] **injective** ensure the uniqueness of the solution, while **stability** ensure a continuity on the data. These are known as the Hadamard's *solvability conditions*.

As we could images, many interesting and fundamental inverse problem are *ill-posed* or *ill-conditioned* in general, even in the following "simple" ones [Kitic 2015]: The solution to the deconvolution problem, where the direct inversion of the transfer function results in instabilities at high frequency; and the solution a linear system $\mathbf{y} = \mathbf{M}\mathbf{x}$ where $\mathbf{M}$ is invertible may lead to erroneous results and numerical instabilities.

Kitic, "Cosparse regularization of physics-driven inverse problems"

Therefore, sometimes ones have to settle for restring the set of solution $\mathcal{C} \subset \mathcal{X}$, where $\mathcal{M}$ is stable and injective[3]. Promoting solution $x \in \mathcal{C}$ is can be achieved through *model priors*, namely prior knowledge about solution, which can be classified in the following methodologies: the usage of *geometric constraints* that deterministically define the solutions; the imposition of *penalization* which "promotes" solution of a certain shape (e. g. *sparse* [4] or

[3] This framework was originally proposed by Tikhonov.

[4] **sparsity** is a fundamental concept of this thesis, better discussed in Part III

smoothness); and casting the problem in a *bayesian framework* which versatilely incorporate prior and posterior density function describing the data.

Let us give two example of practical systems that will be recurrent thought out the entire thesis.

1.2.1 *Selected Audio Inverse Problems*

Here follow some famous problems in the field of audio signal processing with application to speech, music and environmental audio. Given the mixing process defined in § 3.1,

Inverse Problem	*Can we estimate the...*
Audio Source Separation	the signal of the sources s_j from the mixture $\boldsymbol{x}$?
Sound Source Localization	the position $\mathbf{s}_j = [x_{s_j}, y_{s_j}, z_{s_j}]$ of the source s_j from the mixture $\boldsymbol{x}$?
Microphone (Array) Calibration	the position of the microphone (array) position $\mathbf{x}$ from the mixture $\boldsymbol{x}$?
RIR Estimation	the filter between the sources $\boldsymbol{s}_j$ and the mixture $\boldsymbol{x}$ from $\boldsymbol{x}$?
Room Geometry Estimation	the shape of the room in which the mixture $\boldsymbol{x}$ recoding source s_j?

Table 1.1: Selected audio inverse problems

"Everything is connected"
—Douglas Adams, *Dirk Gently's Holistic Detective Agency*

- Depending on the scenario, all these problems exhibits strong inter-connections, namely the solution of one may be (dependent on) the solution of another. Therefore, exploiting expertise and knowledge, interconnect and hierarchical approaches may be built[5]: for instance, many spatial filtering techniques used for Speech Enhancement (SE) rely on Sound Source Localization (SSL) blocks; and in order to achieves Room Geometry Estimation (RooGE), AER must be done.

[5] Machine Learing allows now for end2end approaches

1.3 Thesis Outline and Main Contributions

The dissertation is broken into three largely parts which are largely interconnect, as show in the Figure 1.1:

- Room Acoustic meets Signal Processing

  Chapter 2 This chapter will build a first important bridge: from acoustics to audio signal processing. It first defines sound and how it propagates in the environment § 2.1, teasing out the fundamental concepts of this thesis: the echoes. § 2.2 and the Room Impulse Response (RIR) § 2.3. By assuming some approximations, the RIR will be described in all its parts in relation with methods to compute them. Finally, in § 2.4, how the human auditory system perceives reverberation will be reported.

  Chapter 3 Let us now move from the physics to digital signal processing. At first in § 3.1, this chapter formalizes fundamental concepts of audio signal processing such as signal, mixtures and noise in the time domain. In § 3.2, we will present the signal representation that we will use throughout the entire thesis: the Short Time Fourier Transform (STFT). Finally, after assuming the narrowband approximation, in § 3.3 some important models for the Room Impulse Response (RIR) are described.

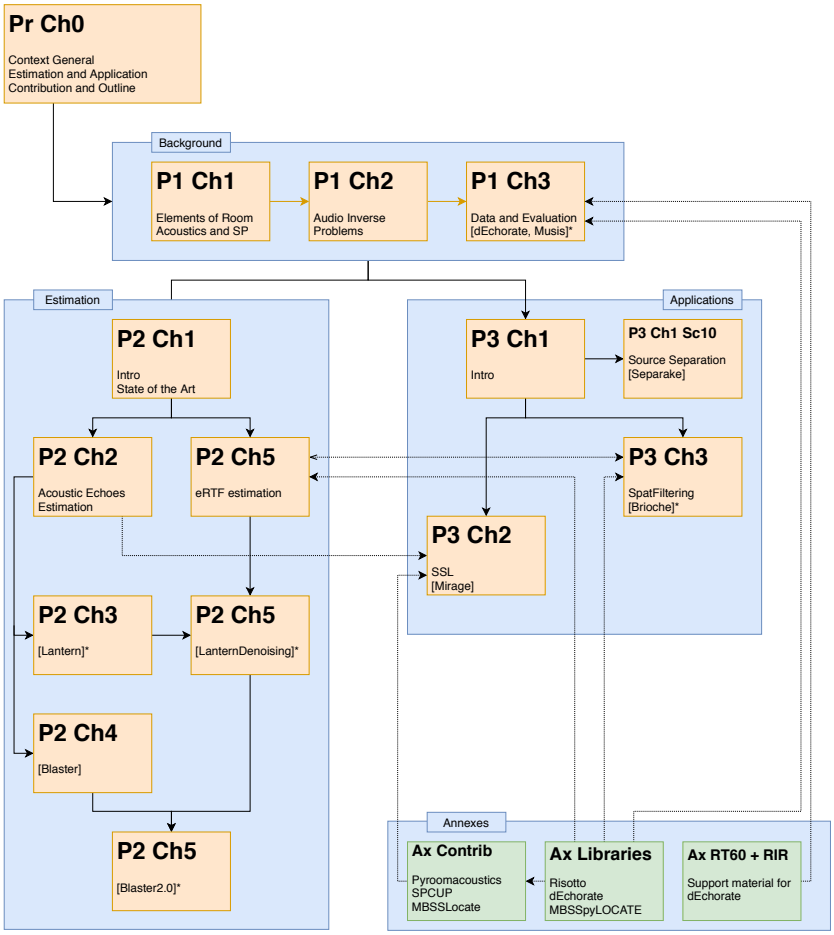


FIGURE 1.1: Schematic rganization of the thesis, dependecies between chapters linked to author contributions.

▶ ACOUSTIC ECHOES ESTIMATION

Chapter 4 This chapter amis to provide the reader with knowledge of the state-of-the-art of Acoustic Echo Retrieval (AER). After presenting the AER problem in § 4.1, the chapter is divided into three main sections: § 4.2 defines the categories of methods thank to which the literature can be clustered and analyzed in detail later in § 4.3. Finally, in § 4.4 some datasets and evaluation metrics for AER are presented.

?? Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 5 Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest

gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

?? Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

► Echo-aware Audio Scene Analysis

?? Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

?? Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

?? Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

?? Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get

no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.4 List of Contribution

This dissertation draws heavily on the earlier work and writing in the following papers, written jointly with several collaborators:

- Di Carlo, Diego, Pinchas Tandeitnik, Sharon Gannot, Antoine Deleforge, and Nancy Bertin (2021). "dEchorate: a calibrated Room Impulse Response database for acoustic echo retrieval". In: *Workin progess*
- Di Carlo, Diego, Clement Elvira, Antoine Deleforge, Nancy Bertin, and Rémi Gribonval (2020). "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 156–160
- Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019b). "Mirage: 2d source localization using microphone pair augmentation with echoes". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 775–779
- Deleforge, Antoine, Diego Di Carlo, Martin Strauss, Romain Serizel, and Lucio Marcenaro (2019). "Audio-Based Search and Rescue With a Drone: Highlights From the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]". In: *IEEE Signal Processing Magazine* 36.5, pp. 138–144
- Lebarbenchon, Romain, Ewen Camberlein, Diego Di Carlo, Clément Gaultier, Antoine Deleforge, and Nancy Bertin (2018). "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge". In: *arXiv preprint arXiv:1812.05901*
- Scheibler, Robin, Diego Di Carlo, Antoine Deleforge, and Ivan Dokmanić (2018d). "Separake: Source separation with a little help from echoes". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901

1.5 Don't Panic!

The reader will have already noticed that a large margin is left free on the right side of each page of the manuscript. We will use it to insert comments, historical notes as well as figures and tables to complete the subject. This graphic charter is inspired by the work of Tufte (2001) and produced using the latex tufte-latex class. We emphasize that the presence of the clickable GitHub logo in the margin indicates the online availability of the codes.

1.5.1 *Quick vademecum*

for the readers:

- Bibliographic references are denoted as [Kuttruff 2016]. Kuttruff, *Room acoustics*
- Figures, Tables and other floating objects as well as equations are numbered within the chapter number.
- Equations are referred as Eq. (2.6)
- The main matter of the Thesis's manuscript starts at page 1, until page 103.
- The back matter covers the list of the candidate's publications and the bibiographic references cited along the text.
- Small notes on the margin might be used to easily navigate through the Example of margin note manuscript. They are meant to summarize paragraphs/blocks of text.
- The end of the chapter is shown by the following sign between horizontal rules.

1.5.2 *The golden ratio of the thesis*

- at most 3 level of sub-headings: section, subsection and new-thought
- usage of dichotomies are preferred
- each paragraph is introduced briefly at the end of the previous one
- definition are provided with stacco
- Not important figures: without numbering

Part II

ROOM ACOUSTIC MEETS SIGNAL PROCESSING

2

Elements of Room Acoustics

"Sound, a certain movement of air."
—Aristotele, De Anima II.8 420b12

▸ Synopsis This chapter will build a first important bridge: from acoustics to audio signal processing. It first defines sound and how it propagates in the environment § 2.1, teasing out the fundamental concepts of this thesis: the echoes. § 2.2 and the Room Impulse Response (RIR) § 2.3. By assuming some approximations, the RIR will be described in all its parts in relation with methods to compute them. Finally, in § 2.4, how the human auditory system perceives reverberation will be reported.
The material on waves and acoustic reflection is digested from classic texts on room acoustics [Kuttruff 2016; Pierce 2019] and on partial differential equations [Duffy 2015].

2.1 Sound wave propagation

Imagine a calm pond. The surface is flat and smooth. Drop a rock into it. *Kerploop!* The surface is now disturbed. The disturbances spread propagate, as waves. The medium here is the water surface.

According to common dictionaries and encyclopedias,

> *sound is the sensation perceived by the ear caused by the vibration of air.*

This definition highlights two aspects of sound: a physical one, characterized by the air particles vibration; and a perceptual one, involving the auditory system. Focusing on the former phenomenon, when vibrating objects excites air, surrounding air molecules starts oscillating, producing zones with different air densities leading to a compressions-rarefactions phenomenon. Such vibration of molecules takes place in the direction of the excitement, with the next layer of molecules excited by the previous one. Pushing layer by layer forward, a *longitudinal mechanical wave*[6] is generated. Notice that therefore sound needs a medium to travel: it cannot travel through a vacuum and no sound is present in outer space.

[6] As opposed to mechanical vibrations in a string or (drum) membrane, acoustic vibrations are *longitudinal* rather than *transversal*, i. e. the air particles are displaced in the same direction of the wave propagation.

Thus sound propagates though a medium, which can be solid, liquid or gaseous. The propagation happens at a certain speed which depends on the physical properties of the medium, such as its density. The medium assumed throughout the entire thesis is air, although extensions of the developed methods to other media could be envisioned. Under the fair assumption of air being homogeneous and steady, the speed of sound can be approximated as follows:

$$c_{\text{air}} = 331.4 + 0.6T + 0.0124H \quad [\text{m/s}], \tag{2.1}$$

where T is the air temperature [°C] and H is the relative air humidity [%].

The air pressure variations at one point in space can be represented by a *waveform*, which is a graphical representation of a sound Figure 2.2.

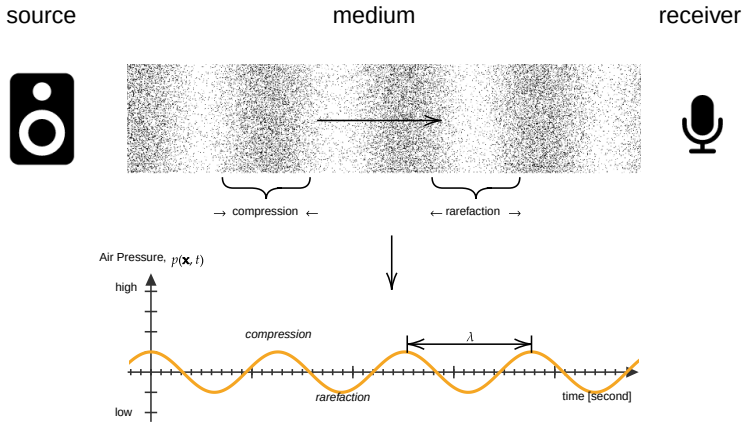


FIGURE 2.2: Illustration of the molecules under sound compression and rarefaction due to longitudinal sound wave and its waveform representation.

We can think of this process in the light of the classic *source-medium-receiver* model of communication theory: the *source* is anything that emits waves[7], the *medium* carries the waves from one point to another, and the *receiver* absorbs them.

[7] example of sources are vibrating solids (e. g. loudspeakers membrane), rapid compression or expansion (e. g. explosions or implosions) or air vortices with characteristics frequencies (e. g. flute and whistles).

2.1.1 *The acoustic wave equation*

The acoustic wave equation is a second-order partial differential equation[8] which describes the evolution of acoustic pressure p as a function of the position $\underline{\mathbf{x}}$ and time t

$$\nabla^2 p(\underline{\mathbf{x}}, t) - \frac{1}{c^2}\frac{\partial^2 p(\underline{\mathbf{x}}, t)}{\partial t^2} = 0, \tag{2.2}$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ stands for the 3-dimensional *Laplacian* operator. The constant c is the sound velocity in the medium and has dimension $\left[\frac{\mathrm{m}}{\mathrm{s}}\right]$. Despite its complicated formulation, the wave equation is linear. Thus it implies the followings:

- the pressure field at any time is the sum of the pressure fields resulting from each source at that time;
- the pressure field emitted at a given position propagates over space and time according to a linear operation.

[8] In 1746, d'Alembert discovered the one-dimensional wave equation for music strings, and within ten years Euler discovered the three-dimensional wave equation for fluids.

Assuming the propagation of the wave in a homogeneous medium, one can obtain the equation above by combining three fundamental physical laws:

- the *conservation of momentum*,
- the *conservation of mass*, and
- the *polytropic process relation*, meaning that the medium is an ideal gas undergoing a reversible adiabatic process.

However, media are not uniform and feature inhomogeneities of two types: scalar inhomogeneities, e. g. due to temperature variation, and vector

inhomogeneities, e. g. due to presence of fans or air conditioning. Although these affect the underlying assumption of the model, the effects are small in typical application of speech and audio signal processing. Therefore they are commonly ignored.

▸ THE HELMHOLTZ'S EQUATION

The equation 2.2 is expressed in the space-time domain $(\underline{\mathbf{x}}, t)$. By applying the temporal Fourier transform, we obtain the *Helmholtz equation*:

$$\nabla^2 P(\underline{\mathbf{x}}, f) + k^2 P(\underline{\mathbf{x}}, f) = 0, \tag{2.3}$$

where $k = \frac{2\pi f}{c}$ is known as *wave number* and relates the frequency f to the propagation velocity c.

Both the wave 2.2 and the Helmholtz's equation 2.3 are source-independent, namely no source is present in the medium. Therefore they are said to be *homogeneous* as the right-hand term is zero. Normally the sound field is a complex field generated by acoustics sources. As consequence, the two equations become inhomogeneous as some non-zero terms needs to be added to the right-hand sides.

In the presence of a sound source producing waves with source function $s(t, \underline{\mathbf{x}})$, the wave equation can be written

$$\nabla^2 p(\underline{\mathbf{x}}, t) - \frac{1}{c^2} \frac{\partial^2 p(\underline{\mathbf{x}}, t)}{\partial t^2} = s(t, \underline{\mathbf{x}}). \tag{2.4}$$

Thus, the corresponding Helmholtz's equation writes

$$\nabla^2 P(\underline{\mathbf{x}}, f) - k^2 P(\underline{\mathbf{x}}, f) = S(\underline{\mathbf{x}}, f). \tag{2.5}$$

For instance one can assume an infinitesimally small pulsating sphere locate at $\underline{\mathbf{s}}$ radiating constant acoustic energy at frequency f, i. e. $S(\underline{\mathbf{x}}) = \delta(\underline{\mathbf{x}} - \underline{\mathbf{s}})$. At the receiver position $\underline{\mathbf{x}} \neq \underline{\mathbf{s}}$, the Helmholtz's equation writes

$$\nabla^2 H(f, \underline{\mathbf{x}} \mid \underline{\mathbf{s}}) - k^2 H(f, \underline{\mathbf{x}} \mid \underline{\mathbf{s}}) = \delta(\underline{\mathbf{x}} - \underline{\mathbf{s}}), \tag{2.6}$$

The function $H(f, \underline{\mathbf{x}} \mid \underline{\mathbf{s}})$ satisfying Eq. (2.6) is called the *Green's function* and is associated to Eq. (2.3), for which it is also a solution.

2.1.2 ... and its Green solution

Green's Functions are mathematical tools for solving linear differential equations with specified initial- and boundary- conditions [Duffy 2015]. They have been used to solve many fundamental equations, among which Eqs. (2.2) and (2.3) for both free and bounded propagation. They can be seen as a concept analogous to *impulse responses*[9] in signal processing. Under this light, the physic so-far can be rewritten using the vocabulary of the communication theory, namely *input*, *filter* and *output*.

According to Green's method, the equations above can be solved in the frequency domain for arbitrary source as follows:

$$P(f, \underline{\mathbf{x}}) = \iiint_{\mathcal{V}_{\tilde{s}}} H(f, \underline{\mathbf{x}} \mid \underline{\mathbf{s}}) S(f, \underline{\mathbf{s}}) \, \mathrm{d}\underline{\mathbf{s}}, \tag{2.7}$$

where $\mathcal{V}_{\tilde{s}}$ denotes the source volume, and $\mathrm{d}\underline{\mathbf{s}} = \mathrm{d}x_{\tilde{s}} \, \mathrm{d}y_{\tilde{s}} \, \mathrm{d}z_{\tilde{s}}$ the differential volume element at position $\underline{\mathbf{s}}$. If one ignores the space integral, one can see

By 1950 Green's functions for Helmholtz's equation were used to find the wave motions due to flow over a mountain and in acoustics. Green's functions for the wave equation lies with Gustav Robert Kirchhoff (1824–1887), who used it during his study of the three-dimensional wave equation. He used this solution to derive his famous *Kirchhoff's theorem* [Duffy 2015].

[9] Impulse responses in time domain, transfer functions in the frequency domain.

the close relation with a transfer function.
The requested sound pressure $p(\underline{\mathbf{x}}, t)$ can now be computed by taking the frequency-directional inverse Fourier transform of Eq. (2.7).

It can be shown [Kuttruff 2016] that the Green's function for Eqs. (2.3) and (2.6) writes

$$H(f, \underline{\mathbf{x}} \mid \underline{\mathbf{s}}) = \frac{1}{4\pi \|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|} \mathrm{e}^{-\frac{\mathrm{i}2\pi f \|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|}{c}} \tag{2.8}$$

where $\|\cdot\|$ denotes the Euclidean norm. By applying the inverse Fourier transform to the result above, we can write the time-domain Green's function as

$$h(t, \underline{\mathbf{x}} \mid \underline{\mathbf{s}}) = \frac{1}{4\pi \|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|} \delta\left(t - \frac{\|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|}{c}\right) \tag{2.9}$$

where $\delta(\cdot)$ is the time-directional Dirac delta function.
As consequence, the *free field*, that is open air without any obstacle, the sound propagation incurs a delay q/c and an attention $1/(4\pi q)$ as function of the distance $q = \|\underline{\mathbf{x}} - \underline{\mathbf{s}}\|$ from the source to the microphone.

Eqs. (2.8) and (2.9) are respectively the free-field transfer function and the impulse response.

According to Eq. (2.9), the sound propagates away from a point source with a spherical pattern. When the receiver is far enough from the source, the curvature of the *wavefront* may be ignored. The waves can be approximated as *plane waves* orthogonal to the propagation direction. This scenario depicted in Figure 2.3 is known as *far-field*. In contrast, when the distance between the source and the receiver is small, the scenario is called *near field*.

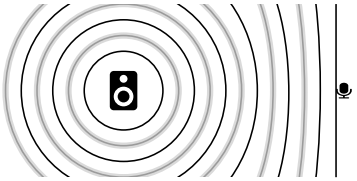


FIGURE 2.3: Visualization of the sound propagation. Since the sensor (i.e. a microphone) is drawn in the far field, the incoming waves can be approximated as plane waves.

2.2 ACOUSTIC REFLECTIONS

The equations derived so far assumed unbounded medium, i. e. free space: a rare scenario in everyday applications. Real mediums are typically bounded, at least partially. For instance in a room, the air (propagation medium) is bounded by walls, ceiling, and floor. When sound travels outdoor, the ground acts as a boundary for one of the propagation directions. Therefore, the sound wave does not just stop when it reaches the end of the medium or when it encounters an obstacle in its path. Rather, a sound wave will undergo certain behaviors depending on the obstacles' acoustics and geometrical properties, including

- *reflection* off the obstacle,
- *diffraction* around the obstacle, and
- *transmission* into the obstacle, causing
  - *refraction* through it, and
  - *dissipation* of the energy.

Reflections typically arise when a sound wave hits a large surface, like a room wall. When the sound meets a wall edge or a slit, the wave diffracts, namely it bends around the corners of an obstacle. The point of diffraction effectively becomes a secondary source which may interact with the first one. The part of energy transmitted to the object may be absorbed and refracted. Objects are characterized by a proper acoustic resistance, called *acoustic*

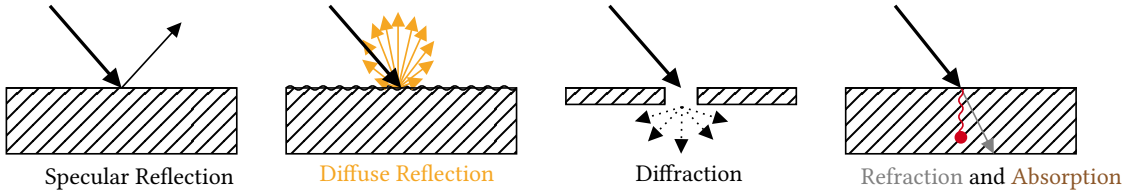


FIGURE 2.4: Different types of sound interact with a surface.

impedance, which describes their acoustic inertia as well as the energy dissipation. The remaining contribution may continue to propagate resulting in the refraction phenomenon.

When sound reflects on an solid surface, two types of acoustic reflections can occur: part of the sound energy

- is reflected *specularly*, i. e., the angle of incidence equals the angle of reflection; and
- is reflected *diffusely* - or *scattered*, i. e., scatter in every direction).

All the phenomena occur with different proportions depending on the acoustics and geometrical properties of surfaces and the frequency content of the wave. In acoustics, it is common to define the *operating points* and different *regimes*, e. g. for instance near- vs. far-field, according to the sound frequencies or the corresponding *wavelength*,

$$\lambda = \frac{2\pi}{k} = \frac{c}{f} \quad [\mathrm{m}], \tag{2.10}$$

where f is the frequency of the sound wave.

As depicted in Figure 2.2, λ measures the spatial distance between two points around which the medium has the same value of pressure.
Using this quantity we can identify the following three responses of objects (irregularities) of size d to a plane-wave, as depicted in Figure 2.6

- $\lambda \gg d$, the irregularities are negligible and the sound wave reflection is of specular type;
- $\lambda \approx d$, the irregularities break the sound wave which is reflected towards every direction;
- $\lambda \ll d$, each irregularities is a surface reflecting specularly the sound waves.

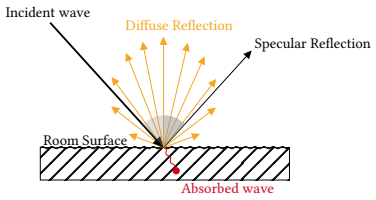


FIGURE 2.5: Specular and diffuse reflection.

" Sabine had previously used ray-based acoustics in the early 1900s to investigate sound propagation paths using Schlieren photography. Their impressive visualizations show wavefronts that are augmented with rays that are perpendicular to the wavefronts."
—[Savioja and Svensson 2015]

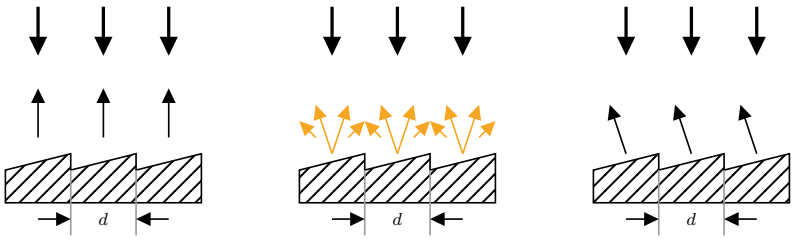
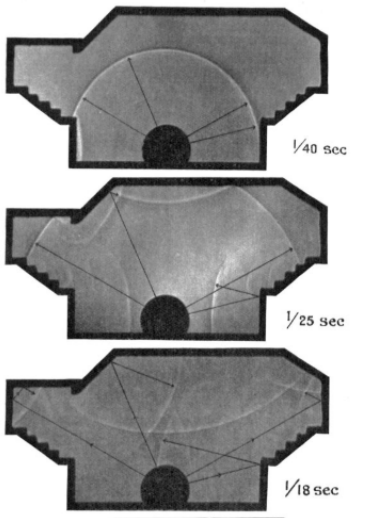


FIGURE 2.6: A reflector having irregularities on its surface with width d much smaller than the sound wavelength λ. Image courtesy of [Kuttruff 2016].

This presented behavior can be described with the wave equation by imposing adequate boundary conditions. A simplified yet effective approach - just as in optics - is to model incoming sound waves as *acoustic rays* [Davis and Fleming 1926; Krokstad et al. 1968]. A ray has well-defined direction and velocity of propagation, and conveys a total wave energy which remains constant. This simplified description undergoes with the name of Geometrical (room) acoustics (GA) [Savioja and Svensson 2015], and share many fundamentals with geometrical optics. This model will be convenient to describe and visualize the reflection behavior hereafter.



Photographs showing successive stages in the progress of a sound pulse in a section of a Debating Chamber. Image courtesy of [Davis and Fleming 1926]

2.2.1 *Large smooth surfaces, absorption and echoes*

Specular reflections are generated by surfaces which can be modelled as infinite, flat, smooth and rigid. As mentioned above, this assumption is valid as long as the surface has dimension much larger than the sound wavelength. Here the acoustic ray is reflected according to the *law of reflection*, stating that (i) the reflected ray remains in the plane identified by the incident ray and the normal to the surface, and (ii) the angles of the incident and reflected rays with the normal are equal.

If the surface S is not perfectly rigid or impenetrable, its behavior is described by the *acoustic impedance*, $Z_S(f) \in \mathbb{C}$. Analytically, it is defined as a relation between sound pressure and particle velocity at the boundary. It consists of a real and imaginary part, called respectively acoustic *resistance* and *reactance*. The former can be seen as the part of the energy which is lost, and the latter as the part which is stored.

- THE REFLECTION COEFFICIENT β can be derived from the acoustic impedance for plane waves, i. e. under assuming a far-field regime between source, receiver and surface.

  > *It measures the portion of energy absorbed by the surface and the incident acoustic wave.*

  Analytically, it is defined as [Kuttruff 2016; Pierce 2019]

$$\beta(f,\theta) = \frac{Z_S(f)\cos\theta - Z_{\text{air}}(f)}{Z_S(f)\cos\theta + Z_{\text{air}}(f)}, \tag{2.11}$$

  where $Z_S(f)$ and $Z_{\text{air}}(f)$ are the frequency-dependent impedance of the surface and the air respectively, and θ is the angle of incidence.

- THE ABSORPTION COEFFICIENT is typically used instead in the context of GA and audio signal processing. It comes from the following approximations [Savioja and Svensson 2015]: (*i*) the energy or intensity of the plane wave[10], is considered instead of the acoustic pressure; (*ii*) dependency on the angle of incidence is relaxed in favor of the averaged quantities; (*iii*) local dependency on frequencies is relaxed in favor of a frequency-independent scalar or at most a description per octave-band. These assumptions are motivated by the difficulty of measuring the acoustic impedance and the possibility to compute an equivalent coefficient a posteriori

[10] Since it is the square magnitude of the acoustic pressure, the phase information is lost.

Therefore, it is customary to use the absorption coefficient, defined as

$$\alpha(f) = 1 - \left|\bar{\beta}(f)\right|^2, \tag{2.12}$$

where $\bar{\beta}$ is the reflection coefficient averaged over the angles θ.

▶ ECHOES ARE SPECULAR REFLECTIONS which stand out in terms of energy strength or timing. Originally this term is used to refer to sound reflections which are subjectively noticeable as a separated repetition of the original sound signal. These can be heard consciously in outdoor scenario, such as in mountain. However, they are less noticeable to the listener in close rooms. In § 2.3.1 a proper definition of echoes will be given with respect to the temporal distribution of the acoustic reflections.

The word echo derives from the Greek *echos*, literally "sound". In the folk story of Greek, Echo is a mountain nymph whose ability to speak was cursed: she only able to repeat the last words anyone spoke to her.

2.2.2 *Diffusion, scattering and diffraction of sound*

Real-world surfaces are not ideally flat and smooth; they are rough and uneven. Examples of such surfaces are coffered ceilings, faceted walls, raw brick walls as well as the entire audience area of a concert hall. When such irregularities are in the same order as the sound wavelength, *diffuse reflections* is observed.

In the context of GA, the acoustic ray associated to a plane-wave can be though of as a bundle of rays traveling in parallel. When it strikes such a surface, each individual rays are bounced off irregularly, creating *scattering*: a number of new rays are created, uniformly distributed in the original half-space. The energy carried by each of the outgoing ray is angle dependent and it is well modeled thought the *Lambert's cosine law*, originally used to describe optical diffuse reflection.

The total amount of energy of this reflection may be computed a-priori knowing the *scattering* coefficient of the surface material. Alternatively, it can be derived a-posteriori with the *diffusion coefficient*, namely the ratio between the specularly reflected energy over the total reflected energy.

Diffraction waves occur when the sound confronts the edge of a finite surface, for instance around corners or through door openings. This effect is shown in Figure 2.8 At first the sound wave propagates spherically from the source. Once it reaches the reflector's apertures, the wave is diffracted, i. e. bended, behind it. It is interesting to note that the diffraction waves produced by the semi-infinite reflector edge allow the area that is "behind" the reflector to be reached by the propagating sound. This physical effect is exploited naturally by the human auditory system to localize sound sources.

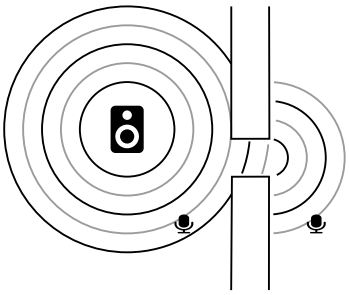


FIGURE 2.8: Schematic representation of sound diffraction. This effect allows to hear "behind walls".

2.3 ROOM ACOUSTICS AND ROOM IMPULSE RESPONSE

Room acoustics is concerned with acoustic waves propagating in air enclosed in a volumes with a set of surfaces (walls, floors, etc.), which an incident wave may be interacts with as described in § 2.2. In this context, a

A room is a physical enclosure containing the medium and with boundaries limiting the sound propagation.

Mathematically, the sound propagation is described by the wave equation (2.2). By solving it, the Acoustic Impulse Response (AIR)[11] from a source

[11] The Acoustic Transfer Function (ATF) is the Fourier transform of the AIR

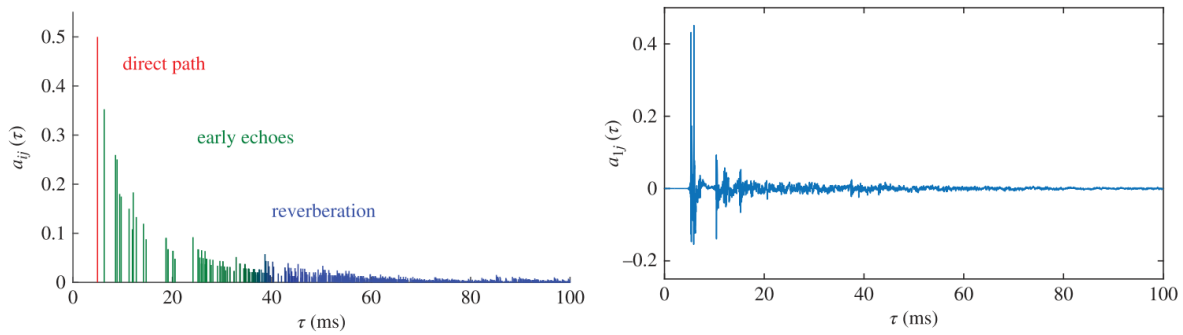


FIGURE 2.9: Schematic illustration of the shape of an RIR and the first 100 ms of a measured one.

to a microphone can be obtained. In the context of room acoustics, it is commonly referred to as the Room Impulse Response (RIR), usually stressing the geometric relation between reflections and the geometry of the scene. In this thesis the two terms will be used indistinctly.

2.3.1 *The room impulse response*

The Room Impulse Response (RIR) is where physical room acoustic and indoor audio signal processing meets and from now on, we will adopt a signal processing perspective. Therefore

the RIR as a causal time-domain filter that accounts for the whole indoor sound propagation from a source to a receiver

Figure 2.9 provides a schematic illustration of the shape of a RIR compared to a measured one. The RIRs usually exhibit common structures. Based on the consideration of § 2.2, they are commonly divided into three partially overlapped components:

$$h(t) = h^d(t) + h^e(t) + h^l(t), \tag{2.13}$$

where

- *the direct path* $h^d(t)$ is the line-of-sight contribution of the sound wave. This term coincides with the "pure delay" modeled by the free-field propagation model (2.9).
- *the acoustics echoes or early refelctions* are included in $h^e(t)$ comprising few disjoint reflections coming typically from room surfaces. They are usually characterized by sparsity in the time domain and greater prominence in amplitude. These first reflections are typically specular and are well modeled in general by the Image Source Method (ISM) explained in § 2.3.3.
- *the late reverberation*, or simply *reverberation*, $h^l(t)$ collects many reflections occurring simultaneously. This part is characterized by a diffuse sound filed with exponentially decreasing energy.

These three components are not only "visible" when plotting the RIR against time, but they are characterized by different perceptual features, as explained § 2.4.

To conclude, let $s(t)$ be the source signal. The received sound writes

$$x(t) = (h \star s)(t), \tag{2.14}$$

where the symbol $\star$ is the continuos-time convolution operator.

Apart for certain simple scenarios, computing RIRs in closed forms is a cumbersome task. Therefore numerical solvers or approximate models are used instead.

2.3.2 *Simulating room acoustics*

The docouementation of the Wayverb acoustic simulator offers a complete overview of the State of the Art (SOTA) in acoustic simulator methods [Thomas 2017].

Most of the simulators available falls in three main categories:

- *Wave-based simulators* aims at solving the wave equation numerically;
- *Geometric simulators* make some simplifying assumption about the wave propagation. They typically ignore the wave physic, instead they adopt much lighter models such as *rays* or *particles*;
- *Hybrid simulators* combining both approaches.

▶ WAVE-BASED METHODS are iterative methods that divide the 3D bounded enclosure into a grid of interconnected nodes [12]. For instance, the Finite Element Method (FEM) divides the space into small volume elements smaller than the sound wavelengths, while the Boundary Element Method (BEM) divides only the boundaries of the space into surface elements. These nodes interact with each other according to the math of the wave equation. Unfortunately, simulating higher frequencies requires denser interconnection-, so the computational complexity increases. The Finite-Difference-Time-Domain (FDTD) method replaces the derivatives with their discrete approximation, i. e. finite differences. The space is divided into a regular grid, where the changes of a quantity (air pressure or velocity) is computed over time at each grid point. Digital Waveguide Mesh (DWM) methods are a subclass of FDTD often used in acoustics problem.

[12] i. e. mechanical unit with simple degrees of freedoms, like mass-spring system or one-sample-delay unit

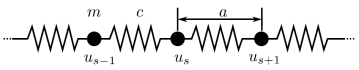


FIGURE 2.10: Example of a mass-spring linear mesh used to simulate a 1D transversal wave.

The main drawback of these methods is discretisation: less dense grids may simplify too much the simulation, while denser grids increase the computational load. Moreover, they require delicate definitions of the boundary condition at the physical level, like knowing complex impedances, which are rarely available in practice. On the other hand these methods inherently account for many effects such as occlusion, reflections, diffusion, diffractions and interferences. In particular, by simulating accurately low-frequencies components of the RIR, they are able to well characterize the *room modes* [13], namely, collections of resonances that exist in a room and characterize it.
As stated in [Välimäki et al. 2016], among the wave-based methods, the DWMs are usually preferred: they run directly in the time domain, requiring typically an easier implementation, and they exhibit a high level of parallelism.

[13] Room modes have the effect of amplifying and attenuating specific frequencies in the RIR, and produce much of the subjective sonic "colour" of a room. Their analysis and synthesis is of vital importance for evaluating acoustic of rooms, such as concert hall and recording studios or when producing musically pleasing reverbs.

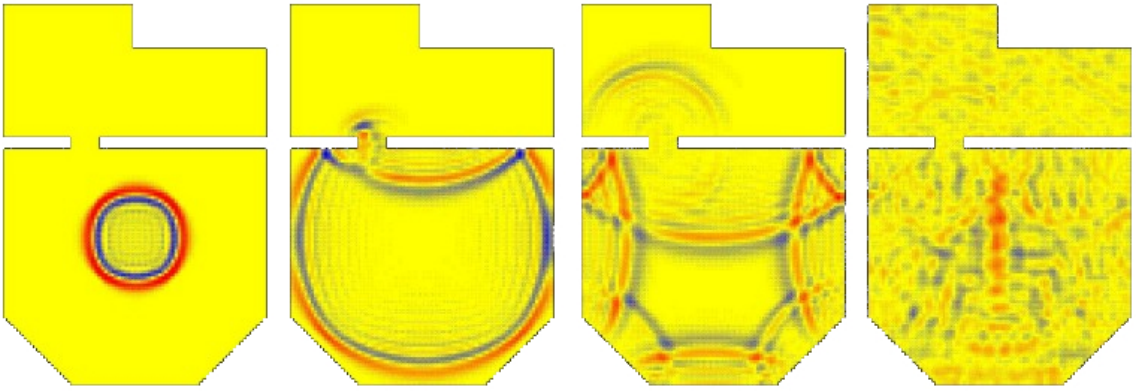


FIGURE 2.11: Simulation of Sound propagation at four consecutive timestamps using the DWM technique. A short, sharp, impulsive sound fired into the larger of two rooms causes a circular wavefront to spread out from the sound source. The the wave is reflected from the walls and part of it passes through a gap into the smaller room. In the larger room, interference effects are clearly visible; in the smaller room, the sound wave has spread out into an arc, demonstrating the effects of diffraction. A short while after the initial event, the sound energy has spread out in a much more random and complex fashion.

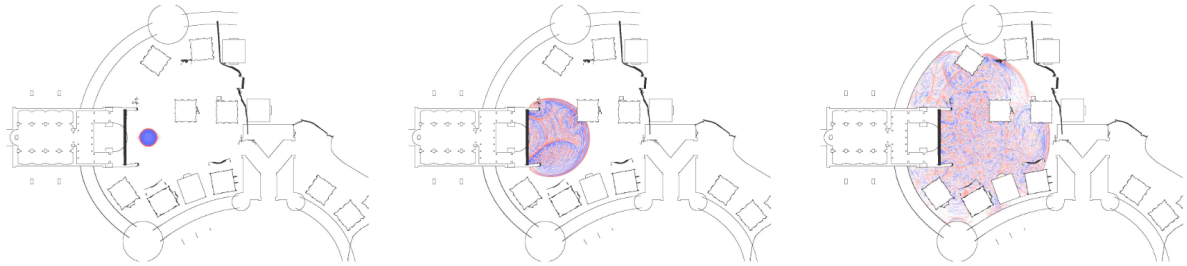


FIGURE 2.12: Sound propagation at three consecutive timestamps using the FDTD-based *Triton* simulator from Microsoft

- GEOMETRIC METODS can be sub-grouped into *stochastic* and *deterministic* approaches. They typically compute the reflection path(s) between the source and the receivers, assuming that the wave behaves like a particle or a ray carrying the acoustic energy around the scene.

  For a detailed discussion about geometric acoustic methods, please refer to [Savioja and Svensson 2015].

  STOCHASTIC METHODS are approximate by nature. They are based on statistical modeling of the RIRs or Monte Carlo simulation methods. The formers writes statistical signal processing models based on prior knowledge, such as probability distribution of the RIR in regions of the time-frequency domain [Badeau 2019]. Rather than the detailed room geometry, these methods generally use high-level descriptors[14] to synthesize RIRs and in some application are preferable. The latters randomly and repeatedly subsample the problem space, e. g. tracing the path of random reflections, recording samples which fulfil some correctness criteria, and discarding the rest. By combining the results from multiple samples, the probability of an incorrect result is reduced, and the accuracy is increased. Typically the trade-off between quality and speed of these approaches is based on the number of samples and the quality of the prior knowledge modeled.

  [14] such as the amount of reverberation or source-to-receiver distance.

  Ray-tracing [Kulowski 1985] is one the most common methods that fall in this category and is very popular in the field of computer graphics for light simula-

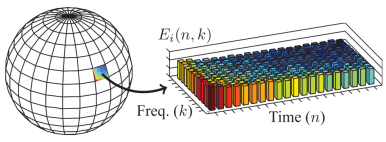


FIGURE 2.13: Directional-time-frequency Energy map resulting form the diffuse rain algorithm [Schröder et al. 2007]. For each direction, that is receiver's spherical bin, a time-frequency histogram collects the energy of incoming rays. Image courtesy of [Schimmel et al. 2009]

tion. The basic idea is to collect "valid" paths of discrete rays traced around the room. Many technique have been proposed to reduce the computational load, among which the *diffuse rain algorithm* [Schröder et al. 2007; Heinz 1993] is commonly used in many acoustic simulators. Each ray trajectory is reflected in a random direction every time it hits a wall and its energy is scaled according to the wall absorption. The process of tracing a ray is continued until the ray's energy falls below a predefined threshold. At each reflection time and for each frequency (bin or band), the ray's energy and angle of arrival are recorded in histograms, namely a *directional-time-frequency energy map* of the room's diffuse sound field for a given receiver location (Cf. Figure 2.13). This map is then used as prior distribution for drawing random sets of impulses which are used to form the RIR. While lacking a detailed description of early reflections and room modes, these methods are good to capture and simulate the statistical behavior of the diffuse sound field at a low computational cost.

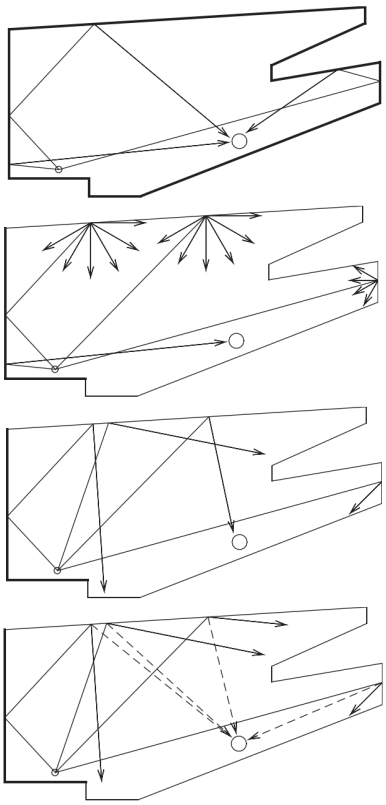


FIGURE 2.14: Visualization of ray-tracing method. From top to bottom: first the method will eventually find specular reflection; then diffuse reflections can be modeled either by splitting a ray into several new rays or a single random one. In the diffuse rain technique a shadow-ray is cast from each diffuse reflection point to the receiver to speed-up convergence of the simulation. Image courtesy of [Savioja and Svensson 2015]

DETERMINISTIC METHODS are good to simulate early reflections instead: they accurately trace the exact direction and the timing of the main reflections' paths. The most popular method is the Image Source Method (ISM), proposed by Allen and Berkley in [Allen and Berkley 1979]. Even if the basic idea is rather simple, the model is able to produce the exact solution to the wave equation for a 3D shoebox with rigid walls. It models only specular reflections, ignoring diffuse and diffracted components. it only approximates arbitrary enclosures and the late diffuse reflections.

The implementation reflects the sound source against all surfaces in the scene, resulting in a set of *image* sources. Then, each of these image sources is itself reflected against all surfaces. There are two main limitations of this method. First, in a shoebox the complexity of the algorithm is cubic in the order of reflections. Therefore when an high order is required, the algorithm becomes impractical. Second it models only the specular reflection, neglecting the diffuse sound field. For these reasons, the image-source method is generally combined with a stochastic method in hybrid methods to model the full impulse response.

- HYBRID METHODS combines the best of these two approaches. As discussed above, the image-source method is accurate for early reflections, but slow and not accurate for longer responses. The ray tracing method is by nature an approximation, but produces acceptable responses for diffuse fields. And in general geometric methods fail to properly model lower frequencies and room modes. The waveguide method models physical phenomena better than geometric methods, but is expensive at high frequencies. All these limitations correspond to three regions in the Time-Frequency (TF) representation of the RIR. As depicted in Figure 2.15,

    - in the time domain, a transition can be identified between the early vs. late reflection, corresponding to the validity of the deterministic vs. stochastic models; and
    - in the frequency domain, between geometric vs. wave-based modeling.

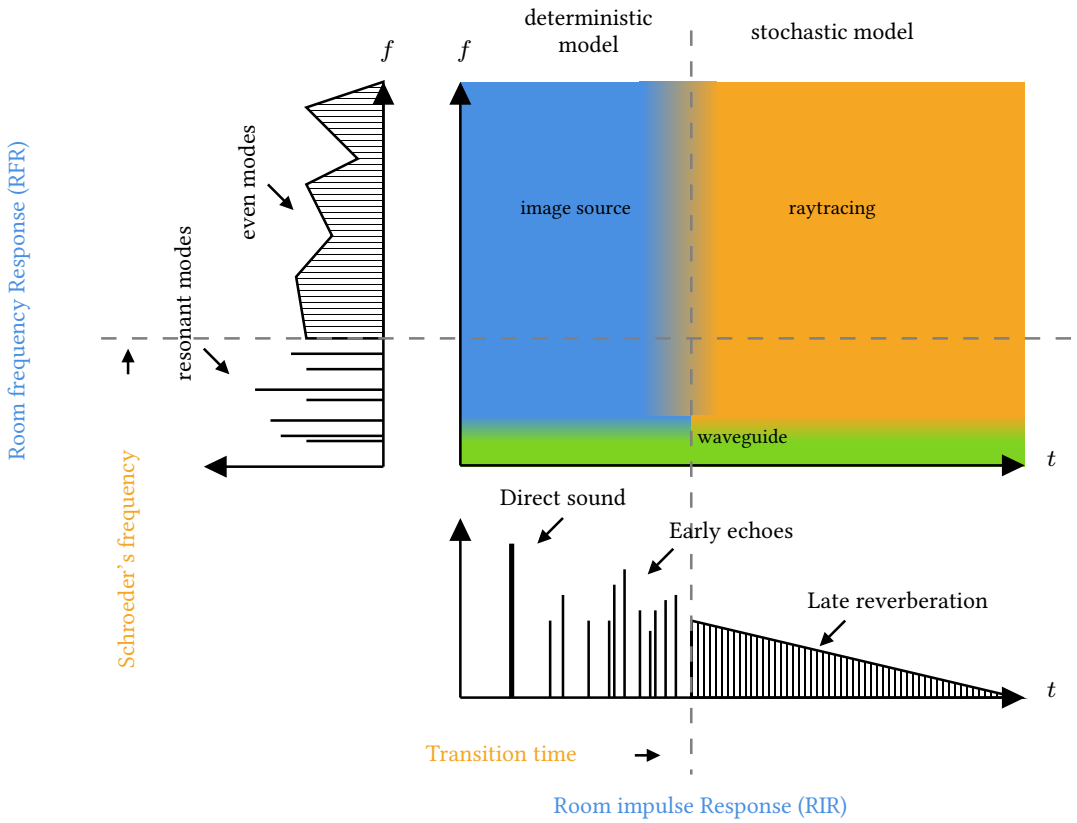


FIGURE 2.15: Time-Frequency regions of the RIR associated to the method that better simulate them. Image adapted from [Thomas 2017; Badeau 2019].

By combining three methods, accurate broadband impulse responses can be synthesized. However, this is possible provided that the time- and frequency-domain *crossover points* are respected and the level of each component is scaled accordingly [Badeau 2019]. The *transition time*, or *mixing time*, identifies the moment after which reflections are so frequent that they form a continuum and, because the sound is partially absorbed by the room surfaces at every reflection, the sound level decays exponentially over time. This point define the cross-fade between the deterministic and the stochastic process The crossover point in the frequency domain is called *Schroeder's frequency* and it split the spectrum of the RIR into a region with a few isolated modes and one denser, called respectively the *resonant* and *even* behaviors. This point define the cross-fade between the geometrical and wave-based model.

Each simulator available has its own way to compute and implement this crossover points as well as mixing the results of the three methods.

2.3.3 *The method of images and the image source model*

The *Method of Images* is a mathematical tool for solving a certain class of differential equations subjected to boundary conditions. By assuming the presence of a "mirrored" source, certain boundary conditions are verified facilitating the solution of the original problem. This methods is widely used in many fields of physics, and interestingly with specific applications to Green's functions. Its application to acoustic was originally proposed by Allen and

Berkley in [Allen and Berkley 1979] and it is know as the Image Source Method (ISM). Now ISM is probably the most used technique for deterministic RIR simulation due to its conceptual simplicity and its flexibility.

The ISM is based on purely specular reflection and it assumes that the sound energy travels around a scene in "rays". In the appendix of [Allen and Berkley 1979], the authors also proved that this method produces a solution the Helmholtz's equation for cuboid enclosures with rigid boundaries.

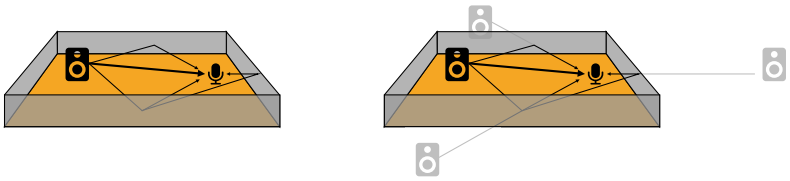


FIGURE 2.16: 3-D representation of the Image Source Method (ISM) and its propagation paths for selected echoes.

The image source defines the interaction of the propagating sound and the surface. It is based on the observation that when a ray is reflected, it spawns a secondary source "behind" the boundary surface.

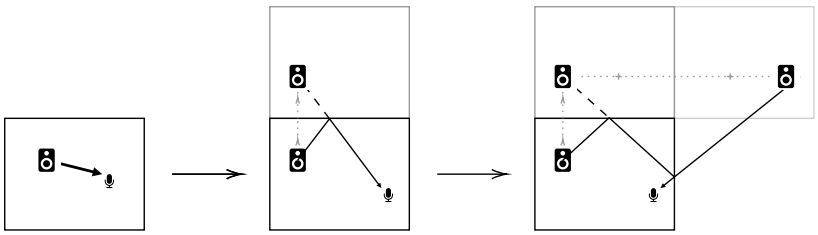


FIGURE 2.17: From left to right, path involving the direct path, one reflection obtained using first-order image, and two reflections obtained using two images. Image inspired from [Habets 2006].

As show in Figure 2.17, this additional source is located on a line perpendicular to the wall, at the same distance from it as the original source, as if the original source had been "mirrored" in the surface. In this way, each wavefront that arrives to the receiver from each reflection off the walls corresponds to the direct path received from an equivalent (or image) source.
The ISM makes use of the following assumptions:

- sound source and receiver are points in a cuboid enclosure;
- purely specular reflection paths between a source and a receiver;
- this process is simplified by assuming that sound propagates only along straight lines or rays; and
- rays are perfectly reflected at boundaries

Finally the RIR is found by summing the contribution from each (image) source, delayed and attenuated appropriately depending on their distance from the receiver. Therefore, in the time domain, the RIR associated to the source at position $\underline{\mathbf{s}}$ and the receiver at $\underline{\mathbf{x}}$ reads

$$h_{\mathrm{ISM}}(t, \underline{\mathbf{x}} \mid \underline{\mathbf{s}}) = \sum_{r=0}^{R} \frac{\bar{\alpha}_r}{4\pi \|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|} \delta\left(t - \frac{\|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|}{c}\right) \tag{2.15}$$

where $\underline{\mathbf{s}}_r$ is the r-th image of the source and $\bar{\alpha}_r$ is the total frequency-independent[15] damping coefficient related to the r-th image. Such coefficient accounts for all the dissipation effects encountered in the reflection path, e. g. absorption, air attention and scattering. In the original formulation of the ISM, $\bar{\alpha}_0 = 1$ is assumed for the direct propagation; while for the first order images, it coincides with the frequency-independent surface absorption coefficient of the surface. For the subsequent orders of images, the product of all the coefficient of the surfaces encounters in the reflection path is considered.

[15] Which is equivalent to consider perfectly rigid and reflective walls

In order to easily incorporate frequency-dependent damping effects, the Fourier transform of Eq. (2.15) is considered instead, where each reflection term is appropriately scaled

$$H_{\text{ISM}}(f, \underline{\mathbf{x}} \mid \underline{\mathbf{s}}) = \sum_{r=0}^{R} \frac{\alpha_r(f)}{4\pi \|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|} \exp\left(-\mathrm{i}2\pi f \frac{\|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|}{c}\right), \tag{2.16}$$

where now the r-th damping coefficient α_r is frequency dependent. Notice that now the damping coefficients correspond to filters, requiring Eq. (2.15) to be written as sum of convolutions. This have a strong implication when modeling and estimating the RIRs as stream of Dirac function. Ideally they consists of scaled Diracs with well defined time locations. The probability that two or more Diracs arrive at the same time is then very small. However, if we now assume that each reflection has a non-flat frequency response, filters are observed in the time domain. Such filters have arbitrary long time-domain description and now the probability that two or more overlap is much higher.

Moreover the reader should notice that the summation in the echo models of Eq. (2.15) and **??** induce an "order" among reflections indexed by r. Reflections are usually sorted for increasing Time of Arrival (TOA), $\tau_r = \|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|/c$, or decreasing amplitudes, $\bar{\alpha}_r/(4\pi\|\underline{\mathbf{x}} - \underline{\mathbf{s}}_r\|)$. Alternatively, one can sort them according to their "image" generation, e. g. direct path, first-, second-order images etc. This would require an arbitrary order within the same generation, based typically on arbitrary wall sequence. Notice that the resulting sorted sequences can differ substantially as show in **??**. This translates into non trivial definition of evaluation metrics for the task of estimating echoes.

? CAN ECHOES BE LOUDER THAN THE DIRECT-PATH? Yes, in certain cases reflections maybe carry energy comparable or stronger than the direct contribution. This happens for instance when directional sources are directed towards reflectors or when multiple reflections arrive within a very short time. Typical scenarios are when a person is presenting facing the slides projected on a wall giving the shoulders to the microphones. When a person is very far from the microphones, the delay between each reflection is very small compare to

2.4 PERCEPTION AND SOME ACOUSTIC PARAMETERS

So far we have analyzed reverberation from a purely mathematical point of view. However in many applications it is important to correlate physical measurements to subjective and perceptual qualities. This will be important in order to define evaluation scenarios later in this thesis [16].

[16] ***Cite Sacks about perception***

2.4.1 *The perception of the RIR's elements*

It is commonly accepted that the RIR components defined in § 2.3.1 play rather separate roles in the perception of sound propagation.

- THE DIRECT PATH is the delayed and attenuated version of source signal itself. It coincides with the free-field sound propagation and, as we will see in Chapter 10, it reveals the direction of the source.

- THE EARLY REFLECTIONS AND ECHOES are reflections which are by nature highly correlated with to the direct sound. They convey a sense of geometry which modifies the general perception of the sound:

  - *The precedence effect* occurs when two correlated sounds are perceived as a single auditory event [Wallach et al. 1973]. This happens usually when they reach the listener with a delay within 5 ms to 40 ms. However, the perceived spatial location carried by the first-arriving sound suppressing the perceived location of the lagging sound. This allows human to accurately localize the direction of the main source, even in presence of its strong reflections.

  - *The comb filter effect* indicates the change in timbre of the perceived sound, named *coloration*. This happens when multiples reflections arrive with periodic patterns and some constructive or destructive interferes may arise. Such phenomena can be well modeled with a comb filter [Barron 1971].

  - *Apparent source width* is the audible impression of a spatially extended sound source [Griesinger 1997]. By the presence of early reflection, the perceived energy increases, providing the impression that a source sounds larger than its true size.

  - *Distance and depth perception* provides to the listener cues about the source location. While the former refers to the spatial range, the latter relates the source to the auditory scene as a whole [Kearney et al. 2012]. A fundamental cue for distance perception is the *direct-to-reverberant ratio* (DRR), i. e. the ratio between the direct path ratio and the remaining portion of the RIR. Regarding the depth perception, early reflections are the main responsible. In the context of virtual reality, correctly modeling of these quantities is essentials in order to maintain a coherent depth impression [Kearney et al. 2012].

- THE LATE REVERBERATION in room acoustics is indicative of the size the environment and the materials within [Välimäki et al. 2016]. It provides the *listener envelopment*, i. e. the degree of immersion in the sound field [Griesinger 1997]. This portion of the RIR is mainly characterized by the sound diffusion, which depends on the surfaces roughness.

2.4.2 *Mixing time*

Perceptually, it define the instant when the reverberation cannot be distinguished from that of any other position of the listener in the room. Analytically,

the the mixing time is the instant that divides the early reflections from the late reverberation in a RIR. Due to this, it is an important parameter also in the context of RIRs synthesis as it defines cross-over point for room acoustics simulator using hybrid methods [Savioja and Svensson 2015][17].

[17] Cf. § 2.3.2

2.4.3 *Reverberation time*

The *reverberation time* measures the time that takes the sound to "fade away" after it ceases. In order to quantify it, acoustics and in audio signal processing use the *Reverberation Time at 60 dB*, i. e. the $\mathtt{RT}_{60}$, the time after which the sound energy relatively dropped by 60 dB. It depends on the size and absorption level of the room (including obstacles), but not on the position of specific position of the source and the receiver. Real measurements of RIRs are affected by noise. As a consequence, it is not always possible to consider a dynamic range of 60 dB, i. e. the energy gap between the direct path and the ground noise level. In this case, the $\mathtt{RT}_{60}$ value must be approximated with other methods.

By knowing the room geometry and the surfaces acoustics profiles, it is possible to use the empirical *Sabine's equation*:

$$\mathtt{RT}_{60} \approx 0.161 \frac{V_{\text{TOT}}}{\sum_l \alpha_l S_l} \quad [\text{s}], \tag{2.17}$$

where V_{TOT} is the total volume of the room $[\text{m}^3]$ and α_l and S_l are the absorption coefficient and the area $[\text{m}^2]$ of the l-th surface.

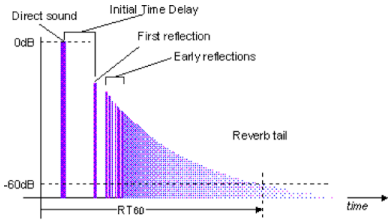


FIGURE 2.18: illustration of the Reverberation Time ($\mathtt{RT}_{60}$) definition. It. Image courtesy of wikipedia.

2.4.4 *Direct-to-Reverberant ratio and the critical distance*

The direct-to-reverberant ratio (DRR) quantifies the power of direct against indirect sound [Zahorik 2002]. It varies with the size and the absorption of the room, but also with the distance between the source and the receiver according to the curves depicted in Figure 2.19 The distance beyond which the power of indirect sound becomes larger than that of direct sound is called the *critical distance*.

These quantities represent an important parameter to assert the robustness of audio signal processing methods, since they basically measure the validity of the free-field assumption.

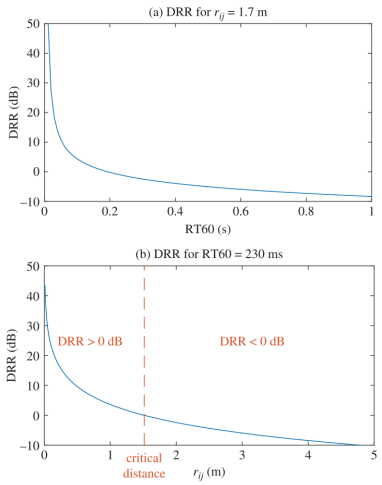


FIGURE 2.19: DRR as a function of the RT60 and the source distance rij based on Eyring's formula (Gustafsson et al., 2003). These curves assume that there is no obstacle between the source and the microphone, so that the direct path exists. The room dimensions are the same as in Figure 3.1.

3

Elements of Audio Signal Processing

"*Signal, a function that conveys information about a phenomenon.* [. . .] *Consider an acoustic wave, which can convey acoustic or music information.*"
—R. Priemer, *Introductory Signal Processing*

▸ Synopsis Let us now move from the physics to digital signal processing. At first in § 3.1, this chapter formalizes fundamental concepts of audio signal processing such as signal, mixtures and noise in the time domain. In § 3.2, we will present the signal representation that we will use throughout the entire thesis: the Short Time Fourier Transform (STFT). Finally, after assuming the narrowband approximation, in § 3.3 some important models for the Room Impulse Response (RIR) are described.
Unless specified, the notation and definitions presented in this chapter for the audio signal model are excerpted from Vincent et al.'s book *Audio source separation and speech enhancement.* The material used for illustrating concepts of digital signal processing are taken from standard book on the topics.

3.1 Signal model in the time domain

In the previous chapter we formalized the physics that rule the sound propagation from the source to the microphone. A raw *audio signal* encodes the variation of pressure over time on the microphone membrane. Mathematically it is denoted as the function

$$\begin{aligned} \tilde{x} : \mathbb{R} &\to \mathbb{R} \\ t &\mapsto \tilde{x}(t), \end{aligned} \tag{3.1}$$

continuous both in time $t \in \mathbb{R}$ and amplitudes.

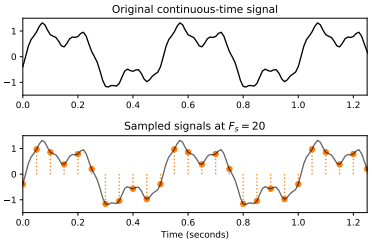


Figure 3.1: Continuous-time signal and its sampled version.

Today signals are typically processed, stored and analyzed by computers as *digital audio signal.* This corresponds to finite and discrete-time signal x_n obtained by periodically sampling the continuous-time signal $\tilde{x}$ at rate F_s [Hz], truncate it to n samples. As common to most measurement models, we assume that the sampling process involves two steps: first, the impinging signal undergoes an ideal low-pass filter $\tilde{\phi}_{\text{LP}}$ with frequency support in $]-F_s/2, F_s/2]$[18]; then it time-support is regularly discretized, $t = n/F_s$ for $n \in \mathbb{Z}$. This is expressed by

$$\hat{x}[n] = \left(\tilde{\phi}_{\text{LP}} \star \tilde{x}\right)\left(\frac{n}{F_s}\right) \in \mathbb{R}, \tag{3.2}$$

where $\star$ is the continuous-time convolution operator. This will restrict the frequency support of signal to satisfy the *Nyquist–Shannon sampling theorem* and avoid aliasing effect.

Strickly speaking, the digital representation of a continuos signal involves sampling and quantization. In this thesis we assume the sampled signals are real-valued, ignoring the *quantization* process.

[18] The ideal low-pass filter is $\tilde{\phi}_{\text{LP}}(t) = \text{sinc}(t) = \sin(\pi F_s t)/(\pi F_s t)$. The term sinc stands for *sinus cardinal* and was introduced by Philip M. Woodward in 1952 in [Woodward and Davies 1952], in which he said that the function "*occurs so often in Fourier analysis and its applications that it does seem to merit some notation of its own*"

Finally, at the end of the discretisation process, the $\tilde{x}(t)$ is represented as the finite time series or a vector,

$$\hat{x}_N \in \mathbb{R}^N, \tag{3.3}$$

with entries $\hat{x}_N[n]$ for $n = 0, \ldots, N-1$.

The choice of F_s depends on the application since it is a trade-off between computational power, processing and rendering quality. Historically the two iconic values are $44.1\,\mathrm{kHz}$ for music distribution on CDs and $8\,\mathrm{kHz}$ for first-generation speech communication. Now multiples of $8\,\mathrm{kHz}$ are typically used in audio processing: $(16, 48, 96, 128\,\mathrm{kHz})$.

Audio signals are emitted by sources and are observed, received or recorded by microphones. A set of microphones is called a microphone *array*, whose signals are sometime referred to as *channels*. In this thesis, these objects are assumed to have been deployed in a indoor environment, called generically *room*. Let us provide some taxonomy, through some dichotomies, useful for describe the mixing process later:

⇄ Sources vs. Mixtures: Sound sources emits sounds. When multiple sources are active at the same time, the sounds that reach our ears or are recorded by microphones are superimposed or *mixed* into a single sound. This resulting signal is denoted as *mixture*.

⇄ Single-Channel vs. Multichannel: The term *channel* is used here to indicate the output of one microphones or one source. A *single-channel* signal ($I = 1$) is represented by the scalar $\tilde{x}(t) \in \mathbb{R}$, while a *multichannel* ($I > 1$) is represented by the vector $\tilde{\boldsymbol{x}}(t) = [\tilde{x}_1, \ldots, \tilde{x}_I]^\intercal \in \mathbb{R}^I$.

⇄ Point vs. Diffuse Sources: *Point sources* are single and well-defined points in the space emitting single-channel signal. In certain application, human speakers or the sound emitted by a loudspeaker can be reasonably modeled as in this way.
Diffuse sources refers for instance to wind, traffic noise, or large musical instruments, which emit sound in a large region of space. Their sound cannot be associate to a punctual source, but rather a distributed collection of them.

⇄ Directional vs. Onmidirectional: An *omnidirectional* source (resp. receiver) will in principle emit (resp. record) sound equally from all directions, both in time and in frequency. Although this greatly simplifies processing models and frameworks, this is not true in real scenario. The physical properties of real sources (resp. receivers) leads to *directivity patterns*, a. k. a. *polarity*, which may be different at different frequencies. In this thesis we will assume omnidirectional sources and receivers.

3.1.1 *The mixing process*

Let us assume the observed signal has I *channels* indexed by $i \in \{1, \ldots, I\}$. Let us assume that there are J sources indexed by $j \in \{1, \ldots, J\}$. Each

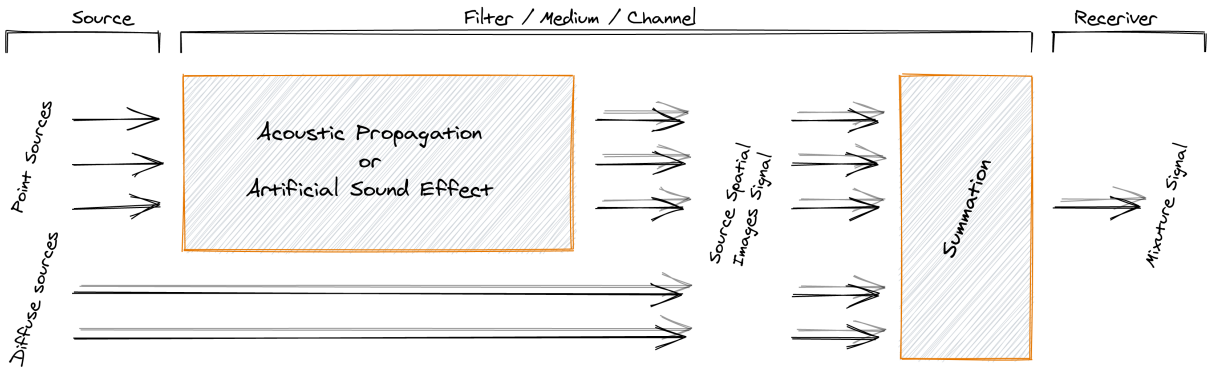


FIGURE 3.2: General mixing process, illustrated in the case of $J = 3$ sources, including three point sources and one diffuse source, and $I = 2$ channels.

microphone i and each source j have a well defined position in the space, $\underline{\mathbf{x}}_i$, $\underline{\mathbf{s}}_j$, respectively.

The mixing process describes then the nature of the mixtures. In order to better formalized it, the authors of [Sturmel et al. 2012] introduced the intermediate representation called *source spatial images*: $\tilde{c}_{ij}(t)$ describes the contribution of the source j to the microphone i. Consequently, the *mixture* $\tilde{x}_j$ is the combination of images associated to the source j. Depending on the "contribution" the image describes, the following type of mixture can be defined:

⇄ NATURAL VS. ARTIFICIAL MIXTURES: The former refers to microphone mixtures recorded simultaneously the same auditory scene, e. g. teleconferencing systems or hands-free devices. By contrast, the latters are created by mixing together different individual, possibly processed, recordings. This are the typical mixtures used professional music production where the usage of long-chain of audio effects typically "hide", willingly or not, the recording environment of the sound sources.

⇄ INSTANTANEOUS VS. CONVOLUTIVE MIXTURES: In the first case, the mixing process boils down to a simple linear combination of the source signals, namely the mixing filters are just scalar factors. This is the typical scenario when sources are mixed using a mixing console. Convolutive mixtures, instead, denote the more general case where the each mixture is the sum of filtered signals. In between are the *anechoic* mixtures involving the sum of scaled and delayed source signals. Natural mixtures are convolutive by nature and ideal free-far-field natural recording are well approximated by anechoic mixtures.

▶ IN THIS THESIS, we will particularly focus on natural mixture: the microphone mixture listens to the propagation of sound in the room and this process is linear (Cf. § 2.1) and time invariant provided a static scenario. Therefore, the resulting mixture is the simple summation of the sound images, which are the collections of convolution between the RIRs and source signal:

instantaneous	$\tilde{c}_i j = a_{ij}\tilde{s}(t)$
anechoic	$\tilde{c}_i j = a_{ij}\tilde{s}_j(t - \tau_{ij})$
convolutive	$\tilde{c}_i j = (\tilde{g}_{ij} \star \tilde{s}_j)(t)$

TABLE 3.1: Taxonomy of linear mixing models for a mixture channel x_i, sources s_j, impulse response $\tilde{g}_{ij}$, scaling factor a_{ij} and delay τ_{ij}.

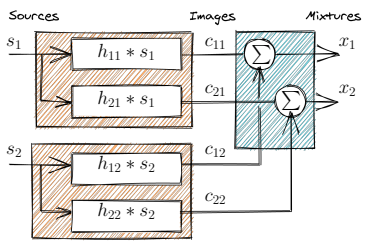


FIGURE 3.3: Graphical representation of the mixing model 3.5 for 2 sources and 2 microphones.

$$\tilde{c}_{ij}(t) = \Big(\tilde{h}_{ij} \star \tilde{s}_j\Big)(t) \tag{3.4}$$

$$\tilde{\boldsymbol{c}}_j(t) = [\tilde{c}_{1j}(t), \dots, \tilde{c}_{Ij}(t)]^\intercal$$

$$\tilde{\boldsymbol{x}}(t) = \sum_{j=1}^{J} \tilde{\boldsymbol{c}}_j(t). \tag{3.5}$$

Considering the time domain description of the RIR derived (and approximated) in the previous chapter, the time-domain *mixing filters* $\tilde{h}_{ij}(t)$ will be modeled as follows:

$$\tilde{h}_{ij}(t) = \sum_{r=0}^{R} \frac{\alpha_{ij}^r}{4\pi c \tau_{ij}^r} \delta\big(t - \tau_{ij}^r\big) + \varepsilon_{ij}(t) \tag{3.6}$$

where $\alpha_{ij}^r \in \mathbb{R}$ and $\tau_{ij}^r \in \mathbb{R}$ are the attenuation coefficient and the time delay of the reflection r. The noise term $\tilde{\varepsilon}_{ij}(t)$ collects later echoes ($r > R$) and the tail of the reverberation. We do not assume $\tilde{\varepsilon}_{ij}(t)$ to be known.

3.1.2 *Noise, interferer and errors*

In Eq. (3.5) no noise is included: all the sources are threated in the same way, including *target*, *interfering* and *noise* sources. While the definition of target sound source is quite self-explanatory and it will denoted by default as the first source, that is $j = 1$, the term interfer and noise depends on the specific use case, problem, application, and research field. Notice that in Eq. (3.6) a noise term is added to gather unknown quantities.

> *Noise is a general term for unwanted (and, in general, unknown) modifications that a signal may suffer during capture, storage, transmission, processing, or conversion* [Tuzlukov 2018].

Therefore, we will define and use the following type of noises:

- INTERFERS identifies the undesired source with properties similar to the target source. For instance, a concurrent speech source for speech application or concurrent music instrument in case of music.
  Later, in this thesis the interfer sources will be denoted as additional source indexed by $j > 1$.

- NOISE collects all the remaining effects, typically nonspeech sources. Moreover we will make a further distinction between the followings.

- DIFFUSE NOISE FIELD describes the background diffuse sources present in the auditory scene, e. g. car noise, indistinct talking or winds. It can be recorded or approximated as Additive White Gaussian Noise (AWGN) with a specific spatial description as described in [Habets and Gannot 2007].

- MEASUREMENT AND MODEL NOISE accounts for general residual miss- and under-modeling error. As common is signal processing and information theory, this error term will be modeled as AWGN.
  In this thesis, it will denoted as $\tilde{\varepsilon}_{ij}(t)$ and will be used to model the approximation of the RIR with the ISM or sensor noise, respectively.

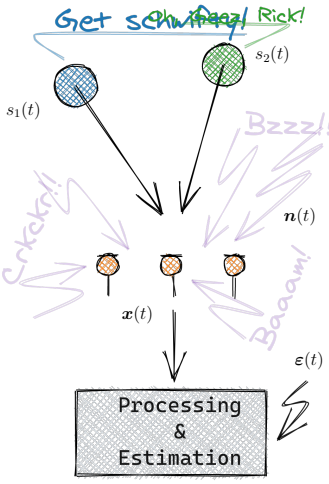


FIGURE 3.4: Graphical representation of the mixing model (3.5): $s_2(t)$ is the *interferer*, $\boldsymbol{n}(t)$ contributes to the *diffuse noise field*, and $\boldsymbol{\varepsilon}(t)$ model acquisition and modeling errors.

By making the noisy terms explicit, the mixing model in Eqs. (3.4) and (3.5) writes:

$$\tilde{c}_{ij}(t) = \left(\tilde{h}_{ij} \star \tilde{s}_j\right)(t) + \tilde{\varepsilon}_{ij}(t) \tag{3.7}$$

$$\tilde{\boldsymbol{c}}_j(t) = [\tilde{c}_{1j}(t), \dots, \tilde{c}_{Ij}(t)]^\intercal$$

$$\tilde{\boldsymbol{x}}(t) = \sum_{j=1}^{J} \tilde{\boldsymbol{c}}_j(t) + \tilde{\boldsymbol{n}}(t) \tag{3.8}$$

3.2 Signal model in the spectral domain

The frequency, or spectral, representation is probably the most famous signal representation used in signal processing: Speech and music signals naturally exhibit harmonic and periodic behaviors and through it are described as combination of sinusoids as function of their frequencies.

It was introduced by Joseph Fourier in his work on the heat equation [Fourier 1822]. His mathematical tool, named later *Fourier Decomposition*, aims at approximating any signal by a sum of sine and cosine waves.

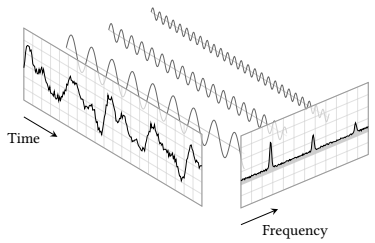


FIGURE 3.5: A signals resolved into its Fourier series: a linear combination of sines and cosines represented as peaks in the frequency domain.

This operation is achieved by the Fourier Transform (FT), $\mathcal{F} : \mathbb{R} \mapsto \mathbb{C}$, which projects a continuous-time-domain signal $\tilde{x}$ onto a space spanned by continuous-frequency complex exponentials:

$$\tilde{X}(f) = (\mathcal{F}\tilde{x})(f) = \int_{-\infty}^{+\infty} \tilde{x}(t) \mathrm{e}^{-\mathrm{i}2\pi f t} \, \mathrm{d}t, \tag{3.9}$$

where $f \in \mathbb{R}$ are the *natural frequency* in Hz and i is the imaginary unit.

A part from providing a space where audio signal reveals their harmonic structures, the Fourier transforms benefits of two fundamental properties: it is linear and it converts time-convolution into element products.
First, linearity allows to write Eq. (3.5) simply as:

$$\tilde{\boldsymbol{x}}(t) = \sum_{j=1}^{J} \tilde{\boldsymbol{c}}_j(t) \quad \overset{\mathcal{F}}{\to} \quad \tilde{\boldsymbol{X}}(f) = \sum_{j=1}^{J} \tilde{\boldsymbol{C}}_j(f) \tag{3.10}$$

Secondly, by the *convolution theorem*, the source spatial images in Eq. (3.4) writes as:

$$\tilde{c}_{ij}(t) = \left(\tilde{h}_{ij} \star \tilde{s}_j\right)(t) \quad \overset{\mathcal{F}}{\to} \quad \tilde{C}_{ij}(f) = \tilde{H}_{ij}(f)\tilde{S}_j(f). \tag{3.11}$$

As discussed in **??**, the FT of a RIR, a. k. a. the Room Transfer Function (RTF), can be computed exactly in closed-form as

$$\tilde{H}_{ij}(f) = \sum_{r=0}^{R} \alpha_{ij}^r \mathrm{e}^{-\mathrm{i}2\pi f \tau_{ij}^r}. \tag{3.12}$$

In practice, the filters $\tilde{h}_{ij}$ are not available in the continuous time domain nor in the continuous frequency domain directly. They must be estimated from the observation of the discrete-time mixtures $\hat{x}_i[n]$, therefore, after the convolution with a source and the measurement process. In practice, we don't have access to continuous signal, neither is time and in frequency domain. Every signal or spectrum the microphones capture are represented by finite- and discrete time signals for which the properties (3.11) are valid with some precautions.

3.2.1 *Discrete time and frequency domains*

The spectral representation of a discrete-time signal, $x[n]$ with $n \in \mathbb{Z}$, is given by the (forward) Discrete-Time Fourier Transform (DTFT), $\mathcal{F}_{\mathrm{F_s}}$:

$$\tilde{X}_{F_s}(f) = (\mathcal{F}_{\mathrm{F_s}}\, x)(f) = \sum_{n=-\infty}^{+\infty} x[n]\mathrm{e}^{-\mathrm{i}2\pi f n/Fs}, \tag{3.13}$$

which is a continuos function of f with period F_s. Notice that the term discrete-time refers to the fact that the transform operates on discrete signal. When this samples are uniformly spaced at rate F_s, it produces a function of continuous frequency that is a periodic summation of the continuous Fourier transform of the original continuous function. Under certain theoretical conditions, described by the *sampling theorem*, both the original continuous signal $\tilde{x}$ and its sampled version $\hat{x}$ can be recovered perfectly from the DTFT. The DTFT itself is a continuous function of frequency which require infinite discrete value to be computed. For these two reason, it is not accessible in practice or computed in the digital domain. Therefore the following representation is used instead.

The spectral representation of a discrete- and finite-time signal $\hat{x}_N$ is given by its (forward) Discrete Fourier Transform (DFT) [19], $\mathbf{F} : \mathbb{R}^N \mapsto \mathbb{C}$:

$$\hat{X}_F[k] = (\mathbf{F}\,\hat{x}_N)[k] = \sum_{n=0}^{N-1} \hat{x}_N[n]\mathrm{e}^{-\mathrm{i}2\pi k n/F}. \tag{3.14}$$

where $k \in [0, F-1]$ is the discrete *frequency bin* and F is the total number of bins. Again we use the subscript F and the brackets $[k]$ to stress the finite and discrete frequency support of the DFT.
The natural frequency f_k in Hz corresponding to the k-th frequency bin can be computed as

$$f_k = \frac{k}{F} F_s. \tag{3.15}$$

[19] This can be intrepreded as the projection onto the space spanned by a finite number of complex exponentials.

3.2.2 *The DFT as approximation of the FT*

An important application of the DFT is to approximate numerically the FT. As mentioned at the beginning of the chapter, with the discretisation process the continuous signal is periodically sampled, low-passed and finally truncated. It can be proved that sampling in the time domain corresponds to limiting the signal bandwidth and periodizing the spectrum.
By assuming sampling at rate F_s, in the continuos-frequency domain the spectrum $\tilde{X}(f)$ is repeated every internals of size F_s Hz . By further assuming that the signal undergoes an ideal low-pass filter, no spectral leakage is presents between each repetition.
So far, the sampled time domain signal, $\hat{x}[n]$, is mapped to the continuous frequency domain $\tilde{X}(f)$. This particular case of the FT is called Discrete-Time Fourier Transform (DTFT) and it is denote with $\tilde{X}_{F_s}[k]$.

$$\tilde{X}(f) = \int_{-\infty}^{+\infty} \tilde{x}(t)\mathrm{e}^{-\mathrm{i}2\pi f t}\,\mathrm{d}t \quad \rightarrow \quad \tilde{X}_{F_s}(f) = \sum_{n=-\infty}^{\infty} \hat{x}[n]\mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}}. \tag{3.16}$$

Here the continuous integral the FT is approximated by Riemann sum over the discrete points $n \in \mathbb{Z}$: To be more rigorous, when computing a Riemann

sum approximation, the length of the discretisation interval multiply the summation. In our application, this quantity always set to F_s and for readability reason such term is dropped.
The quality of this approximation w. r. t. the original continuous spectrum is regulated by the choice of F_s: the higher F_s, the better the approximation. The upper bound to the possible value F_s is the results known as the Nyquist–Shannon's sampling theorem.

Furthermore, we consider only the finite sequence $\hat{x}_N$ consisting of N samples. This would reduce the summation ranges the right part of Eq. (3.17). Instead, we can keep the infinite summation by multiplying the sampled signal by a discrete-time window function $\hat{w}$ selecting the non-zero porting of $\hat{x}$, $\hat{x}_N[n] = \hat{w}[n]\hat{x}[n]$. By the *convolution theorem*, the multiplication in the time domain translates in a convolution between the corresponding spectra. As a consequence, the spectrum of the truncated signal is distorted by the spectrum of the window function. In math,

$$\tilde{X}_N(f) = \sum_{n=0}^{N-1} \hat{x}_N[n] \mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}} \quad \leftrightarrow \quad \tilde{X}_{F_s}(f) = \sum_{n=-\infty}^{\infty} \hat{x}[n]\hat{w}[n] \mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}}. \tag{3.17}$$

By the convolution theorem, we have that

$$\hat{x}_N[n] = \hat{x}[n]\hat{w}[n] \quad \leftrightarrow \quad \tilde{X}_N(f) = (\tilde{X}_{F_s} \star \tilde{W}_{F_s})(f) \tag{3.18}$$

where $\tilde{W}_{F_s}$ is the DTFT of the sampled window function $\hat{w}[n]$.
Assuming the window function to be an ideal door function[20], its DTFT is a ideal low-pass filter, which acts on the original spectrum as a smoothing function. As a consequence, the quality of this approximation is then based on the spectral leakage of the chosen window function, $w[n]$. As a rule of thumb, here the longer the segment, the better the approximation[21]

[20] door function here

[21] When short excerpt are considered instead (e. g. in case of the Short Time Fourier Transform (STFT)), particular types of window function are used but their analysis are out of the scope of this thesis.

Finally, we cannot access the DTFT directly because that involves an infinite number of frequencies $f \in \mathbb{R}$. Therefore, taking F uniformly-spaced frequency $f_k \in \mathbb{R}$ as in Eq. (3.15), we finally obtain the DFT as in Eq. (3.14), that is

$$\tilde{X}_N(f_k) = \sum_{n=0}^{N-1} \hat{x}_N[n] \mathrm{e}^{-\mathrm{i}2\pi f \frac{n}{F_s}} \quad \leftrightarrow \quad \hat{X}_F[k] = \sum_{n=0}^{N-1} \hat{x}_N[n] \mathrm{e}^{-\mathrm{i}2\pi kn/F}. \tag{3.19}$$

Notice that the F_s term disappeared in the right part of the equation above as it cancels out when using Eq. (3.15). By increasing F, we can sample more densely $\hat{X}_F[k]$ which leads to a better approximation to $\tilde{X}_N$. However this does not eliminates the distortion of the previous steps, due to $\tilde{W}_{F_s}$.

Again, we sampled a domain. Thus, according to the defined sampling process, this involve using a ideal low-pass filter. This filter acts now on the discrete spectrum, smoothing it and limiting the support of its transformation in the dual domain. Therefore, the inverse DFT of $\hat{X}_F[k]$ is not properly $\hat{x}_N[n]$, but its periodic version repeated every F samples. In fact, sampling in one of the two domain is equivalent to a periodization in the other domain while truncating

lead to convolving with a window function. Moreover, the chain of operation (sampling in time and truncation in time and sampling in frequencies) are valid in both way. Thus one can arbitrarily first sample and truncate frequency domain and finally sample in time. The only difference is in the interpretation of the windowing function, which in one case smooth the spectrum and in the other smooth the signal. All this relation and approximation that connects the FT to the DFT are well explained in explanatory material presented in [22].

[22] https://krasjet.com/rnd.wlk/poisson.pdf

3.2.3 *Signal model in the discrete Fourier domain*

Conscious of the above approximations, we can now rewrite our signal model for the discrete case. Hereafter we will always consider finite-length sequences and the index N will be dropped to lighten the notation.
The DFT is linear, so the discrete version of Eq. (3.10) becomes

$$\hat{\boldsymbol{x}}[n] = \sum_{j=1}^{J} \hat{\boldsymbol{c}}_j[n] \quad \overset{\mathbf{F}}{\rightarrow} \quad \hat{\boldsymbol{X}}[k] = \sum_{j=1}^{J} \hat{\boldsymbol{C}}_j[k] \tag{3.20}$$

Secondly, by using naïvely the discrete convolution theorem, one could translate Eq. (3.4) as

$$\hat{c}_{ij}[n] = (\hat{h}_{ij} * \hat{s})[n] \quad \overset{\mathbf{F}}{\rightarrow} \quad \hat{C}_{ij}[k] = \hat{H}_{ij}[k]\hat{S}[k], \tag{3.21}$$

where $*$ is the finite-time linear convolution operator[23].
The filter $\hat{H}_{ij}[k]$ is the DFT of the room impulse response. As mentioned in the § 3.2.2, this just approximates the RTF of Eq. (3.27). Thus we can write,

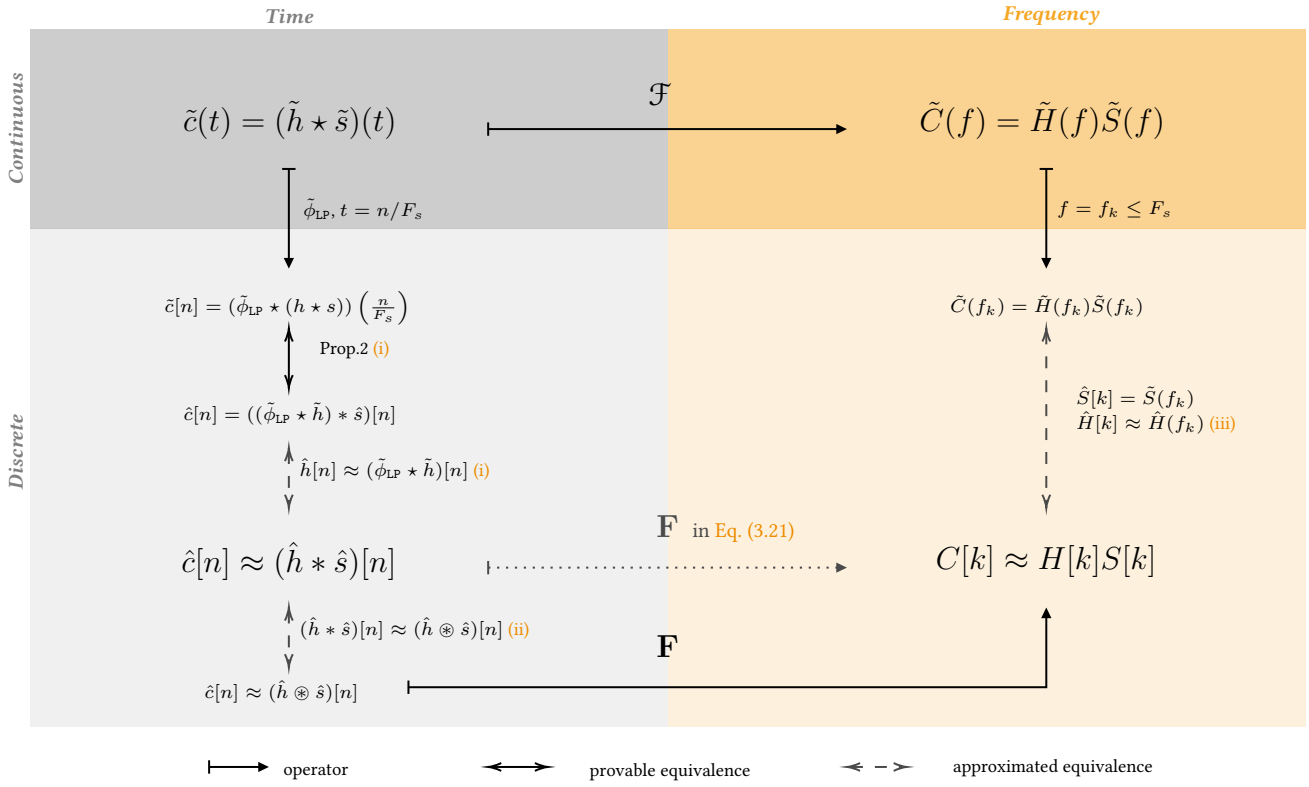
$$\hat{H}_{ij}[k] \approx \sum_{r=0}^{R} \frac{\alpha_{ij}^{r}}{4\pi c \tau_{ij}^{r}} \mathrm{e}^{-\mathrm{i}2\pi k F_s \tau_{ij}^{r}/F}. \tag{3.22}$$

[23] The finite-time linear convolution for two vectors $\hat{u} \in \mathbb{R}^L$ and $\hat{v} \in \mathbb{R}^D$ is $(\hat{u} * \hat{v})[n] = \sum_{l=0}^{L-1} \hat{u}[l]\hat{v}[L-1+n-j]$ for $n = 0, \cdots, D-L$.

Although used in practice, the model (3.21) makes use of other approximations that are worth presenting. In particular, the work by [Tukuljac et al. 2018] properly discuss them in the context of the echo estimation problem. The paper mention three approximations, which are depicted in the following diagram.

The diagram shows a chain of operators (sampling and transforms) with provable and approximated equivalences that lead to Eq. (3.21) used in practice. In order,

(i) In [van denBoomgaard and van derWeij 2001], the Proposition 2 shows that if the signal $\tilde{s}(t)$ is band-limited by F_s, then sampling the continuous convolution is exactly equivalent to *linearly convolving* the infinite discrete signal $\hat{s}[n]$ and the discrete and low-passed version of the filter. While the source signal is band-limited by nature, $\tilde{h}(t)$ is not (in fact the RIR is modeled as a summation of spikes, which has infinite spectrum). Thus, the first approximation (i) considers $\hat{h}[n] \approx (\tilde{\phi}_{\mathrm{LP}} * \tilde{h})[n]$, in words we assume that the filter is band-limited by $\pm F_s/2$.
Tukuljac et al. made an important observation here: even if infinite number of samples are available, after the measurement process, the discrete-time filter $\hat{h}[n]$ consists of infinite-length decimated combinations of sinc functions.



In the context of this thesis, this observation tell us that even in ideal conditions, that is without noise, possibly knowing the transmitted signal, and processing infinitely many samples, the exact estimation of the echo properties of the RIR is challenging task itself. This is a fundamental difference between RIR estimation and estimating the time of arrivals of the early echoes.
Note, for instance, that we wrote the echo model only in the continuous-time domain or with its closed-form form discrete frequencies. The discrete-time domain was avoided on purpose since the echoes' arrival time are naturally off the sampling grid, namely not integer multiple F_s.

(ii) The discrete-time convolution theorem applies to the *circular convolution*, which can be approximated by the *linear convolution* that is $(\hat{h} \circledast \hat{s})[n] \approx (\hat{h} * \hat{s})[n]$. This second approximation is reasonably good when many samples are available and when one of the two signals is periodic, which are typical cases for audio signals.

(iii) The third approximation regards the closed-form of $h_{ij}(f)$ of Eq. (3.22) which would require infinitely many samples and unlimited frequency support to be computed[24].

Nevertheless, it is important to notice that approximations (ii) and (iii) become arbitrarily precise as the number of samples N grows to infinity.

[24] This formula would results from the Discrete-Time Fourier Transform (DTFT) of $\tilde{h}_{ij}(t)$

While the raw audio signal encodes the amplitude of a sound as a function of time, its spectrum represents it as a function of frequency. In order to jointly

account for both temporal and spectral characteristic, joint time-frequency representations are used.

3.2.4 *Time-Frequency domain representation*

Time-Frequency (TF) representations aim to jointly describe the signal in the time and frequency domains. Instead of considering the entire signal, the main idea is to consider only a small section of the signal. To this end, one fixes a so-called *window* function, $\hat{w}_N[n]$, which is nonzero for only a period of time $L_{\texttt{win}}$ shorter than the entire signal length, $L_{\texttt{win}} \ll N$. This function iteratively shifts and multiplies the original signal, producing consecutive *frames*. Finally, the frequency information are extracted independently from each frame. The choice of a window function $w[n]$ depends on the application since its contribution reflects in the TF representation together with the one of the signal.

▶ THE DISCRETE STFT is the most commonly used TF-representation in audio signal processing. This representation encodes the time-varying spectra into a matrix $X[k,l] \in \mathbb{C}^{F,T}$ with frequency index k and time frame index l. More formally, the process to compute the complex STFT coefficients is given by

$$X[k,l] = \sum_{n=0}^{L_{\texttt{win}}-1} w[n]x[n+lL_{\texttt{hop}}]\mathrm{e}^{-\mathrm{i}2\pi kn/F} \quad \in \mathbb{C} \tag{3.23}$$

The STFT was introduced by Dennis Gabor in the 1946, the person behind Holography and Gaborlets.

where $L_{\texttt{win}}$ is the window length and $L_{\texttt{hop}}$ is the *hop size* which specifies how much the window needs to be shifted across the signal. Equivalently, Eq. (3.23) can be expressed as DFTs of windowed frames, $X[k,l] = \mathbf{F}\,\hat{x}[n,l]$ where $\hat{x}[n,l] = \hat{x}[n+lL_{\texttt{hop}}]\hat{w}[n]$.

Since each STFT coefficient $x[k,l]$ lives in the complex space $\mathbb{C}$, the squared magnitude of the STFT, $|\hat{X}[k,l]|^2$ is commonly used for visualization and for processing The resulting two-dimensional representation is called (log) *spectrogram*. It can be visualized by means of a two-dimensional image, whose axes represent time frames and frequency bins. In this image, the (log) value $|\hat{X}[k,l]|^2$ is represented by the intensity or color in the image at the coordinate $[k,l]$. Throughout this works both estimation and processing will be typically conducted in the STFT domain, unless specified. This is a common approach in the audio signal processing community, but it is not the only one: many algorithm are designed directly in the time domain or in alternatives TF representation, e. g. Mel-Scale, Filter-Banks, or the quadratic STFT transform used in **??**.

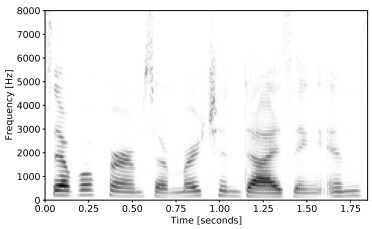
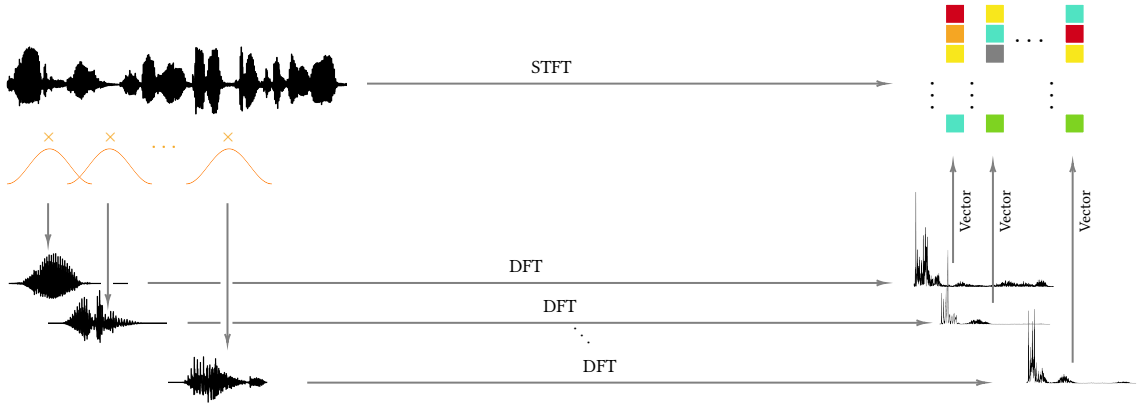


FIGURE 3.6: STFT spectrogram of an example speech signal. Higher energies are illustrated with darker colors.

As discussed [Vincent et al. 2018], the STFT has the following useful properties for audio processing:

- the frequencies f_k is a linear function of the frequency bin k;
- the resulting matrix allows easy treatment of the phase $\angle\hat{X}[k,l]$, the magnitude $|\hat{X}[k,l]|$ and the power $|\hat{X}[k,l]|^2$ separately;
- the DFT can be efficienciently computed with the Fast Fourier Transform (FFT) algorithm;
- the STFT is simple to invert;

For more mathematical detailed description of STFT and STFG can be found in [Oppenheim 1999] for audio-processing-oriented and music-processing-oriented explanation please refer to Chapter 2 of [Vincent et al. 2018] (Chapter2) and Chapter 2 of [Müller 2015], respectively.



- the STFT inherits the linearity and convolution property of the DFT under some condition about the length of the signals.

3.2.5 *The final model*

The model (3.21) shows how in practice the RIRs are treated in the frequency-domain. However this does not generalize straightforwardly to the time-frequency domain: it depends on the length of the filter w. r. t. to the length of the analysis window on of the STFT. Issues arise with "long" filters, which are common in highly reverberant or time-varying scenarios. To circumvent this issue, the *convolutional STFT* for arbitrary window functions have been proposed[25] [Gilloire and Vetterli 1992]. Although mathematically exact, it is computationally and memory intensive.

[25] It translates the time-domain convolution into inter-frame and inter-band convolutions, rather than pointwise multiplication of Fourier transforms.

In this thesis, we will assume that the filter length is shorter than the analysis window length. In the literature, this is known as the *narrowband approximation*, namely the time-domain filtering can be approximated by complex-valued multiplication in each time-frequency bin $[l, k]$:

$$\boldsymbol{C}_j[l,k] \approx \hat{\boldsymbol{H}}[k] S_j[l,k], \tag{3.24}$$

where the $\hat{\boldsymbol{H}}_j[k] = \left[\hat{h}_{1j}[k], \cdots, \hat{h}_{Ij}[k]\right]^\intercal$ is the $I \times 1$ vector of the RTFs for source j. It is sometimes practical to concatenate all these vectors into an $I \times J$ matrix $\hat{\mathbf{H}}[k] = [\boldsymbol{H}_1(f), \cdots, \boldsymbol{H}_J(f)]$ called *mixing matrix*.

With the above notation and considerations, mixing process including noise terms can be written in the STFT domain compactly as:

$$\boldsymbol{X}[l,k] = \mathbf{H}[l,k]\boldsymbol{S}[l,k] + \boldsymbol{U}[l,k] \tag{3.25}$$

where $\boldsymbol{U}[l,k] = \boldsymbol{N}[l,k] + \varepsilon(l,k)$ includes the contribution of both diffuse noise sources, modeling and measurement errors.

3.3 Other (room) impulse response spectral models

RIRs are complicated quantities to model, compute and estimate. The representations of the RIR discussed so far explicitly models early echoes and

reverberation deterministically. Furthermore, alternative models are common in the audio processing literature.

3.3.1 *Steering vector model*

In the absence of echoes and reverberation, namely assuming free-field propagation, the RIRs simplify to *steering vectors*, namely the DFT of Eq. (2.9):

$$\boldsymbol{D}_j[k] = \left[\frac{1}{4\pi q_{1j}} \mathrm{e}^{-\mathrm{i}2\pi f_k q_{1j}/c}, \cdots, \frac{1}{4\pi q_{Ij}} \mathrm{e}^{-\mathrm{i}2\pi f_k q_{Ij}/c},\right] \tag{3.26}$$

Furthermore, assuming far-field regimes, the microphone-to-source distance q_{ij} is larger than the inter-microphone distance $d_{ii'}$ making the attenuation factors $1/4\pi q_{ij}$ approximately equal, hence ignored.

3.3.2 *Relative transfer function and interchannel models*

Let us consider now only two channels and only one source signal in the model Eq. (3.25). Dropping the dependency on j for readability and taking the first channel as reference, the Relative Transfer Function (ReTF) associated to the i-th channel is defined as the element-wise ratio of the (D)FTs of the two filters [Gannot et al. 2001]

$$\hat{G}_i[k] = \frac{\hat{H}_i[k]}{\hat{H}_1[k]}. \tag{3.27}$$

The continuous-time domain counterpart is called as Relative Impulse Response (ReIR) and can be interpreted as the filter "transforming" the i-th impulse response into the one of the reference channel. Considering the noisy observation $\tilde{x}_i$ and $\tilde{x}_1$, their signals can be re-written in term of $\tilde{g}_i$ as follows

$$\begin{cases} \tilde{x}_1 = \tilde{h}_1 \star \tilde{s} + \tilde{u}_1 \\ \tilde{x}_i = \tilde{h}_i \star \tilde{s} + \tilde{u}_i \end{cases} \quad \rightarrow \quad \begin{cases} \tilde{x}_1 = \tilde{h}_1 \star \tilde{s} + \tilde{u}_1 \\ \tilde{x}_i = \tilde{\tilde{g}}_i \star \tilde{h}_i \star \tilde{s} + \tilde{u}_i \end{cases}. \tag{3.28}$$

Notice that $\tilde{h}_i = \tilde{\tilde{g}}_i \star \tilde{h}_1$ corresponds to Eq. (3.27) in the frequency domain. Moreover although the real-world RIRs h_1 and h_i are causal, their ReTF needs not be so.

The ReTF benefits of several interesting properties that will be of fundamental importance for this thesis. In particular:

- the ReTF associated to the reference channel ($i = 1$) is equal to 1 for each frequency bin k.
- The problem of estimating the ReTF can be considered "easier" than RIRs estimation. In fact, in the noiseless case, it holds that $\tilde{x}_i = \tilde{\tilde{g}}_i \star \tilde{x}_1$.
- The ReTF encode properties of the related impulse responses and there are many efficient methods to estimate them[26]. Therefore, it may be used as a proxy for the estimations of (components of) RIRs.
- A RIR can be seen as a special case of ReTF where the non-reference microphone is a virtual one whose output is the original (non-spatial) source signal s. In fact, if $h_1 = \delta$ then $\tilde{g}_i = h_i$[27].

[26] In **??** methods for estimation the ReTF will be discussed

[27] In practice this virtual microphone is sometimes substituted by a microphone that is very close to the source.

- As discussed below, ReTF simplify to special steering vectors in free- and far-field conditions, which have interesting geometrical properties.

In the general case of multiple microphone arrays ($I > 2$) and multiple sources, the vector of ReTF $\boldsymbol{G}_j[k] = \left[\hat{G}_{1j}[k], \cdots, \hat{G}_{Ij}[k]\right]^\mathsf{T}$ for the j-th source is defined as

$$\hat{\boldsymbol{G}}_j[k] = \frac{1}{\hat{G}_{1j}[k]} \hat{\boldsymbol{G}}_j[k]. \tag{3.29}$$

▶ THE RELATIVE STEERING VECTORS are obtained by combining Eqs. (3.26) and (3.27) as

$$\hat{\boldsymbol{D}}_j[k] = \left[1, \mathrm{e}^{-\mathrm{i}2\pi f_k (q_{2j} - q_{i'j})/c}, \ldots, \mathrm{e}^{-\mathrm{i}2\pi f_k (q_{Ij} - q_{i'j})/c}\right] \tag{3.30}$$

where $(q_{ij} - q_{1j})/c$ is the Time Difference of Arrival (TDOA) between the i-th and the reference microphones. The TDOAs will be the protagonists of Chapter 10 as they are fundamental quantities for sound source localization.

▶ IN THE CONTEXT OF SPATIAL AUDITORY PERCEPTION and Computational Auditory Scene Analysis (CASA), the ReTF is related to the *interchannel cues*[28]. In fact, the ReTF encodes the so-called Interchannel Level Difference (ILD) and the Interchannel Phase Difference (IPD)

[28] sometimes refers to as *interaural cues* when a stress is put on the fact that the two ears are considered as receivers

$$\begin{aligned} \mathtt{ILD}_{ij}[k] &= 20 \log_{10} |\tilde{g}[k]| \quad [\mathrm{dB}] \\ \mathtt{IPD}_{ij}[k] &= \angle \tilde{g}[k] \quad [\mathrm{rad}] \end{aligned} \tag{3.31}$$

As shown in Figure 3.7, the ILD and the IPD cluster around the direct path, associated to the direct path component. However early echoes and reverberation make them significantly diverge.

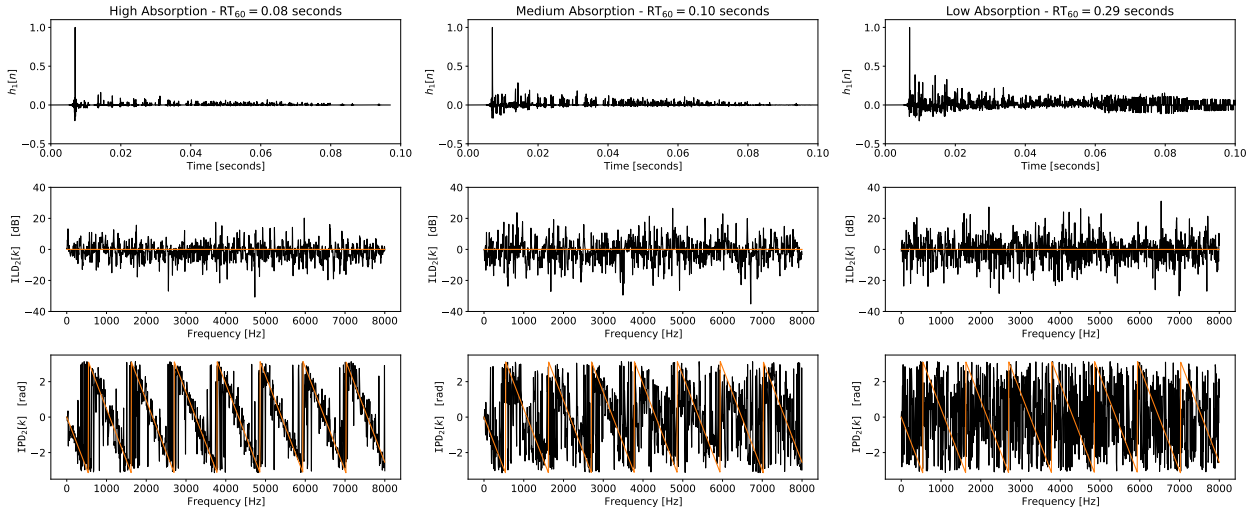


FIGURE 3.7: RIR, ILD and IPD corresponding to the pair of synthetic impulse responses of Figure 2.9 for different absorption conditions. Orange lines denote the theoretical far- and free- field ILD and IPD as defined by the relative steering vectors of Eq. (3.30)

Part III

ACOUSTIC ECHO RETRIEVAL

4

Acoustic Echo Retrieval

"[. . .] *dicebat Bernardus Carnotensis nos esse quasi nanos gigantium humeris insidentes.*"
—Giovanni of Salisbury, *Metalogicon (III, 4)*

▸ Synopsis This chapter amis to provide the reader with knowledge of the state-of-the-art of Acoustic Echo Retrieval (AER). After presenting the AER problem in § 4.1, the chapter is divided into three main sections: § 4.2 defines the categories of methods thank to which the literature can be clustered and analyzed in detail later in § 4.3. Finally, in § 4.4 some datasets and evaluation metrics for AER are presented.

4.1 Problem Formulation

The continuos-time multi-channel signal model for a signal source and I channels writes

$$\begin{aligned} \tilde{x}_i(t) &= (\tilde{h}_i \star \tilde{s})(t) \\ \tilde{h}_i(t) &= \sum_{r=0}^{R} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \tilde{\varepsilon}_i(t), \end{aligned} \tag{4.1}$$

where $\tilde{h}_i(t)$ is the echo model for the RIR between the i-th channel and the source. The sum comprises the line-of-sight propagation and the earliest R echoes we want to account for, while the error term $\varepsilon_i(t)$ collects later echoes and the reverberation tail.

The Acoustic Echo Retrieval (AER) problem consists in estimating the echoes' timings $\{\tau_i^{(r)}\}_{i,r}$ and attenuations (or gains) $\{\alpha_i^{(r)}\}_{i,r}$ of Eq. (4.1). Depending on the field of research, the echoes's timings are also known as time delays, Time of Arrival (TOA) or locations.
The term AER is not an established name for such problem and, depending on the field of research and the prior knowledge available, it can be referred to with different name. In fact AER can be seen as general case of *TOAs estimation*, or an instance of *acoustic channel estimation* (or *shaping*), *spike retrieval* and *onset detection*. As opposed to AER, the task of TOAs Estimation, is only focused in estimating the echos' timings $\{\tau_i^{(r)}\}_{i,r}$. The only knowledge of TOAs is sufficient for typical application related to SSL and RooGE.
Moreover knowing $\{\tau_i^{(r)}\}_{i,r}$, the attenuations $\{\alpha_i^{(r)}\}_{i,r}$ can be estimated in closed-from as showed in [Condat and Hirabayashi 2015].
TOAs estimation is sometimes called *time delays estimation*, when the origin of time is taken w. r. t. the first TOA and not to the time of emission. Hereafter we will make distinction between the two.

The AER may be confused with the *acoustic echo cancellation* problem of telecommunication and telephony which refers to the problem of estimating and suppressing feedbacks due to close speaker to microphone proximity.

4.2 Taxonomy on of Acoustic Echo Retrieval methods

In general, we can identify four main categories which differ on whether the source signal is known and on whether the estimation of the RIR is performed.

⇄ Active vs. passive approaches. *Active* methods assume active scenarios, namely, they use one or more loudspeakers to probe the environment and one or more microphones to record the propagated probe sound. Therefore, they assume that the source reference signal is known. They falls into the big categories of *deconvolution problems* since a "clean" reference signal is used *deconvolve* the observed one. Two are the main advantages of these approaches. First, provided a proper probe signal, a good estimation of the RIR can be achieved. Second, these methods can be used on single-channel recordings.
Instead, *passive* approaches use sets of passive sensors to record the sound field. To decouple environment from the source, they rely either on prior knowledge about the source signal or by comparing the signals received at two (or more) spatially-separated microphones. The methods are also referred to as *blind*, as the are source agnostic and are far more challenging. Passive scenarios are more common in real applications. A great deal of efforts has been devoted to these problems and research is still active in topic. Moreover, an obvious advantage is that these approaches are non-intrusive since only already existing sounds are used in the estimation.

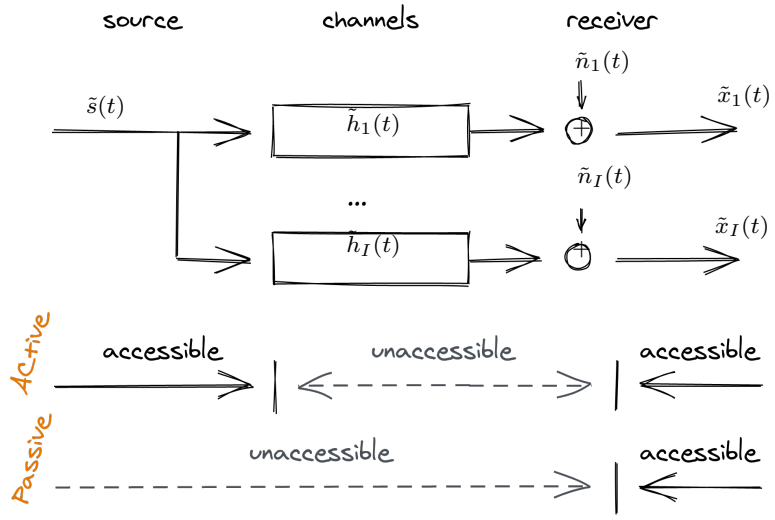


Figure 4.1: Schematic of active and passive approaches.

⇄ RIR-based vs. RIR-agnostic approches. *RIR-based* methods estimate the echoes' properties after estimating the (full or partial) RIR(s). By modeling the early part of the RIR as in Eq. (4.1), solving the AER problem can be seen as solving two subsequent tasks: RIR estimation followed by echo extraction.

The former can be seen as an instance of *channel estimation* (a. k. a. *system identification*) problems, while the latter as a *spike retrieval*, *pick picking* or *onset detection* problems. Other methods estimate the RIRs partially using assumptions derived by the application. It is the case of *impulse response shaping* or *shortening*. In the context of room acoustics, the aim is to reduce the late reverberations allowing some few early reflections which are perceptually useful [Betlehem et al. 2012].
RIR-agnostic methods, instead, try to surpass the challenging task of estimating the acoustic channel and tuning peak-picking methods. They attempt to estimated echo properties directly in the parameter space of echos' TOAs and amplitudes.

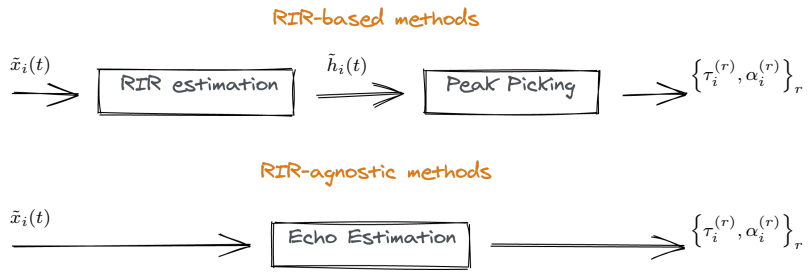


FIGURE 4.2: Schematic of RIR-based and RIR-agnostic approaches.

Given the above categories, we can now review the AER methods presented in the literature.

4.3 LITERATURE REVIEW

4.3.1 *Active and RIR-based method*

In this categories fall all the methods that first attempt for a "good" estimation of RIRs for which the reference signal is known.

- THE RIR ESTIMATION STEP is typically modeled as a deconvolution problem whose performances depend on the type of transmitted signal. When the transmitted signal is arbitrary, several methods were developed to measure real RIRs. Since the RIR identifies the room response to a perfect impulse, one can measured it by producing an impulse sound, e. g. a clap, piercing a ballon, or a gun shot. Even though this methods are commonly used, they show clear limitations in term of reproducibly and safety. Moreover a perfect impulsive and point source is difficult to reproduce in practice. Instead, modern computational technique are used, involving the computation of deconvolution (or *correlation*) between an known emitted signal and the recorded output.
The Minimum Length Sequence (MLS) technique was first proposed by Schroeder [Schroeder 1979] and it is based on the excitation of the acoustical space by a periodic pseudo-random signal, called MLS. The RIR is then calculated by circular correlation between the measured output and the original MLS signal. This method was further improved in order to achieve better RIR estimation in [Dunn and Hawksford 1993; Aoshima 1981]. Unfortunately this technique introduces several artifacts which yield to spurious peaks in the estimation. Moreover, it is sensible to the harmonic distortions introduced by the playback device, e. g.

the loudspeakers.
To overcome these issues, the Exponential Sine Sweep (ESS) technique was introduced by Farina [Farina 2000; Farina 2007]. The probe signal is the ESS signal, a. k. a. *chirp signal*, which benefits of the following properties: the signal spans a user-defined frequency range; it is *self-orthogonal*, namely it compress into Dirac's impulse during autocorrelation; and its Fourier inverse is available in closed form, allowing the user to not record and invert the probe signal. The reader can find a review of the presented techniques in [Szöke et al. 2019] applied to RIR measurements.

Sometimes the reference signal is known, but none of the above techniques can be used. In such a scenario, the RIR estimation problem needs to be addressed as a more general deconvolution problem, typically solved through optimization methods [Lin and Lee 2006]. This approaches is well studied in literature and can be solved using standard Linear Least Squares with closed-form solution. However, in the case of a narrowband signal (e. g. speech or music) or low SNR, it becomes ill-conditioned and prior knowledge about the RIR is used to improve the estimation [**needForAGoodCitationHere**].

- Echo Retrieval from RIR. As discussed in Part II, acoustic echoes can be identified as peaks in the early part of the RIR. In general, due to the measurement process, such peaks are not necessarily positive, thus, to better visualize them, the *echogram* [Kuttruff 2016], $|\tilde{h}(t)|$, or the energy envelop[29] [Schroeder 1979] are used instead.
  Provided a good estimation of the RIR, the echoes' location and amplitudes could be extracted manually by experts. However, even in ideal scenario, the automation of this process and the correct identification of such quantities are not straightforward tasks. As showed in [Tukuljac et al. 2018], since the TOAs are not necessarily multiple of the sampling grid, their true locations (and amplitudes) are blurred by spurious side peaks. This issue is referred to as *basis mismatch* in the *compressed sensing* literature. Although it can be alleviated by increasing the sampling frequency, it is bound to occur in practice. Moreover, the harmonic distortion due to the non-ideal source-receiver coupling may introduce other spurious spikes as well. Furthermore, as noticed in [Defrance et al. 2008b], even small errors of echoes timing estimation yield to significant differences in echo-based applications.

[29] The energy envelope of a signal is computed as the magnitude of its analytic representation computed with the Hilbert transform.

The existing methods for extracting echoes from RIRs can be further dichotomized into two broad categories: on-gird and off-grid approaches. The methods belonging to the former group are the most used in practice, and advance technique are used to cope with the presence of spurious peaks [Kuster 2008; Crocco et al. 2017; Remaggi et al. 2016; Defrance et al. 2008a; Bello et al. 2005; Cheng et al. 2016; Defrance et al. 2008a; Annibale et al. 2012; Kelly and Boland 2014; Usher 2010].
The most straightforward approach is to deploy iterative and adaptive thresholding algorithm on the RIR, followed by robust and manually tuned peak finders [Kuster 2008; Crocco et al. 2017].
To better inform the peak-picking, several strategies have been proposed. In the work of [Remaggi et al. 2016], based on a algorithm presented in [Naylor

et al. 2006], peaks are clustered according to changes in the phase slope of the RIR spectrum. Other works apply onset detection techniques used in music information retrieval and transcription based on edge-detection wavelet filters [Bello et al. 2005], non-negative matrix factorization [Cheng et al. 2016], or considering the RIR's Kurtosis [Usher 2010].
By noticing that the reflection in the RIRs exhibit similar shape of the direct path, the author of [Defrance et al. 2008a] first proposed the use of *Matching Pursuit* (and improvements) to identify such shapes. Here the direct sound part was used as pattern (or atom) to be retrieved across the RIR. Unfortunately, in its pure form, this approach is unsuitable for RIRs because of the non-stationary nature of the reflections due to the frequency dependent characteristic of the room absorption material. In order to improve the detection, [Kelly and Boland 2014] extents this approach employing *Dynamic Time Warping* to account for the non-uniform compression, dilation and concurrency of the echoes. Nevertheless, the idea of exploiting the direct path component to isolate the source-receiver coupling and thus identify first prominent reflection through deconvolution was used in [Annibale et al. 2012]. This technique is also referred to as *matching filter* or *direct-path compensation.*
Alternatives approaches, detect the echo timings in other signal domain. In [Vesa and Lokki 2010] the echoes are localized in the Time-Frequency (TF) domain using the cross-wavelet transform based on previous works [Guillemain and Kronland-Martinet 1996; Loutridis 2005]. Curiously, the works [Ristić et al. 2013; Pavlović et al. 2016] use (multi-)fractal analysis to detect echoes in the Time-Frequency (TF) domain. Alternatively, the authors of [Ferguson et al. 2019; Jia et al. 2017] propose to identify echoes properties in the cepstral domain. The *cepstrum* is the spectrum of a logarithmic spectrum and is used to detect periodicity in the spectral domain, typically in hydraulic and mechanic application. This approach seems promising since time-domain spikes are mapped as complex sinusoids in frequency. However this representation is highly sensible to external and sampling noise and the accuracy is limited by the approximation of the DFT operator.

All the above mentioned works aims at detecting echoes on the sampling grid. In order to face the inherent limitations of these approaches, off-grid frameworks have been proposed, e. g. [Condat and Hirabayashi 2013]. This approach can be related to classical Maximum Likelihood (ML) estimation approaches, which consist in selecting the model which is most likely to explain the observed noisy data. In this category fall classical spectral estimation techniques, e. g. Multiple Signal Classification (MUSIC) [Loutridis 2005], Estimation of Signal Parameters via Rational Invariance Techniques (ESPRIT) [Roy et al. 1986], which are fast but statistically suboptimal. The method presented in [Condat and Hirabayashi 2013] focuses on the general problem of estimating a finite stream of Dirac's pulses from uniform, noisy and lowpass-filtered samples. The authors showed that this particular problem can be reformulated as a *matrix denoising*, from which the echoes location and amplitudes can be retrieved in closed-from. Although this method reach the statistical optimality in the ML sense, the exact knowledge of number of Diracs needs to be known in advance. If this number is unknown or approximated, huge errors in the estimation are observed. This results in a huge drawback since the exact

number of echoes is difficult to know a priori and false-positive spikes are present even in clean RIRs.

That have being said, AER is far from trivial and solved even when clean RIR estimates are provided. It is important to note that, for every TOA estimator, a practical trade off exists between the number of missed TOAs and the number of spurious TOAs wrongly selected. This trade-off is only partially dependent on the noise level since, many factors can provide spurious peaks. For instance, side lobes due to finite signal bandwidth, echo distortions due to frequency dependent attenuations and coalescing peaks due to close TOAs can affect peak estimation. This fact is often a source of unavoidable outliers that make the robustness of subsequent steps in echo-aware application a delicate and very important issue. A way to overcome to this is to overestimate the echoes in the RIR by including some false-positive and prune them using echo labeling afterwards.

- Echo Labeling or TOAs Disambiguation is the task of assigning acoustic echoes to different image sources or reflectors. Many methods have been proposed in the context of SSL [Scheuing and Yang 2006; Zannini et al. 2010], microphone calibration [Parhizkar et al. 2014; Salvati et al. 2016] and RooGE [Antonacci et al. 2010; Filos et al. 2011; Venkateswaran and Madhow 2012; Antonacci et al. 2012; Dokmanić et al. 2013; Crocco et al. 2014; Jager et al. 2016; El Baba et al. 2017]. A brief review of these methods is provided in [Crocco et al. 2017].
  In the context of SSL, the disambiguation is typically performed in the TDOAs space [Scheuing and Yang 2006; Zannini et al. 2010]. Moreover these works focus on actively localizing (the direction of arrival of) multiple sources while discarding reflection, rather than localizing the actual image sources.
  The other disambiguation schemes are typically used for RooGE. In [Venkateswaran and Madhow 2012] the pruning of the combinatorial candidate-image-source search is done through Bayesian inference. A similar approach can be found in [Dokmanić et al. 2013; Parhizkar et al. 2014] where the validity check is based on a particular structured matrix called *Euclidean Distance Matrix* and further improved using compatibility graphs in [Jager et al. 2016]. These methods rely on a combinatorial search with potentially high number of candidates, which leads to intractable computational complexity when multiple reflection are considered. Moreover these methods require that all the distances between each microphone are known with precision, which may not be true in practice.
  In the works of [Antonacci et al. 2010; Filos et al. 2011; Antonacci et al. 2012], the reflectors are modelled as planes tangent to the ellipsoids with foci given by each pair of microphone/source. By solving a non-convex optimization problem based on geometrical reasoning and the Hough transform [30], they are able to disambiguate TOAs and reconstruct reflectors position and inclination. However, they all require a very specific acquisition setup and the non-convex cost function are sensible to local minima.

[30] A mathematical operator that maps points into curves in a 2-D space. If a set of points belongs to the same line, the corresponding curves will intersect in a single point. In computer vision, this tranform is typically used as feature extractor to detect lines and edges in pictures.

In general, all the above methods do not have specific strategies to cope with missing or spurious echoes' estimates. This is due to malfunctioning of the peak finder or by erroneous selection of peaks corresponding to higher reflection

order. A way to solve this issues is to exploit particular prior knowledge. For instance, the approach presented in [El Baba et al. 2017] exploits the shapes of linear and compact arrays of loudspeakers, which provide a natural ordering among the loudspeakers. By stacking side-by-side the measured RIRs in a matrix, they can be visualized as an image. Here the wavefront of each reflection draw specific pattern which can be identified easily and robustly even in presence of (a few) spurious and missing peaks. Moreover, this approach avoids the combinatorial search, but still requires a very specific setup for recordings.
In the work [Crocco et al. 2014] an iterative strategy is used. First the direct path arrivals are used to estimate a first guess of the microphone and source positions. Then, the whole set of extracted peaks are used to estimate the planar reflectors positions which are then used to refine the microphone and source localization. Alternating between the geometrical space of microphone and source coordinates and the signal space of the echoes' TOAs, the ambiguous peaks are pruned during the optimization.

4.3.2 *Active and RIR-agnostic method*

This class of methods uses the signal at the microphones to directly estimate the echoes reflections, rather than estimating the RIRs. Here two different approaches can be identified: optimization-based approaches [Jensen et al. 2019; Saqib et al. 2020] and cross-correlation-based approaches [Crocco et al. 2014; Al-Karawi and Mohammed 2019].

The former approaches exploit the strong relation between the TOA of a echo with its Direction of Arrival (DOA). When multiple microphones are used and their geometry is known, the relation between these two quantities can be express in closed-form and used in a ML-based frameworks. By modeling the Directions of Arrival (DOAs), such approaches are able to implicitly reduce the ambiguity of the estimated echoes. This idea is rooted in existing methods used in multipath communication systems, denoted as Joint Angle and Delay Estimation (JADE) [Vanderveen et al. 1997; Verhaevert et al. 2004].

Alternatively, the echoes contribution can be extracted from the *correlation* between the observed and the reference signals. The cross-correlation analysis is a mathematical tool for the identification of repeated patterns in a signal as function of a certain time lag. Due to indoor sound propagation, the received signal consists in repeated copies of the emitted signal. Therefore, the received signal may correlate with the emitted one for certain time lags. Therefore, peaks in the cross-correlation function can be observed. By the extraction of these peaks, echoes' TOAs and relative amplitudes can be identified. This approach was used in [Tervo et al. 2011; Crocco et al. 2014; Al-Karawi and Mohammed 2019].

Cross-correlation and convolution operations are very similar, an mathematically they differ just by the inversion of the reference signal. While, the former measures the similarity between two signals as function of a translation, the latter measure the effect of one signal on the other signal.

When the array geometry is known, the time lag axes of cross-correlation functions between channels can be mapped to possible 2D directions of arrivals (elevation and azimuth), namely from TOAs to 2D-DOAs. The identification of strong reflections can be then performed in the so-called *angular spectrum* domain DOAs [DiBiase et al. 2001]. With a proper clustering approach, the

reflections can be inspected, disambiguated and their TOAs deduced. This approach is used in [O'Donovan et al. 2008; O'Donovan et al. 2010; Tervo and Politis 2015] and can be generalized by spatial filtering methods, such as steered-response power-based beamforming. In [O'Donovan et al. 2008], it was referred to as *acoustic camera* since it benefits of the following visual interpretation: As shown in Figure 4.3, the 2D-polar coordinates can be mapped into cartesian ones so that the angular spectrum can be superimposed to a panoramic picture of the audio scene taken by the barycenter of the recording arrays.

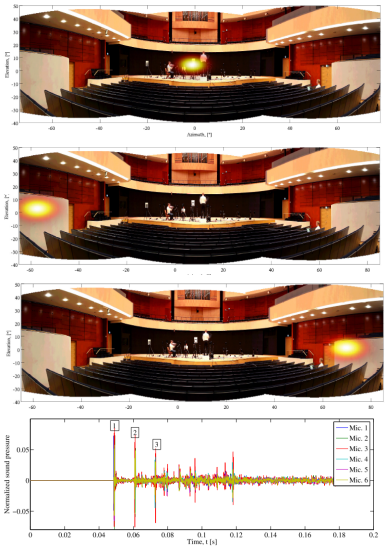


FIGURE 4.3: Visualization of the *audio camera*: The angular spectrum is overlapped to the corresponding images. Also shown are the impulse responses for 6 microphones. The numbered boxes indicate events shown in the audio camera. Images from [Tervo 2011].

4.3.3 *Passive and RIR-based method*

Passive approaches rely on external sound sources in the environment to conduct the estimation. In the literature, this problem belongs to the broad and deeply studied category of Blind Channel Estimation (BCE) (or Blind Sistem Identification (BSI)) problems. In the particular case of a single source, it is referred to as SIMO BCE. Common to all these methods is the assumption that RIRs are discrete Finite Impulse Response (FIR) filters defined on the sampling grid, namely, vectors in the Euclidean space. In the general setting of arbitrary signals and filters, rigorous theoretical ambiguities under which the problem is unsolvable have been identified [Xu et al. 1995]. Some well-known limitations of these approaches are their sensitivity to the chosen length of the filters, and their intractability when the filters are too large. FIR SIMO BCE can be broadly dichotomized into the class of *statistical methods* and the class of *blind methods*.

The innovative idea of passive Blind Channel Estimation (BCE) can be traced back to [Sato 1975]. A review of the the evolution of Single Input Multiple Output (SIMO) BCE can be found in [Huang and Benesty 2003].

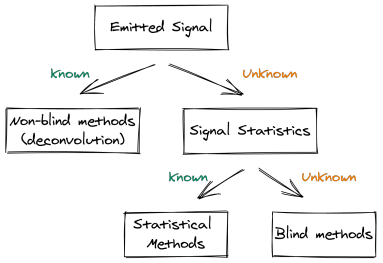


FIGURE 4.4: Classification of the State of the Art in channel estimation.

- STATISTICAL METHODS exploit knowledge about the emitted signal. Since the nature of the source signal is by definition not deterministic, their statistics can modeled based on the signal category, e. g. speech or music, and modeled accordingly. Two main approaches can be identified [Tong and Perreau 1998]:

  - *Second Order Moments approaches* derive closed-form solution for which the knowledge of the source auto-correlation (or variance) function is required.

  - *Maximum Likelihood approaches* require instead the source probability function. Even though they are optimal in the ML sense, they optimize non-convex cost functions, typically via Expectation Maximization (EM). In this category one may include the methods developed for multichannel blind source separation [Ozerov and Févotte 2009; Duong et al. 2010; Leglaive et al. 2016; Leglaive et al. 2018; Scheibler et al. 2018d]. These methods are built on the well-studied framework of Multichannel Nonnegative Matrix Factorization (NMF) [Ozerov and Févotte 2009] which lends itself to account for various type of side information. Here the source signals are typically modeled as Gaussian distributions centred in zero and with unknown variance. Using pre-trained dictionaries for modeling the variance of the sources, they are able to estimate both the acoustic channels and the source contribution. In particular, the work of [Duong et al. 2010] extends this framework to reverberant recordings using physic-based models for the late reverberations, while

the works of [Leglaive et al. 2016] considers explicitly the contribution of early echoes, further improved in [Leglaive et al. 2018].

Even if statistical methods have reported a considerable success in the field of Sound Source Separation, they play a minor role in RIR estimation. This is due to the difficulty in achieving reliable statistics of the emitted signals or a good initialization point required by the EM. Moreover, although the final estimated RIRs may match the real ones in the statistical sense, they lack of a sufficient details, indispensable for AER.

- BLIND METHODS comprises two main groups: *subspace* methods [Abed-Meraim et al. 1997] and *cross-relation methods* [Tong et al. 1994; Xu et al. 1995; Lin et al. 2007; Lin et al. 2008; Kowalczyk et al. 2013; Crocco and Del Bue 2015; Crocco and Del Bue 2016a]. The formers are based on the key idea that the channel (or part of it) vector spans a one-dimensional subspace of (a block of of) noiseless observations. These methods have the attractive property that the channel estimates can often be obtained in a closed-form by optimizing a quadratic cost function. However they may be not robust, especially when the channel covariance matrix is close to being singular. The second disadvantage is that they are typically computationally expensive.

The second family of methods rely on the clever observation that in noiseless case, for every pair of microphone (i, i'), it holds

$$(\tilde{x}_{i'} \star \tilde{h}_i)(t) = (\tilde{x}_i \star \tilde{h}_{i'})(t) = ((\tilde{h}_{i'} \star \tilde{h}_i) \star s)(t), \tag{4.2}$$

by the commutativity of the convolution operator. This principle is called the *cross-relation* and it was firstly introduced by [Tong et al. 1994]. In this work, the RIR are estimated by solving a Least Square minimization of the sum of square cross relation errors. In [Xu et al. 1995; Tong and Perreau 1998], sufficient and necessary conditions for channel identification are discussed. This approach has received significant attention as it does not require any assumption about the source signal. Later, the accuracy of estimated RIRs has been subsequently improved using prior knowledge of the filters: in particular, the authors of [Lin et al. 2007] have proposed to use sparsity penalty and non-negativity constraints to increase robustness to noise as well as Bayesian-learning methods to automatically infer the value of the hyper-parameters in [Lin et al. 2008]. Even if sparsity and non-negativity could be seen as a strong assumption, works in speech enhancement [Ribeiro et al. 2010b; Dokmanić et al. 2015b] and room geometry estimation [Antonacci et al. 2012; Crocco et al. 2017] have proven the effectiveness of this approach. On a similar scheme, in [Kowalczyk et al. 2013], the (4.2) is solved using an adaptive time-frequency-domain approach while [Aissa-El-Bey and Abed-Meraim 2008] proposes to use the ℓ_p-norm instead of the ℓ_1-norm. A successful approach has been presented by Crocco *et al.* in [Crocco and Del Bue 2015; Crocco and Del Bue 2016a], where the anchor constraint is replaced by an *iterative weighted* ℓ_1 equality constraint to better balance sparsity penalty and the model constraints[31]. Finally, the very recent work [Qi et al. 2019] extends cross-relation approaches under the umbrella of the Kalman filter which was previously used for echo-cancellation applications.

[31] These approaches will be further formalized and detailed in Chapter 5.

An alternative approach is used in [Čmejla et al. 2019], where the RIR estimation problem is treated as special case of RTF estimation. As mentioned in § 3.3.2, in the noiseless case, the RTF identifies the RIR when the reference microphone is placed very close to the source. RTF estimation found its root in the field of Speech Enhancement (SE) [Gannot et al. 2001] and many techniques have been proposed since then [Gannot et al. 2001; Koldovskỳ et al. 2015; Koldovsky and Tichavsky 2015; Kodrasi and Doclo 2017][32]. In general, by its definition, RTF describes the relative filter between two observations and not directly their RIRs and may differs in case of noise. The main limitation of this approach is that it is possible only in measurement scenarios, where the user has the possibility to place the microphone arbitrarily in the room and in presence of high SNR levels. Nevertheless, in this context, this particular setup is found to be useful not only for RTF estimation, but also for microphone calibration, since it allows to solve geometrical ambiguities, yielding a closed-form solution, as done in [Crocco et al. 2012].

[32] Methods for RTF estimation will be detailed in **??**.

In general, the main drawbacks of FIR SIMO BCE works is that they rely on on-grid estimation, sparsity-enforcing regularizers and peak-picking which need to be tuned manually. As described in § 3.2.3, due to the sampling process involving a sinc function, the filters are non-sparse and non-negative. This general bottle-neck has been referred to as *basis mismatch* and was notably studied in the *compressed sensing* community [Chi et al. 2011]. In particular, the true peaks in the RIR do not necessarily correspond to the true echoes as shown in **??**

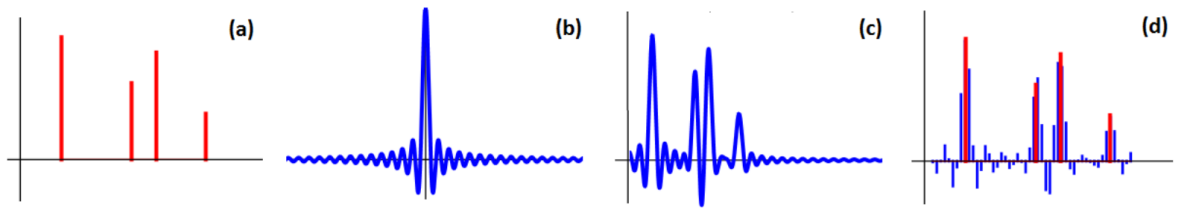


FIGURE 4.5: (a) Continuos-time stream of Diracs $\tilde{h}(t)$, (b) sinc kernel $\tilde{\phi}_{\text{LP}}(t)$, (c) smoothed stream of Diracs, (d) original stream of Dirac $\tilde{h}(t)$ (red) and its sampled (i. e., smoothed and discrete) version (blue). Image courtesy of [Tukuljac et al. 2018]

Since these methods are fundamentally on-grid, the estimated echo locations are integer multiples of the sampling period $1/F_s$. This prevents subsample resolution, which may be important in applications such as RooGE [Crocco et al. 2017] or acoustic parameter estimation [Defrance et al. 2008b]. Moreover, these methods strongly rely on the knowledge of the length of the filters. When this parameter is underestimated or overestimated, identifiability and computational issues may arise, affecting the estimation. Nevertheless, despite this slight mismatch between theoretical assumptions and real data, for some scenarios, the position of the estimated peaks by the methods [Crocco and Del Bue 2016a] reproduces the positions of the ground truth peaks with remarkable precision as demonstrated in our work [Di Carlo et al. 2020].

Di Carlo et al., "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval"

4.3.4 *Passive and RIR-agnostic methods*

Methods in this categories bypass the onerous task of estimating the (full or partial) acoustic channel and, to the best of our knowledge, only a few have been identified. As for the active and RIR-agnostic case, the audio camera based on the cross-correlation function can be used in passive settings. Exploiting the geometrical knowledge of the microphone array, TDOAs extracted from robust correlation function can be mapped to DOAs [DiBiase et al. 2001; O'Donovan et al. 2008; O'Donovan et al. 2010]. Assuming a single source scenario, difference DOAs can be disambiguated using geometrical prior knowledge and can be associated to image sources, hence reflectors. These methods typically ignore the echoes amplitudes and in general do consider only angles on the unit sphere, ignoring the distance from the source. Without proper prior knowledge, their application to AER is far from trivial as RooGE and *reflector estimation* methods need to be used to convert DOAs back to echoes timing.

Recently a fully blind, passive, off-grid and RIR-agnostic method was proposed by authors of [Tukuljac et al. 2018] for stereophonic recordings, namely using only 2 microphones. They proposed a method, called Multichannel Annihilation (MULAN), based on the properties of the *annihilation filter*[33], [Condat and Hirabayashi 2013] and the theory of Finite Rate of Innovation (FRI). If the source signal is known, starting from the cross-relation identity, the AER problem translates into finding the annihilation filter for the RIRs, which can be recasted into an eigenvalue problem. In the fully blind case, the problem is solved with non-convex optimization, iterating between the estimation of the two filters and the signal until convergence. The method was later extended to the multichannel case in [Tukuljac 2020] using the generalization of Cadzow denoising framework [Condat and Hirabayashi 2015]. This method is shown to outperform conventional approaches by several orders of magnitude in precision in noiseless case, with synthetic data and when the correct number of echoes is known a priori. However its effectiveness was not tested on challenging real scenarios featuring external noise and partial knowledge on the number of echoes.

[33] For a sequence of Fourier coefficients $a_N \in \mathbb{C}^N$ (describing a signal or a filter), its annihilation filter $b_L \in \mathbb{C}^L$ is such that the linear convolution between the sequence and the filter coefficients is identically zero: $\sum_{l=0}^{L-1} b_L[l]a_N[n-l] = 0 \quad \forall n = -N+L, \dots, N$.

4.4 Data and Evaluation

AER is a relatively recent problem which is typically addressed in the context of much broader applications, e. g. SE, RooGE, SSL. Therefore the literature lacks of standard datasets as well as standard evaluation frameworks.

4.4.1 *Datasets*

As listed in [Szöke et al. 2019] and in [Genovese et al. 2019], a number of recorded RIRs corpora are freely available online, each of them meeting the demands of certain applications, usually SE and Finite Rate of Innovation (ASR). However, even if these datasets feature reverberation and strong early reflections, they lack of proper annotations, making them difficult to use for testing AER methods. For this reason, to bypass the complexity of recording real annotated RIR datasets, simulators based on the ISM are extensively used instead. While simulated datasets are more versatile, simple and quicker to

Database Name	Annotated			Number of				Key characteristics	Purpose
	Pos.	Echoes	Rooms	RIRs	Rooms	Mic×Pos.	Src		
[Dokmanić et al. 2013]	✓	∼	∼	15	3	5	1	Non shoebox room	RooGE
[Crocco et al. 2017]	✓	∼	✓	204	1	17	12	Accurate 3D calibration Many mic and src positions	RooGE
[Remaggi et al. 2016]	✓	∼	✓	∼1.5k	4	48×2	4-24	Circural dense array Circular placement of sources	RooGE SE†
[Remaggi et al. 2019]	✓	∼	✓	∼1.6k	4	48×2 +2×2	3-24	Circural dense array Binaural Recordings	RooGE† SE
BUT Reverb [Szöke et al. 2019]	✓	✗	∼	∼1.3k	8	(2-10)×6	3-11	Accurate metadata different device/arrays various rooms	SE/ASR
VoiceHome [Bertin et al. 2019]	✓	✗	✗	188	12	8×2	7-9	Various rooms, real homes	SE/ASR
`dEchorate` Chapter 7	✓	✓	✓	∼1.8k	1	30	6	Accurate annotation Different Echo-energy	RooGE SE/ASR

TABLE 4.1: Comparison of some existing RIR databases which account for early acoustic reflections. Receiver positions are indicated in terms of number of microphones per array times number of different positions of the array (∼ stands for partially available information). The reader is invited to refer to [Szöke et al. 2019; Genovese et al. 2019] for a more complete list of existing RIR datasets.
†The dataset in [Remaggi et al. 2016] is originally intended for RooGE and further extended for (binaural) SE in [Remaggi et al. 2016] with a similar setup.

obtain, they fail to fully capture the complexity and the richness of real acoustic environments. Due to this intrinsic issues, methods trained or validated on them may fail to generalize to real conditions, as will be shown in Chapter 7.

A good dataset for AER should include a variety of environments (rooms geometries and surface materials), of microphone placings (close to or away from reflectors, scattered or forming ad-hoc arrays) and, most importantly, precise annotations of the scene's geometry and echo parameters within the RIRs. Moreover, in order to be versatile and used in echo-aware applications, the provided annotations should match the ISM, *i.e.*, TOAs should be expressed in terms of image sources and vice-versa. Such data are difficult to collect since they require precise measurements of the positions and orientations of all the acoustic emitters, receivers and reflective surfaces inside the environment with dedicated planimetric equipment. We identified here two main classes of related RIR datasets in the literature: SE/ASR-oriented datasets, e.g. [Szöke et al. 2019; Bertin et al. 2019; Čmejla et al. 2019], and RooGE-oriented datasets, e.g. [Dokmanić et al. 2013; Crocco et al. 2017; Remaggi et al. 2016]. The formers include acoustic echoes as highly correlated interfering sources coming from close reflectors, (e. g. a desk in meeting rooms or the close wall), however their proper annotations are not provided. The latter group deals with sets of distributed, synchronized microphones and loudspeakers in a room. These setups are not exactly suitable for SE methods, which typically involve compact or ad hoc arrays. Table 4.1 summarizes some existing datasets that can be used in the context of AER.

4.4.2 *Metrics*

The metrics used in AER depend on the application and the methods used to estimate the echoes. When addressed as a FIR SIMO BCE problem, the ground-truth acoustic channels are considered as a discrete vector $h \in \mathbb{R}^L$, and similarly their estimates, that is, $\hat{h} \in \mathbb{R}^L$. To assess the quality of the estimated discrete filters, the following metrics have been proposed in the literature:

- *The Root Mean Square Error (RMSE)* measures the distance between points in the Euclidean space, defined by vector coordinates:

$$\texttt{RMSE}(\hat{h}, h) \stackrel{\text{def}}{=} \sqrt{\sum_{n=0}^{L-1} \left|\hat{h}[n] - h\right|^2} \quad \text{[seconds (or, samples)]}, \tag{4.3}$$

  where $|\cdot|$ denotes the absolute value. This metrics is known to be highly sensitive to scaling and little translation.

- *The Normalized Projection Misaligment (NPM)* was originally proposed in [Morgan et al. 1998] to solve the limitation of the RMSE. In the formulation provided in [Huang and Benesty 2003; Ahmad et al. 2006], it writes as

$$\texttt{NPM}(\hat{h}, h) \stackrel{\text{def}}{=} 20 \log_{10} \left(\left\| h - \frac{h^\intercal \hat{h}}{\hat{h}^\intercal \hat{h}} \right\|_2 / \|h\|_2 \right) \quad \text{[dB]}, \tag{4.4}$$

  where $\|\cdot\|_2$ denotes the Euclidean norm. By projecting $\hat{h}$ onto h and defining a projection error, only the intrinsic misalignment of the channel estimate is considered, disregarding an arbitrary gain factor and the length difference of both vectors. However it is not translation invariant.

- *The Hermitian angle* is similar to NPM and was used in the context of RTF estimation in [Varzandeh et al. 2017; Tammen et al. 2018]

$$\Delta\Theta(\hat{h}, h) = \arccos \left(\frac{h^\mathsf{H} \hat{h}}{\|h\|_2 \|\hat{h}\|_2} \right). \tag{4.5}$$

  As for NPM, this metrics is invariant to possible scaling factors and length differences between the ground-truth and the estimated vectors.

In the context of RooGE, SSL and microphone calibration, echoes' timings are typically mapped to reflectors or image source positions, either in cartesian or polar coordinates. Therefore, the models for AER are evaluated in the geometrical space, rather than in the space of echoes' parameters. For instance, for the task of reflectors localization, the accuracy is measured in terms of *plane-to-plane distance* between estimated and ground-truth surfaces and the *angular error* between their normals. In the case of SSL and microphone calibration, the *Euclidean distance* between the 3D coordinates is typically computed as RMSE between ground-truth and estimated DOAs. This metrics considers only echoes' TOA, ignoring their amplitudes which interest a previous peak peaking and echo labeling steps.

To the best knowledge of the author, the literature lacks of metrics properly defined for AER. As for the application mentioned above, echoes' amplitudes in a single RIR or between them, are typically ignored or considered for peak picking only. More attention is paid on the echoes' timing which are evaluate using regression/classification metrics of *information retrieval* and *machine learning*.
Let be $\hat{\boldsymbol{\tau}} = \{\hat{\tau}_r\}_{r=0}^R$ and $\boldsymbol{\tau} = \{\tau_r\}_{r=0}^R$ the sets of estimated and reference echoes' TOAs. The following metrics are used:

- the *RMSE* is defined as

$$\texttt{RMSE}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) \overset{\text{def}}{=} \sqrt{\sum_{r=0}^{R} |\hat{\tau}_r - \tau_r|^2} \quad \text{[seconds (or, samples)]}, \tag{4.6}$$

  This metric describes the mean error between estimated and reference of echoes' TOAs. Unfortunately, the RMSE is proportional to the size of the squared error, thus is sensitive to outliers. In the context of AER, the RMSE is computed only on the matched TOAs.

- the *Precision, Recall, and F-measure* are standard metrics used in information retrieval for evaluating classification problems, e. g. in onset detection [Böck et al. 2012]. Here the real valued estimates and ground-truth need to be converted into binary values indicating a *match*. Typically, hard thresholding is used to assess whether estimated TOAs match the reference one. In the context of AER, *precision* expresses the fraction of matching TOAs among all the estimated ones, while *recall* measure the fraction of matching TOAs that are correctly estimated. Finally, the *F-measure*, defined as the harmonic mean of precision and recall, is used to summarize precision and recall in one value.
  Depending on the application, precision and recall can have different impact. RooGE methods are more sensible to missing TOAs than to their misalignment which can be redefined with geometrical reasoning. Thus they are more incline to privilege recall over precision and allow for some false-positive which can be pruned using echo labelling methods. Instead, echo-aware SE methods prefer to accurately select the relevant echoes, thus favoring an higher precision.

Since these metrics rely on decision thresholds, their usage is not straightforward. In fact, in order to compare echoes, first both estimated and reference echoes need to be labeled, pruned and matched. As discussed at the end of § 2.3.3, echoes can be sorted differently according to their amplitudes, their TOAs or reflection order. AER tends to return echoes' parameter sorted by the echoes' amplitudes which can be distorted by the measurement process and modelling errors (Cf. **??**). This matching and labeling process introduces strong biases in the evaluation process which is currently unsolved without a proper echo labeling step.

4.5 Proposed Approaches

So far, we presented a view of current methods for solving the AER problem. In the following two chapters, we will explore two novel approaches which follows two paradigms which occurs in the recent years of signal processing: *knowledge-driven* and *data-driven* methods.

▸ Knowledge-driven methods
take advantage of prior information which may have deterministic (e. g. physical equation) or asymptotic behaviors (e. g. statistical models). In this context, AER exploits prior information about the sources, the mixing process and the

physic of the acoustic propagation, along with the audio. This knowledge is typically used to build models which lead solution computed through closed-forms or optimization-based algorithm. All the literature presented in this chapter follows this approach. In general, the advantages and the disadvantages of these approach depends on the kind of knowledge is known and on the context in which is applied.
Regarding our contribution, the Chapter 5 proposes a new knowledge-driven method for solving AER based on the theory of Continous Dictionary (CD).

- Data-driven methods

, instead, are based on machine learning algorithm where information is automatically "learned" in *supervised* ways. Providing comprehensive and exhaustive annotated training datasets, such methods can learn function that maps an input to an output based on example input-output pairs. Due to its recent success, machine learning, and in particular *deep learning*, has been applied in many signal processing (sub-)task. Along side with the huge benefits of having black-box models that are able to learn by their own, this paradigm hides a few limitations.
First these models rely on the information encoded in training data which is sometimes insufficient represent the real-world data. In order to solve this issues, many strategies have been proposed, e. g. using data augmentations techniques or knowledge-driven generating models, based on simulators. This leads to the second limitation of these approaches, that is *overfitting* to the data obeying the generating model and able to generalize. Finally, machine learning models learn black-box functions. Although they can reach incredible performance, it is difficult to predict their behavior when facing new type of data. Despite this issues, data-driven methods are currently intensively studied and interlaced with knowledge-driven approaches. In this direction, we propose our contribution in **??**, as a new data-driven method for solving AER based on virtually supervised learning.

5

Knowledge-driven Acoustic Echo Retrieval & `Blaster`

Keywords: Blind Channel Identification, Super Resolution, Sparsity, Acoustic Impulse Response.
Resources:

- Paper
- Code
- Open-access paper with supplementary material
- Slides
- Presentation

▶ Synopsis This chapter proposes a novel approach for *off-grid* AER from a stereophonic recording of an unknown sound source such as speech. In contrast with existing methods, the proposed approach, named Blind and Sparse Technique for Echo Retrieval (BLASTER). It is built on the recent framework of Continous Dictionary (CD) and it does not rely on parameter tuning nor peak picking techniques by working directly in the parameter space of interest. The accuracy and robustness of the method are assessed on challenging simulated setups with varying noise and reverberation levels and are compared to two state-of-the-art methods. While comparable or slightly worse recovery rates are observed for the task of recovering 7 echoes or more, better results are obtained for fewer echoes and the off-grid nature of the approach yields generally smaller estimation errors.

The material presented in this chapter was previously published in [Di Carlo et al. 2020] and results from a collaboration with a colleague whose domain of expertise is in the Continous Dictionary (CD) framework. The CD framework applied for AER is extracted from the related publication and briefly commented, while more attention is paid in grasping the motivation behind it. Finally, this chapter recall the main findings of the paper bringing additional insight in the existing models for AER.

Di Carlo et al., "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval"

5.1 Introduction

Acoustic Echo Retrieval (AER) consists in estimating the properties of the early (strong) acoustic reflections only in multi-path environments, and sometimes referred to as time delay estimation [Chen et al. 2006]. To achieve this, several methods rely on a known source signal [Park et al. 2017; Jensen et al. 2019]. In contrast, when multiple receivers attend an unknown single source, AER can be seen as an instance of Single Input Multiple Output (SIMO) Blind Channel Estimation (BCE) problem, i. e. estimating the filters entailing an unknown input observed output of a system. A common approach for solving AER in the context of SIMO-BCE is to first blindly estimate a discrete version of the acoustic channels using the so-called cross-relation identity [Xu et

al. 1995; Crocco and Del Bue 2016a]. The location of the echoes are then chosen among the strongest peaks with ad-hoc peak-picking techniques. Such methods are generally *on-grid* in the sense that the estimation relies on a fixed grid of points and *a priori* chosen filter lengths. However, in practice, the true timings of echoes rarely match the sampling grid, thus leading to pathological issues called basis-mismatch in the field of compressed sensing. To circumvent this issue, the authors of [Tukuljac et al. 2018] proposed to leverage the framework of finite-rate-of-innovation sampling to make one step towards off-grid approaches. Despite promising results in the absence of noise and with synthetic data, the quality of the estimation highly relies on an initialization point.

Of particular interest in this paper is the recently proposed framework of Continous Dictionary (CD) [Candès and Fernandez-Granda 2014]. By formulating an inverse problem as the recovery of a discrete measure over some parameter space, CD has allowed to overcome imaging device limitations in many applications such as super-resolution [Candès and Fernandez-Granda 2014] or PALM/STORM imaging [Denoyelle et al. 2019]. In this work, we formulate the problem of AER for stereophonic mixtures, i. e. using only microphone pairs, within the framework of continuous dictionaries. The resulting optimization problem is convex and thus not prone to spurious minimizers. The proposed method is coined *Blind and Sparse Technique for Echo Retrieval (BLASTER)* and requires no parameter tuning. The method is compared to state-of-the art on-grid approaches under various noise and reverberation levels using simulated data.

5.2 Background in Acoustic Echo Estimation

5.2.1 *Signal and measurement model*

We consider here the common setup of stereophonic mixtures, that is 2-channel microphone recordings. Using the notation presented Chapter 3, recorded signal at microphone $i \in \{1, 2\}$ reads

$$\tilde{x}_i = \tilde{s} \star \tilde{h}_i^\star + \tilde{u}_i \tag{5.1}$$

where $\star$ denotes the (continuous) convolution operator, $\tilde{n}_i$ models some additive noise in the measurement process and $\tilde{h}_i^\star$ denotes the Room Impulse Response (RIR). In the remainder of this chapter, the superscript $\star$ refers to the ground truth. Assuming the echo model, the RIRs read

$$\tilde{h}^\star(t) = \sum_{r=0}^{R_i} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) \tag{5.2}$$

where R_i is the (unknown) number of echoes.
In the noiseless case, that is when $\tilde{u}_i = 0$ for $i \in \{1, 2\}$, we have the cross-relation identity

$$\tilde{x}_1 \star \tilde{h}_2^\star = \tilde{x}_2 \star \tilde{h}_1^\star \tag{5.3}$$

by commutativity of the convolution operator.

However, in practice, only sampled versions of the two recorded signals are available. More precisely, we consider the measurement model introduced

in Chapter 3: the incoming signal undergoes a (ideal) low-pass filter $\tilde{\phi}_{\text{LP}}$ with frequency support $[-F_s/2, F_s/2]$ before being regularly sampled at the rate F_s. We denote $x_1, x_2 \in \mathbb{R}^{2N}$ the two vectors of $2N$ (consecutive) samples and $i \in \{1, 2\}$ by

$$x_i[n] = \left(\tilde{\phi}_{\text{LP}} * \tilde{x}\right)\left(\frac{n}{F_s}\right) \qquad \forall n \in \{0, \dots, 2N-1\}. \tag{5.4}$$

5.2.2 *Existing works*

Starting from the identity Eq. (5.3), the common SIMO-BCE cross-relation framework aims to compute $\tilde{h}_1, \tilde{h}_2$ solving the following problem in the discrete-time domain [Lin et al. 2008]:

$$\widehat{h}_1, \widehat{h}_2 = \underset{h_1, h_2}{\arg\min}\ \tfrac{1}{2}\|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2 + \lambda\|h\|_1 \tag{5.5}$$

where x_i and h_i are the discrete, sampled version of $\tilde{x}_i, \tilde{h}_i$ respectively. $\mathcal{T}(x_i)$ is the $(2N + L - 1) \times L$ Toeplitz matrix (build as shown is Figure 5.1) associated to convolution where $2N$ and L respectively denote microphone and filter signal length.
This type of problem can be seen as an instance of Least Absolute Shrinkage and Selection Operator (LASSO) problem [Tibshirani 1996], written in the form:

$$\underset{\mathbf{u}}{\arg\min}\ \tfrac{1}{2}\|\mathbf{v} - \mathbf{A}\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1. \tag{5.6}$$

This type of well-known optimization problem are convex and, despite the non-differentiability of the ℓ_1-norm, they can be easily tackled by standard optimization tool. Later, in § 5.2.3, we show how to express Eq. (5.5) as a standard LASSO problem.

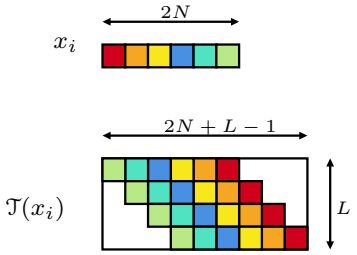


FIGURE 5.1: Graphical representation of the construction of $\mathcal{T}(x_i)$ form x_i

The accuracy of estimated RIRs has been subsequently improved using a priori knowledge of the filters. In particular, the authors of [Lin et al. 2007] have proposed to use sparsity penalty and non-negativity constraints to increase robustness to noise and avoid trivial solution. Therefore, let us define a more general formulation for Eq. (5.5), such that

$$\widehat{h}_1, \widehat{h}_2 = \underset{h_1, h_2}{\arg\min}\ \mathcal{J}(h_1, h_2) + \mathcal{P}(h_1, h_2) \text{ s.t. } \mathcal{C}(h_1, h_2) \tag{5.7}$$

where $\mathcal{J} = \frac{1}{2}\|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2$ is the cost function to optimize. $\mathcal{P}(h_1, h_2)$ and $\mathcal{C}(h_1, h_2)$ are respectively a regularization term used to promote sparse solution and a constrained set. Let us define $\mathbf{h} = [h_1^\intercal, h_2^\intercal]$ as the concatenation of the two vectorized discrete filter. Thank to this formulation, current state of the art approaches can be summarized as in the Table Table 5.1.
The constraint $h_i[0] = 1$ is called an *anchor constraint* and it is used to replace the ℓ_2-norm while keeping the problem convex. The non-negativity $\mathbf{h} \geq 0$ constraint may not be satisfied due to effects such as measurement process, the filtering in the propagation media or the imperfect frequency response of a microphone. However, when those effects are common to both channels, they can be viewed as distortions to a common source. Therefore, the non-negativity assumption is seems reasonable for real acoustic environments. Secondly, the sparsity assumption does not hold for the "tail" of the RIR.

Reference	$\mathcal{P}(h_1, h_2)$	$\mathcal{C}(h_1, h_2)$	Note
[Tong et al. 1994]	✗	$\|\mathbf{h}\|_2^2 = 1$	Equivalent to a smallest-eigenvalue problem.
[Kowalczyk et al. 2013]	$\lambda\|\mathbf{h}\|_1$	$\|\mathbf{h}\|_2^2 = 1$	Non-convex due to the quadratic constraint.
[Lin et al. 2008]	$\lambda\|\mathbf{h}\|_1$	$\|h_1[0]\| = 1$	With Bayesian learning for optimal λ.
[Lin et al. 2007]	$\lambda\|\mathbf{h}\|_1$	$h_1[0] = 1, \mathbf{h} \geq 0$	Non-negativity and anchor constraints.
[Aissa-El-Bey and Abed-Meraim 2008]	$\lambda\|\mathbf{h}\|_p^p$	$\|\mathbf{h}\|_2 = 1$	Sparsity enforced by ℓ_p-norm
[Crocco and Del Bue 2015]	$\lambda\left\|\mathbf{p}^{(z)} \odot \mathbf{h}\right\|_1$	$\|\mathbf{h}\|_1 = 1, \mathbf{h} \geq 0$	Iterative weighted ℓ_1-norm
[Crocco and Del Bue 2015]	$\lambda\|\mathbf{h}\|_1$	$\mathbf{p}^{(z)\intercal}\mathbf{h} = 1, \mathbf{h} \geq 0$	Iterative weighted ℓ_1 constraint.

TABLE 5.1: Some state of the art penalties and constraint used in model Eq. (5.7).

Nevertheless, applications concerning RooGE require just the recovery of lower order reflections, i.e. the sparse portion of the RIR. Likewise works in speech enhancement have proven to work under such assumption, thus proving the effectiveness of this approach.

On a similar scheme, in [Kowalczyk et al. 2013], Eq. (5.5) is solved using an adaptive time-frequency-domain approach while [Aissa-El-Bey and Abed-Meraim 2008] proposes to use the ℓ_p-norm instead of the ℓ_1-norm. A successful approach has been presented recently by the authors of [Crocco and Del Bue 2016a], where the anchor constraint is replaced by an *iterative weighted* ℓ_1 equality constraint, i. e., such that at each iteration z, $\mathbf{p}^{(z)\intercal}\mathbf{h}^{(z)} = 1$ [34]. In particular, the method is initialized using the solution of [Lin et al. 2007] and iterated enforcing sparsity using the solution of the previous problem, that is $\mathbf{p}^{(z)} = \hat{\mathbf{h}}^{(z-1)}$. The reader can can found a comprehensive review of these methods in [Crocco and Del Bue 2015; Crocco and Del Bue 2016a].

[34] Note that when $\mathbf{p}^{(z)} = 1$, the constraint returns to the ℓ_1 penalty.

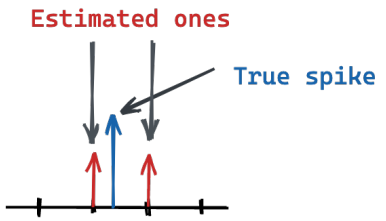


FIGURE 5.2: Schematics of the *bodyguard effect* affecting on-grid approaches.

▸ THE LIMITATION OF THE DISCRETE-TIME METHODS described above are the followings:

- *Basis mismatch*: As explained in § 4.3.3, the filter are not sparse in practice due to the *basis mismatch*. This implies that the true peaks of the filter do not necessarily correspond to the true echoes and lead to followings drawbacks.
- *The estimation is on-grid*: These methods operates fundamentally *on-grid* and return echoes' timings which are integer multiples of the F_s. *Bodyguard effect.* In addition to affecting the AER performance, on-grid methods may converge slowly to suboptimal solutions. In fact, as show in Figure 5.2, instead of estimating the peak at its true location, two smaller "bodyguard" peaks are estimated around it instead. The estimation procedure may stop at this point returning two wrong peaks. Having smaller coefficients, this peaks may not be selected by the subsequent peak pickings technique. Alternatively, the optimization procedure may continue, alternating tuning the weights of the two "bodyguards", without converging to a solution.
- *Computational bottleneck.* A way to cope with the above limitations is to increase the F_s. However this results into a memory and computational bottleneck as several huge (Toeplitz) matrices needs to be built, one

for each pair of microphones. In addition, this leads the risk that the optimization problem becomes ill-conditioned.

5.2.3 *From cross-relation to LASSO*

Integrating the sparse penalty and the constraints proposed by [Lin et al. 2008] (See Table 5.1) in Eq. (5.7), the Blind Sparse Nonnegative Channel Identification (BSN) problem proposed reads

$$\widehat{h}_1, \widehat{h}_2 = \underset{h_1,h_2}{\arg\min}\ \tfrac{1}{2}\|\mathcal{T}(x_1)h_2 - \mathcal{T}(x_2)h_1\|_2^2 + \lambda\|\mathbf{h}\|_1 \text{ s.t. } \begin{cases} \mathbf{h} \geq 0 \\ h_1[0] = 1 \end{cases}. \tag{5.8}$$

This cross-relation based optimization problem can rewritten in the LASSO formulation of Eq. (5.6) as

$$\mathbf{u} = \underset{\mathbf{u}}{\arg\min}\ \tfrac{1}{2}\|\mathbf{v} - \mathbf{B}\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1 \quad \text{s.t.} \quad \mathbf{u} \geq 0,$$

where

$$\mathbf{v} = \mathtt{X}_2\mathbf{e}_1, \quad \mathbf{u} = \begin{pmatrix} h_1[1:] \\ h_2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} -\mathtt{X}_2[:,1:] & \mathtt{X}_1 \end{pmatrix}, \quad \mathtt{X}_i = \mathcal{T}(x_i).$$

Here we used the light, yet common, Python notation for indexing the matrices and vectors.

5.3 Proposed Approach

5.3.1 *Cross-relation in the Fourier domain*

The cross-relation identity Eq. (5.3) ensures that the relation

$$\tilde{\phi}_{\text{LP}} * \tilde{x}_1 * \tilde{h}_2^\star = \tilde{\phi}_{\text{LP}} * \tilde{x}_2 * \tilde{h}_1^\star \tag{5.9}$$

holds even during the introducted measurement process, hence

$$\mathcal{F}(\tilde{\phi}_{\text{LP}} * \tilde{x}_1) \cdot \mathcal{F}\,\tilde{h}_2^\star = \mathcal{F}(\tilde{\phi}_{\text{LP}} * \tilde{x}_2) \cdot \mathcal{F}\,\tilde{h}_1^\star \tag{5.10}$$

where $\mathcal{F}$ denotes the Fourier Transform (FT) described in **??**
In contrast with SIMO-BCE methods that operates in the time domain, here we propose to use Eq. (5.10) in a penalized least-square problem. Such a formulation in the Fourier domain may even be considered as more convenient since the convolution operator is no longer involved. While the FT of $\tilde{h}_i^\star$ can be expressed in closed-form (see Eq. (5.13) below), the FT of $\tilde{\phi}_{\text{LP}} * \tilde{x}_i$ is not available due to the measurement process. To circumvent this issue, following approximation detailed in § 3.2.2, we consider the Discrete Fourier Transform (DFT) of the $\tilde{x}_i$:

$$\mathcal{F}(\tilde{\phi}_{\text{LP}} * \tilde{x}_i)\left(\tfrac{k}{2N}F_s\right) \approx X_i[k] \tag{5.11}$$

for all integers $k \in \{0, \dots, N\}$, where

$$X_i[k] = \sum_{n=0}^{2N-1} x_i[n]\mathrm{e}^{-\mathrm{i}2\pi\frac{kn}{2N}} \tag{5.12}$$

is the DFT of the real vector $\tilde{x}_i$ as defined in **??** for positive frequencies only.

Denoting Δ_τ the following parametric vector of complex exponential

$$\Delta_\tau \stackrel{\text{def}}{=} \left(\mathrm{e}^{-\mathrm{i}2\pi \frac{k}{2N} F_s \tau} \right)_{0 \leq k \leq N} \in \mathbb{C}^{N+1}, \tag{5.13}$$

equation the Fourier-domain cross-relation of Eq. (5.10) evaluated at $f = \frac{k}{2N} F_s$ where $k \in \{0, \dots, N\}$ reads

$$\sum_{r=0}^{R_2-1} \alpha_2^{(r)} X_1 \odot \Delta_{\tau_2^{(r)}} = \sum_{r=0}^{R_1-1} \alpha_1^{(r)} X_2 \odot \Delta_{\tau_1^{(r)}} \tag{5.14}$$

where $\odot$ denotes the component-wise Hadamard product.

5.3.2 *Echo localization with continuous dictionaries*

By interpreting the FT of a Dirac as a parametric *atom*, we propose to cast the problem of RIR estimation into the framework of Continous Dictionary (CD). To this aim, let us define the so-called *parameter set*

$$\Theta \stackrel{\text{def}}{=} [0, T] \times \{1, 2\} \tag{5.15}$$

where T is the length (in time) of the filter. Then, the two desired filters $\tilde{h}_1^\star, \tilde{h}_2^\star$ given by Eq. (5.2) can be uniquely[35] represented by the following discrete measure over Θ

$$\mu^\star = \sum_{i=1}^{2} \sum_{r=0}^{R_i-1} \alpha_i^{(r)} \delta_{(\tau_i^{(r)}, i)}. \tag{5.16}$$

where $\delta_{(\tau_i^{(r)}, i)}$ denotes the Dirac measure which is different from the Dirac function used when modeling the RIRs. The need of defining a measure over the parameter set Θ makes easier the parametrization of the problem in the context of CD. For instance, it is possible to define standard operation over measure used in algorithms to solve the such problems.
Moreover, the rationale behind Eq. (5.15) and Eq. (5.16) is as follows. A couple of filters is now represented by a single stream of Diracs, where we have considered an augmented variable i indicating to which filter the spike belongs. For instance, a Dirac at $(\tau, 1)$ indicates that the first filter contains a Dirac at τ.

[35] Uniqueness is ensured as soon as we impose $\alpha_i^{(r)} > 0 \; \forall i, r$.

The set $\mathcal{M}_+(\Theta)$ of all unsigned and discrete Radon measures over Θ (i. e., the set of all couples of filters) is equipped with the total-variation norm (TV-norm) $\|\mu\|_{\text{TV}}$ [36]. We now define the *linear* observation operator $\mathcal{A} \colon \mathcal{M}_+(\Theta) \to \mathbb{C}^{N+1}$, which is such that

$$\mathcal{A}\delta_{(\tau,i)} = \begin{cases} -X_1 \odot \Delta_\tau & \text{if } i = 1 \\ +X_2 \odot \Delta_\tau & \text{if } i = 2. \end{cases} \qquad \forall (\tau, i) \in \Theta. \tag{5.17}$$

[36] See [Rudin 1987] for a rigorous construction of measures set and the TV-norm.

Therefore, by the linearity of the observation operator $\mathcal{A}$, the relation Eq. (5.14) can be rewritten as

$$\mathcal{A}\mu^\star = \mathbf{0}_{N+1}, \tag{5.18}$$

where $\mathbf{0}_{N+1}$ is a $N+1$-length vector of zeros.
Before continuing our exposition, we note that the anchor constraint can be

written in a more convenient way. Indeed, the constraint $\mu(\{(0,1)\}) = 1$ ensures the existence of a Dirac at 0 in the filter 1. Then, the targeted filter reads

$$\mu^\star = \delta_{(0,1)} + \bar{\mu}^\star \tag{5.19}$$

where $\bar{\mu}^\star$ is a (finite) discrete measure verifying $\bar{\mu}^\star(\{(0,1)\}) = 0$. Denoting $y \overset{\text{def}}{=} -\mathcal{A}\delta_{(0,1)} \in \mathbb{C}^{N+1}$, the relation Eq. (5.18) becomes

$$\mathcal{A}\bar{\mu}^\star = y. \tag{5.20}$$

For the sake of clarity, we use these conventions hereafter and omit the tilde. Now, following [De Castro and Gamboa 2012; Candès and Fernandez-Granda 2014], one can expect to recover the desired filter $\mu^\star$ by solving

$$\widehat{\mu} = \underset{\mathcal{M}_+(\Theta)}{\arg\min} \|\mu\|_{\mathtt{TV}} \quad \text{s.t.} \quad \begin{cases} \mathcal{A}\mu = y \\ \mu(\{(0,1)\}) = 0. \end{cases} \tag{5.21-$\mathcal{P}^0$TV}$$

Note that (5.21-$\mathcal{P}^0$TV) has to be interpreted as a natural extension of the well-known *basis pursuit* problem to the continuous setting. Indeed, for *any* finite discrete measure $\mu = \sum_{r=0}^{R-1} \alpha^{(r)} \delta_{(\tau^{(r)}, i^{(r)})}$, the TV-norm of μ returns to the ℓ_1-norm of the coefficients, i. e., $\|\mu\|_{TV} = \sum_{r=0}^{R-1} |\alpha^{(r)}|$.

Finally, Eq. (5.20) can be exploited to take into account noise during the measurement process (i. e., $n_i \neq 0$ in Eq. (5.1)), as well as approximation errors (see Eq. (5.11)-Eq. (5.14)). In that case, the first equality constraint in (5.21-$\mathcal{P}^0$TV) is relaxed, leading to the so-called Beurling-LASSO (BLASSO) problem

$$\widehat{\mu} = \underset{\mu \in \mathcal{M}_+(\Theta)}{\arg\min} \ \tfrac{1}{2}\|y - \mathcal{A}\mu\|_2^2 + \lambda \|\mu\|_{\mathtt{TV}} \quad \text{s.t.} \quad \mu(\{(0,1)\}) = 0. \tag{5.22-$\mathcal{P}^\lambda_{\mathtt{TV}}$}$$

We emphasize that although continuous Radon measures may potentially be admissible, the minimizers of Eq. (5.22-$\mathcal{P}^\lambda_{\mathtt{TV}}$) are *guaranteed* to be streams of Diracs [Bredies and Carioni 2020, Theorem 4.2]. In addition, although problem Eq. (5.22-$\mathcal{P}^\lambda_{\mathtt{TV}}$) seems to depend on some regularization parameter λ, we describe in § 5.3.5 a procedure to automatically tune it to recover a desired number of spikes. Finally, note that problem Eq. (5.22-$\mathcal{P}^\lambda_{\mathtt{TV}}$) is convex with linear constraints over the parameter set Θ. Therefore, theoretically, the problem can be solved exactly. However, in practice, optimization over space of measures, still complicated because many steps can only be done up to a prescribed precision.

5.3.3 *From LASSO to BLASSO*

In order to better understand the proposed approach based on the BLASSO algorithm, we can present it in light of the LASSO formulation.

$$\underset{\mathbf{u}}{\arg\min} \ \tfrac{1}{2}\|\mathbf{v} - \mathbf{A}\mathbf{u}\|_2^2 + \lambda \|\mathbf{u}\|_1 \ \text{s.t.} \ \mathbf{u} \geq 0$$

$$\downarrow$$

$$\underset{\mathbf{u}}{\arg\min} \ \tfrac{1}{2}\|y - \mathcal{A}\mu\|_2^2 + \lambda \|\mu\|_{\mathtt{TV}} \ \text{s.t.} \ \mu \in \mathcal{M}_+(\Theta)$$

Now, some parallelism can be envisioned:

- *From dictionary to operator*: The matrix $\mathbf{A}$ is typically referred to as *dictionaries*. Then selecting the l-th column of the dictionaries, i. e. $\mathbf{A}\mathbf{e}_l$, means selecting an echo at location l-th w. r. t. the vector $\mathbf{u} = \mathbf{h}[1:]$. In the context of CD, the dictionary is translated into the operator $\mathcal{A}$ thanks to the closed-from of the atom based on the Fourier theory. Therefore, $\mathcal{A}(\delta_\tau)$ can be seen as the selection of an echo at location $\tau \in [0, T]$ ms.
- *Solution*: The LASSO-like approach promotes a solution $\mathbf{u} = \mathbf{h}[1:]$ which is sparse and non-negative vector. The last one, ensured by the non-negativity constraint. In the BLASSO, this is translated assuming the spike measure assumed for the channels, namely, $\mu = \sum_r \alpha^{(r)}\delta(t-\tau^{(r)})$.
- *Sparsity*: while in the initial case, the sparsity is enforced by the ℓ_1-norm, in the second case it is pursued with the TV-norm.
- *Solver*: the former optimization problem can be solved with standard LASSO solvers, while for the latter a gradient-descent algorithm is used.

5.3.4 *The resulting algorithm*

The algorithm used to solve Eq. (5.22-$\mathcal{P}_{\text{TV}}^{\lambda}$) is an instance the sliding Frank-Wolfe algorithm proposed in [Denoyelle et al. 2019] to solve Eq. (5.22-$\mathcal{P}_{\text{TV}}^{\lambda}$). Detailed descriptions of the steps of the algorithm are given in **??**. In a nutshell, the algorithm iterative over the following steps until a condition on the cost function is met.

1. *Anchor constrain.* At first the anchor constraint in added arbitrarily on one of the two filters. This is used to initialize the two filters;
2. *Local cost based on Cross-relation* For both the filters, a local cost-function derived from the cross-relation for both the filters is computed; At this step either the initialization or previously found solution are used;
3. *Find the maximizer* a new candidate echo's location is found as maximizer among the two local cost functions of the previous step;
4. *Update the amplitudes* By solving a non-negative LASSO problem, all the echo's amplitude coefficients estimated until this point are updated;
5. *Joint refinement* The position and the coefficient of the current solution are jointly refined to ease numeric resolution using the original cost function.
6. *Current solution and repeat* The algorithm stops as soon as an iterate satisfies the first order optimality condition associated to the convex problem; if not, the algorithm iterates from step 2. using the current solution as input.

These steps are illustrated in Figure 5.3.

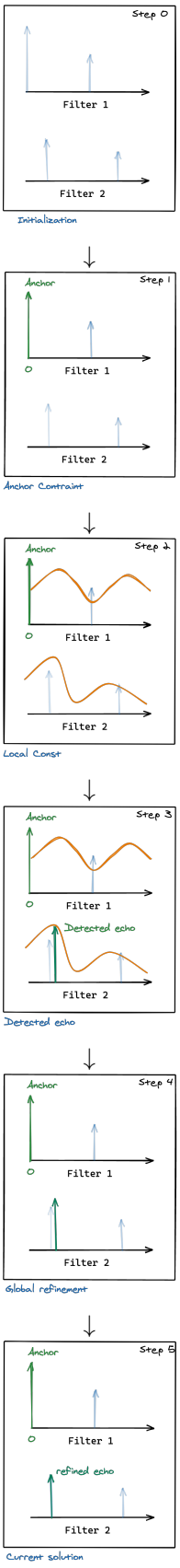


FIGURE 5.3: Illustration of the sliding Frank-Wolfe algorithm proposed in [Denoyelle et al. 2019] in BLASTER.

5.3.5 *Homotopic path for λ estimation*

Existing works, as well as the proposed one, relies of the regularization λ parameter to weight the sparsity penalty. However, this becomes an hyperparameter that needs to be carefully tuned according to the input data. Instead, we propose to compute a *path of solutions* to automatically estimate it in Eq. (5.22-$\mathcal{P}_{\text{TV}}^{\lambda}$). In the context of sparse optimization this technique is also referred to as *homotopic path*. More precisely, let $\lambda_{\max}$ be the smallest value of λ such that the null measure is the solution to Eq. (5.22-$\mathcal{P}_{\text{TV}}^{\lambda}$). It can be shown that $\lambda_{\max}$ is upper bounded by $\max_{\theta\in\Theta}|y^{\intercal}\mathcal{A}\delta_{\theta}|$ (See **??**). Starting from $z = 1$ and the empty filter, we consider a sequential implementation where the solution of Eq. (5.22-$\mathcal{P}_{\text{TV}}^{\lambda}$) is computed for $\lambda^{(}z) = 10^{-0.05z}\lambda_{\max}$ until the desired number of spikes is found in each channel when incrementing z. For each $\lambda^{(}z)$, we search for a solution of Eq. (5.22-$\mathcal{P}_{\text{TV}}^{\lambda}$) with the solution obtained for $\lambda^{(z-1)}$ as a warm start.

5.4 Experiments

The proposed method (BLASTER) is compared against the non-negative ℓ_1-norm method (BSN) of [Lin et al. 2007] and the iterative ℓ_1-norm approach (IL1C) described in [Crocco and Del Bue 2016a]. The problem is formulated as estimating the time location of the first $R = 7$ strongest components of the RIRs for 2 microphones listening to a single sound source in a shoebox room. It corresponds to the challenging task of estimating first-order early reflections. The robustness of the methods is tested against different level of noise (SNR) and reverberation time ($\texttt{RT}_{60}$).

The quality of the AER estimation is assessed in terms of precision[37] in percentage as in the literature of onset detection [Böck et al. 2012] and the RMSE in samples. Both metrics evaluate only the *matched* peaks, where a *match* is defined as being within a small window $\tau_{\max}$ of a reference delay. These two metrics are similar to the ones used in [Crocco and Del Bue 2015].

[37] Since only K time locations are considered in both the ground truth and the estimation, precision and recall are equal.

For this purpose we created three synthetic datasets of 1000 observations each, which are summarized in Table Table 5.2.

Dataset	Signals	SNR [dB]	$\texttt{RT}_{60}$ [s]
$\mathcal{D}^{\text{(valid)}}$	broadband noise	⚄	⚄
$\mathcal{D}^{\texttt{SNR}}$	broadband noise, speech	⚂	400 ms
$\mathcal{D}^{\texttt{RT}_{60}}$	broadband noise, speech	20 dB	⚂

TABLE 5.2: Summary of the dataset used for evaluation. ⚄ and ⚂ stands for randomly sampled from a continuous and discrete set of values respectively

$\mathcal{D}^{\text{(valid)}}$ is used for tuning the hyperparameter λ and the peak-picking parameters for IL1C and BSN using $\texttt{RT}_{60}$ and SNR randomly drawn from $\mathcal{U}[0, 1]$ (sec) and $\mathcal{U}[0, 20]$ (dB) respectively; $\mathcal{D}^{\texttt{SNR}}$ features SNR value uniformly sampled in $[0, 6, 14, 20, \infty]$ while the $\texttt{RT}_{60}$ is kept fixed to 400 ms; akin the $\mathcal{D}^{\texttt{RT}_{60}}$ is built sampling $\texttt{RT}_{60}$ value uniformly in $[200, 400, 600, 800, 1000]$ ms keeping SNR fix to 20 dB. Moreover, while for $\mathcal{D}^{\text{(valid)}}$ broadband signals (white noise) are used as the source, for $\mathcal{D}^{\texttt{SNR}}$ and $\mathcal{D}^{\texttt{RT}_{60}}$ speech utterances from the TIMIT dataset are also included. The signal duration is kept fixed to 1 s with sampling frequency $F_s = 16$ kHz. For a given $\texttt{RT}_{60}$ value and room with random

dimensions, a unique absorption coefficient is assigned to all surfaces based on the Sabine's formula. Then, the two microphones and the source are randomly positioned inside the room. The parameters of such audio scene are then passed as input to the `pyroomacoustics` simulator [Scheibler et al. 2018b], which returns the corresponding RIRs as well as the *off-grid* echo delays and attenuation coefficients computed with the Image Source Method (ISM) [Allen and Berkley 1979]. Note that when generating the data, no samples have been pruned to match any minimal separation condition. To generate the microphone signals, an over-sampled version of the source signal is convolved with ideal RIRs at high frequency ($F_s = 1024$ kHz) made up of on-grid Diracs. The results are later resampled to meet the original F_s and Gaussian white noise is added to meet the given SNR value.

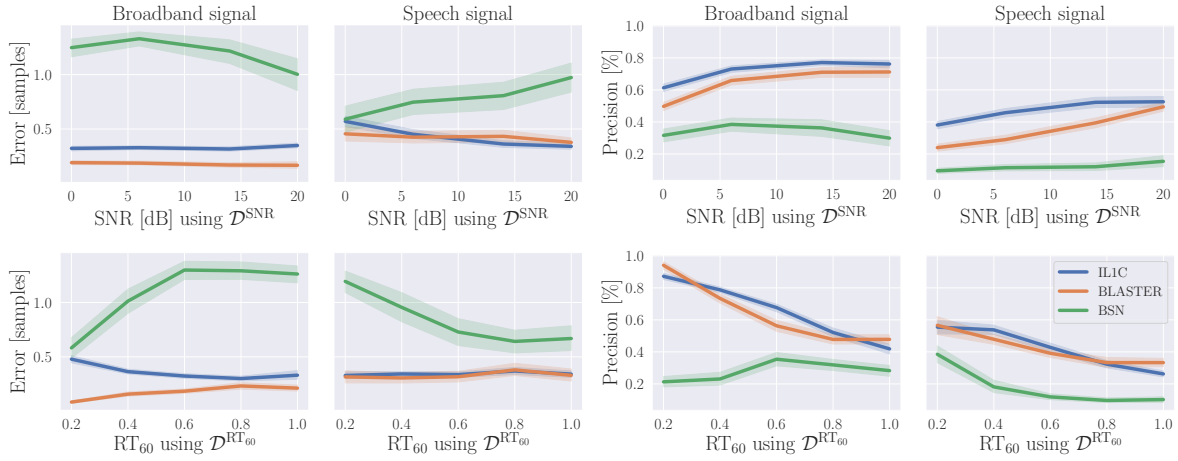


FIGURE 5.4: Line plot with error bands for error (left) and precision (right) versus SNR level (top) and RT$_{60}$ level (bottom) using broadband and speech signals for the task of recovering $R = 7$ echoes. A threshold of $\tau_{\text{max}} = 2$ samples is used to compute the precision.

	Precision [%]									
	R = 2 echoes					R = 7 echoes				
τ_{max}	0.5	1	2	3	10	0.5	1	2	3	10
BSN	8	9	27	46	62	5	8	38	54	73
IL1C	51	55	55	56	58	42	53	55	56	58
BLASTER	**68**	**73**	**74**	**75**	**75**	46	53	56	57	61

TABLE 5.3: Precision for different threshold τ_{max} in samples for the recovery of $R = 2$ and 7 echoes, RT$_{60}$ = 200 ms and SNR = 20 dB.

▶ QUANTITATIVE RESULTS are reported in Figure 5.4, Figure 5.5 and Table 5.3. Here, for both RMSE and Precision and for both broadband and speech signal, the metrics are displayed against the dataset parameters. We observe that BSN performs worst in all tested conditions, possibly due to its strong reliance on the peak picking step. For $R = 7$ or higher, BLASTER yields similar or slightly worse performance than IL1C for the considered noise and reverberation levels, with decreasing performance for both as these levels increase. Using speech rather than broadband signals also yields worse results for all methods. However, the echo timing RMSE is significantly smaller using BLASTER due to its off-grid advantage. We also note that BLASTER significantly outperforms IL1C on the task of recovering $R = 2$ echoes. As showed in Tab. 5.3, in mild conditions, up to 68% of echoes can be retrieved by BLASTER with errors lower than half a sample in that case. This is promising since the practical advantage

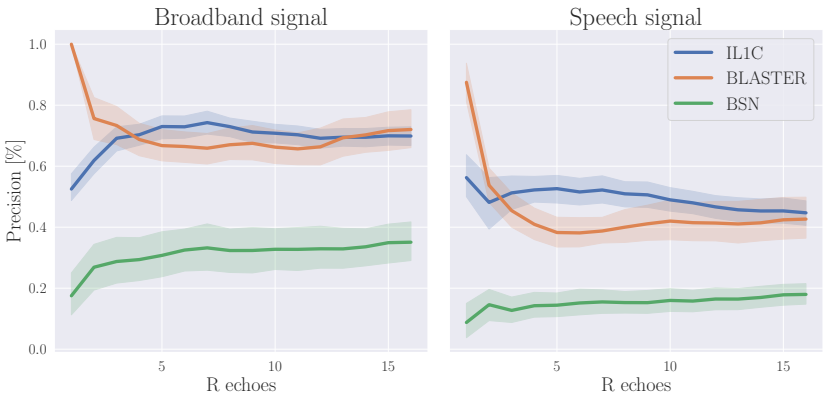


FIGURE 5.5: Line plots with error bands of precision versus number of echoes R to be retrieved for broadband (left) and speech (right) signals with $\mathtt{RT}_{60}$ = 400 ms and SNR = 20 dB.

of knowing the timing of two echoes per channel has been demonstrated in [Di Carlo et al. 2019b; Scheibler et al. 2018d].

5.5 Conclusion

A novel blind, off-grid, multichannel echo retrieval method has been proposed based on the framework of continuous dictionaries. Comparisons with state-of-the-art approaches on various noise and reverberation conditions show that this method performs best when the number of echoes to retrieve is small. While some robustness to noise, reverberation, and non-broadband signals is observed, our experiments reveal that room for improvement exists for this challenging and emerging topic. Future works will include an extension to more than two channels and experiments on real-world data.

6

Data-driven Acoustic Echo Retrieval & `Lantern`

Keywords: Acoustic Echo Retrieval, TDOA Estimation, Supervised Learning, Deep Learning, Regression.
Resources:

- Paper
- Code
- Poster

Di Carlo et al., "Mirage: 2d source localization using microphone pair augmentation with echoes"

▶ Synopsis This chapter

The material presented in this chapter is part of the previously published work [Di Carlo et al. 2019b] and of a technical report for HONDA® [*HRI-JF collaboration - Final Phase II Deliverable*].

6.1 Introduction

The following sections gives a review of machine learning theory knowledge required by the reader in order to understand the implementations related to machine learning in this chapter. The review includes basic theory behind neural networks and deep learning including layer-types, optimization and loss functions, as well as aspects related to training on Room Impulse Response (RIR). This section also gives a brief review of why and how to use autoencoders.

6.1.1 *Supervised Learning*

▶ End-to-end learning

▶ 2-stage learning

▶ Virtually Supervised Learning

6.1.2 *Neural Networks*

▶ Convolutional Neural Networks and Deep Learning

6.1.3 *For the RIR and the AER?*

6.2 Proposed Learning-based AER

6.2.1 *Simple Case: $R = 2$*

Our approach is to train a deep neural network (DNN) on a dataset simulating the considered close-surface scenario. We model the problem as multi-target regression, with *interaural level difference* (ILD) and *interaural phase difference* (IPD) as input features, and $V \in \mathbb{R}^3$ as output parameters. ILD and IPD features are defined in the frequency domain as follows:

$$\begin{cases} ILD(f) = & \frac{1}{T}\sum_{t=1}^{T} \log\left|\frac{M_2(f,t)}{M_1(f,t)}\right| \\ IPD(f) = & \frac{1}{T}\sum_{t=1}^{T} \frac{M_2(f,t)/|M_2(f,t)|}{M_2(f,t)/|M_1(f,t)|} \end{cases} \tag{6.1}$$

More precisely, the input of the network is $\mathbf{x} = [ILD, \mathrm{Re}(IPD)], \mathrm{Im}(IPD)]$, where Re and Im denote real and imaginary part operators, respectively. Note that for the IPD, the frequency $f = 0$ is discarded because it is constant for every observation. In general, the mapping between V and the proposed feature is not unique. In particular, this happen when $\tau_2^1 = \tau_1^1$. In order to avoid this, we preventively pruned all the entries with $|\tau_2^1 - \tau_1^1| < 10^{-6}$ from the dataset.

We use a simple fully-connected DNN architecture consisting of a D-dimensional input layer, a 3-dimensional output layer, and 3 fully connected hidden layers with respective input sizes 500, 300 and 50. Rectified linear unit (ReLU) activation functions are used except at the output layer, and each hidden layer has a dropout probability $p_{\text{do}} = 0.3$. We use the mean square error loss function for training and the Adam optimizer [**kingma2014adam**]. The normalized root mean square error (nRMSE) is taken as validation metric[1]. The network is manually tuned on a validation set to find the best combination of number of hidden layers, their sizes and p_{do}. Once time delay estimates $\hat{V}$ are returned by the DNN, they are converted to synthetic local angular spectra and passed to Ψ_{SRP} (See Sec. § 10.2.2) together with the relative positions of true and image microphones which are assumed known. We call this algorithm MIRAGE. The synthetic local angular spectra consist of Gaussians centered at $\hat{V}$ and with variances equal to the prediction errors made by the DNN on the validation set.

[1] The nRMSE takes values between 0 (perfect fit) and ∞ (bad fit). If it is equal to 1, then the prediction is no better than a constant.

6.3 Robust learning for the case $R = 2$

The neural network follows the convolutional neural network (CNN) architecture in Figure **??**, which is the one also used in [**Nguyen2018**] and similar to the one used in [**Chakrabarty2017**]. It consists in two convolutional modules made of one-dimensional convolutional layer (1DConv) followed by max-pooling along the frequencies, followed by rectified linear unit (ReLu) activation function and batch-normalization. The second part consists in a cascade of fully connected feed-forward (FF) layers. Note that dimension of the input is re-arranged so that the second dimension is considered as channel for the 1DConv. After each layer a dropout probability $p_{\text{do}} = 0.3$ is applied.

The proposed novel loss function is the negative Student-T log-likelihood, which is implemented as follows:

$$\mathcal{L}(\Theta) = \sum_{x \in B} \sum_{t \in V} \frac{1}{2} \log(\nu_t \pi_t) + \frac{1}{2} \log(\lambda_t^2) - \log \Gamma\left(\frac{\nu_t + 1}{2}\right) + \log \Gamma\left(\frac{\nu_t}{2}\right) + \frac{\nu_t + 1}{2} \log\left(1 + \frac{\|\mu_t, x_i)\|}{\nu_t \lambda_t^2}\right) \tag{6.2}$$

where Θ are the CNN parameters and Γ is the Gamma function. The summation over i corresponds to the sum among of all the sample x of the batch B. and the summation over t corresponds to the sum among the three quantities in V (TDOA, iTDOA, TDOE). It follows that for each each input the network will return the parameters of 3 Student-T distribution $(\mu_t, \nu_t, \lambda_t)$ for each variable $t = \text{TDOA}, \text{iTDOA}, \text{TDOE}$. Hereafter we denote with $V_{\mathcal{ST}}$ the set of the 9 network outputs.

We use the Adam optimizer ant the normalized root mean square error (nRMSE) is taken as validation metric (see Section **??**). The network is manually tuned on a validation set to find the best combination of number of hidden layers and their sizes

Once an estimate $\hat{V_{\mathcal{ST}}}$ of the parameters of the 3 distribution is returned by the CNN, they are converted to synthetic local angular spectra and passed to an SRP-PHAT method together with the relative positions of true and image microphones which are assumed known. We call this algorithm MIRAGE. The synthetic local angular spectra consist of Student-t distribution with parameters μ, ν and λ.

For training and validation of the CNN we generate many random shoe-box room configurations using the software presented in [Schimmel et al. 2009]. This software implements both the image-method for simulating reflections and a ray-tracing algorithm for diffusion. Room widths are uniformly drawn at random in $[3, 9]$ m, heights in $[2, 4]$ m. Random source/microphones positions and absorption coefficients for the 6 surfaces are used, respecting the close-surface scenario. In particular, the microphones are at most 30 cm from the close-surface, placed 13 cm from each other, the absorption coefficients of the other walls are uniformly sampled in $(0.5, 1)$ and the one of the close-surface is in $(0, 0.5)$. The same realistic diffusion profile [Gaultier et al. 2017] is used for all surfaces. Around $20,000$ audio scenes are generated this way, yielding reverberation times (RT_{60}) between 20 ms and 250 ms.

Schimmel et al., "A fast and accurate "shoe-box" room acoustics simulator"

Gaultier et al., "VAST: The virtual acoustic space traveler dataset"

For training and validation, the RIRs are convolved with 1 sec of white-noise with additional noise with SNR in $(0, 20)$ dB. All signals and RIRs are sampled at 16 kHz. The STFT is performed on 1024 point with 50% overlap. Finally the features are computed as in (**??**) yielding a vector of size $D = 1534$ for each observation. While we validate the CNN on a portion of the dataset in a *holdout* fashion, the test is conducted on 200 new RIRs convolved with both speech utterances. This set is generated similarly to the training and validation sets. Moreover the recordings are perturbed by external white noise as in the training set. The speech signals are normalized speech utterances of various lengths (from 1 s to 6 s), randomly selected from the TIMIT corpus. A re-implement version of SRP-PHAT is used to aggregate local angular spectra obtained from the DNN's output and as a baseline. However the original MATLAB code for SRP-PHAT can be found at `http://bass-db.gforge.inria.`

`fr/bss_locate/`. A sphere sampling with 1 degree resolution and coordinates $\theta \in [-179, 180]$ and $\phi \in [0, 90]$ degrees is used for the DOA search.

6.4 TOWARDS THE CASE $R > 2$

6.4.1 *Better features: RTF*

6.4.2 *Better architecture: Physical-based learning and unfolding*

6.5 CONCLUSION AND PERSPECTIVE

7

Datasets for Acoustic Echo Estimation & `dEchorate`

> "*Signal, a function that conveys information about a phenomenon. [. . .] Consider an acoustic wave, which can convey acoustic or music information.*"
> —R. Priemer, *Introductory Signal Processing*

▶ Synopsis This chapter presents `dEchorate`: a new database of measured multichannel room impulse response (RIRs) including annotations of early echoes and 3D positions of microphones, real and image sources under different wall configurations in a cuboid room. These data provide a tool for benchmarking recent methods in *echo-aware* speech enhancement, room geometry estimation, RIR estimation, acoustic echo retrieval, microphone calibration, echo labeling and reflectors estimation. The database is accompanied with software utilities to easily access, manipulate and visualize the data as well as baseline methods for echo-related tasks.

Keywords: Room impulse response, Early reflection, Acoustic echoes, Audio database, Microphone arrays.
Resources:

- Code
- Repository

The material presented in the chapter are results of a work done while visiting prof. Sharon Gannot and ing. Pinchas Tandeitnik at the Bar'Ilan University, Israel. The work described here, together with its continuation described in **??** will be submitted as a journal article to the EURASIP special edition *Data-driven ASP: Methods and Apps*.

7.1 Introduction

As discussed § 4.4.1, many RIRs datasets are available online. However, most of them are specifically designed for application either for Speech Enhancement (SE)W or for Room Geometry Estimation (RooGE). The main common drawback of these datasets in that they can not be easily used for other tasks than the one which they were designed for. In particular, SE-oriented dataset lack of of proper annotation of echoes in the RIRs or the absolute position of object inside the rooms. Alternatively, the dataset for RooGE focuses typically scenarios which are not suitable for SE application. The `dEchorate` was designed to fill this gap: a fully calibrated multichannel RIR database with accurate annotation of the geometry and echoes in different configurations of a cuboid rooms with varying wall acoustic profiles. The database currently features 1800 annotated RIRs obtained from 6 arrays of 5 microphones each, 6 sound sources in 10 different acoustic conditions. All the measurements were realized in the acoustic lab at Bar-Ilan university following a consolidated protocol previously established for the realization of two other multichannel RIRs databases: the BIU's Impulse Response Database [Hadad et al. 2014] gathering RIRs of different reverberation levels sensed by uniform linear arrays

FIGURE 7.1: Broad-view picture of acoustic lab at Bar-Ilan university.

(ULAs); and MIRaGE [Čmejla et al. 2019] providing a set of measurements for a source position that can be place in in dense position grid. dEchorate is designed for AER with linear arrays, and is more generally aimed at analyzing and benchmarking RooGE and echo-aware signal processing methods on real data. In particular, it can be used to assess robustness against the number of reflectors, the reverberation time, additive spatially-diffuse noise and non-ideal frequency and directive characteristics of microphone-source pairs and surfaces in a controlled way. Due to the amount of data and recording conditions, it could also be used to train machine learning models or as a reference to improve RIR simulators. The database is accompanied with a Python toolbox that can be used to process and visualize the data, perform analysis or to annotate new datasets.

7.2 Database realization

7.2.1 *Recording setup*

The recording setup is situated in a cuboid room with dimension 6 m $\times$ 6 m $\times$ 2.4 m. The 6 facets of the room (walls, ceiling, floor) are covered by acoustic panels allowing controllable reverberation time ($\mathtt{RT}_{60}$). We placed 4 directional loudspeakers (direct sources) facing the center of the room and 30 microphones mounted on 6 non-uniform linear arrays (nULA) of 5 sensors each. An additional channel is used for the loop-back signal, which serves to compute the time of emission and detect errors. Each loudspeaker and each array was positioned close to one of the walls in such a way that the nature of the strongest echo can is easily identifiable. Moreover, their positioning was chosen to cover a wide distribution of source-to-receiver distances, hence, a wide range of direct-to-reverberant ratio (DRR). Further, 2 more loudspeakers were positioned pointing towards the walls (indirect sources). This was done to study the case of early reflections being stronger than the direct-path. Each linear microphone arrays consists in 5 microphones with non-uniform inter-microphone spacings of $[4, 5, 7.5, 10]$ cm[38]. Each array is steered towards a different vertical edge of the room for calibration and reproducibility purposes.

[38] i. e., $[-12.25, -8.25, -3.25, 3.25, 13.25]$ cm w.r.t the barycenter

Loudspeakers	(directional, direct) 4× Avanton (directional, indirect) 2× Avanton (omnidirectional) 1× B&G (babble noise) 4× 6301bx Fostex
Microphones	30× AKG CK32
Array	6× nULA (5 mics each, handcrafted)
A/D Converter	ANDIAMO.MC
Indoor Positioning	Marvelmind Starter Set HW v4.9

TABLE 7.1: Technical specification of the measurements equipment used in the recordings.

7.2.2 *Measurements*

The main feature of this room is the capability to change the acoustic profile of the each of its facet by flipping double-sided panels with one reflective and one absorbing face. This allows to achieve precise values of $\mathtt{RT}_{60}$ that ranges from 0.1 to almost 1 second. In this dataset the panels of the floor were kept always absorbent.

Two types of sessions were considered, namely, *one-hot* and *incremental*. For the first type, a single facet was placed in reflective mode while all the others were kept absorbent. For the second type, starting from fully-absorbent mode, facets were progressively switched to reflective one after the other until all but the floor are reflective, as shown in Table 7.2.

The dataset features an extra recording session. For this session, office fornitures where positioned in the room to simulate the a typical meeting room with chairs, tables (See Figure 7.1). This recordings will be used in future works for asserting the robustness of echo-aware methods in case of real-world scenario.

For each room configuration and loudspeaker, three different excitation signals were played and recorded in sequence: chirps, white noise and speech utterances. The former consists in a repetition of 3 ESS signals of duration 10 seconds and frequency range from 100 Hz to 14 kHz interspersed with 2 seconds of silence. Such frequency range was chosen to match the characteristics of the loudspeakers. To prevent rapid phase changes and "popping" effects, the signals were linearly faded in and out over 0.2 seconds with a Tuckey taper

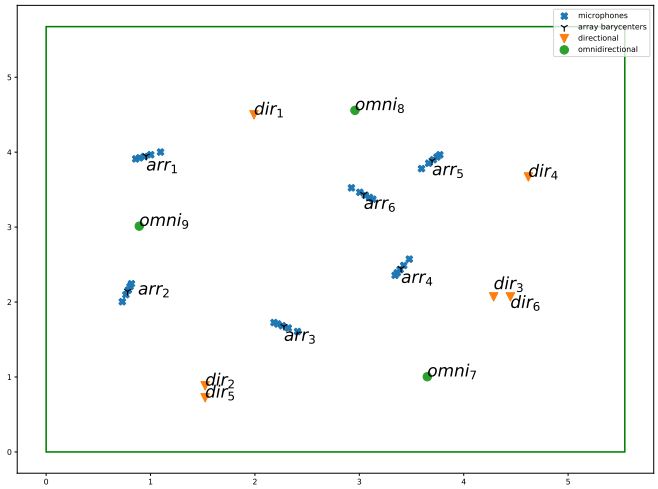


FIGURE 7.2: Illustration of the recording setup - top view.

(a) Overall setup

(b) Microphone array

(c) Revolving panels

FIGURE 7.3: Picture of the acoustic lab. From left to right: the overall setup, one microphone array, the setup with revolved panels.

	Surfaces:	Floor	Ceiling	West	South	East	North
one-hot	000000	✗	✗	✗	✗	✗	✗
	010000	✗	✓	✗	✗	✗	✗
	001000	✗	✗	✓	✗	✗	✗
	...			...			
	000001	✗	✗	✗	✗	✗	✓
incremental	011000	✗	✓	✓	✗	✗	✗
	011100	✗	✓	✓	✓	✗	✗
	...			...			
	011111	✗	✓	✓	✓	✓	✓

TABLE 7.2: Surface coding in the dataset: each binary digit indicates if the surface is absrobent (0, ✗) or reflective (1, ✓).

window[39] Secondly, 10 seconds bursts of white noise and 3 anechoic speech utterances from the Wall Street Journal (WSJ) dataset [Paul and Baker 1992] were reproduced in the room. Through all the recordings, at least 40 dB of sound dynamic range was asserted and room temperature of $24^\circ \pm 0.5^\circ$ and humidity of 80% were registered. Moreover 1 minute of *room tone* (silence) and 4 minutes of diffuse babble noise were recorded for each session. The latter was simulated by transmitting different chunks of the same single-channel babble noise recording from additional loudspeakers facing the four corners of the room.

All the microphone signals were synchronously acquired and digitally converted to 48 kHz with 32 bits/sample using the equipment listed in Table 7.1. The polarity of each microphone was registered by clapping a book in the middle of the room.

[39] The code to generate the reference signal and to process them is available at the Database pepository. Such code is based on `pyrirtool` Python library.

7.3 Dataset annotation

RIRs are estimated with the ESS technique [Farina 2007]: the signal of a microphone recording an ESS source is deconvolved by division in the frequency domain. Notice that the Fourier transform of the ESS signal used at the denominator is available in closed form.

7.3.1 *RIRs annotation*

The objective of this database is to feature annotations in the "geometrical space", namely the microphone and source positions, *fully consistent* with

annotations in the "signal space", namely the echo timings within the RIRs. This results is achieved as follows:

(i) First, a the ground truth positions of array and source centres are acquired via a Beacon indoor positioning system (bIPS). This system consists in 4 stationary bases positioned at the corners of the ceiling and a moving probe used for measurements which can be located within errors of ± 2 cm. The elements of this system are shown in Figure 7.4.

FIGURE 7.4: Picture of the Beacon indoor positioning system used for measuring array and loudspeaker 3D position.

(ii) The estimated RIRs are superimposed on synthetic RIRs computed with the ISM from the geometry obtained in the previous step. A Python GUI[40] (showed in Figure 7.5), was used to manually tune a peak finder and *label* there echoes, that is annotate their positions and their correspondent wall.

(iii) By solving a "simple" multi-dimensional scaling (MDS) problem [Dokmanić et al. 2015; Crocco and Del Bue 2016b; Plinge et al. 2016], refined microphone and source positions were computed. The non-convexity of the problem was alleviated by using a good initialization (obtained at the previous step), by the high SNR of the measurements and, later, by including the additional image sources in the formulation. The prior information about the arrays' structures reduced the number of variables of the problem, corresponding to the 3D positions of the sources and of the arrays' barycenters in addition to the the arrays' tilt on the azimuthal plane.

(iv) By employing a multilateration algorithm [Beck et al. 2008], where the position of one microphone per array served as anchors and the TOAs are converted into distances, it was possible to localize the image sources along side with the real. This step will be further discussed in **??**.

Knowing the geometry of the recording room, we were able to manually label the echoes by iterating through steps (ii), (iii) and (iv).

[40] This GUI is available in the dataset package.

▶ THE FINAL GEOMETRICAL AND SIGNAL ANNOTATION was chosen as a compromise between the bIPS measurements and the MDS output. While the formers are noisy but consistent with the scene's geometry, the latters match the TOAs but not necessarily the physical world. In particular, the geometrical ambiguities such as global rotation, translation and up-down ambiguities were observed. Instead of manually correcting this error, we modified the original problem from using only the direct distances (dMDS) to considering the image sources' TOA of the ceiling in the cost function (dcMDS). Table 7.3 shows numerically the *mismatch* (in cm) between the geometric space (defined by the bIPS measurements) and the signal space (the one defined by the echo timings, converted in cm). To better quantify it, we introduce here the *goodness of match* (GoM): it measures the fraction of (first-order) echo timings annotated on the RIRs matching the annotation produced by the geometry within a threshold. Including the ceiling information, MDS produces a geometrical configuration which has a small mismatch (0.41 cm in average) in both the signal *and* geometric spaces with 98.1% of matching first order echoes

	Metrics	bIPS	dMDS	dcMDS
Geom.	Max.	-	6.1	1.07
	Avg.±Std.	-	1.8 ± 1.4	0.39 ± 0.2
Signal	Max.	5.86	1.20	1.86
	Avg.±Std.	1.85 ± 1.5	0.16 ± 0.2	0.41 ± 0.3
Mismatch	GoM (1.0 ms)	97.9%	93.4%	98.1%
	GoM (0.1 ms)	26.6%	44.8%	53.1%
	GoM (0.05 ms)	12.5%	14.4%	30.2%

TABLE 7.3: Mismatch between geometric measurements and signal measurements in terms of maximum (Max.), average (Avg.) and standard deviation (Std) of absolute mismatch in centimeters. The *goodness of match* (GoM) between the signal and geometrical measurements is reported as fraction of matching echo timing for different threshold in milliseconds.

within 1 ms window. Nevertheless, it is interesting to see that already the bIPS measurements produces a good but less precise annotation.

7.3.2 *Other tools for RIRs annotation*

Finally, we want to mention that the following tools and techniques were found helpful in annotating the echoes:

- THE skyline VISUALIZATION consists in presenting multiple RIRs as an image, such that the wavefronts corresponding to echoes can be highlighted [Baba et al. 2018]. More precisely, it is the visualization of the $L \times N$ matrix $\mathbf{H}$ created by stacking column-wise N normalized echograms[41], that is $\mathbf{H}_{l,n} = \bar{\eta}_n(l) = |h_n(l)| / \max |h_n(l)|$, where $l = 0, \ldots, L-1$ is the sample index and n is an arbitrary indexing of the all microphones for a fix room configuration. 4 RIR skylines for 4 directional sources for the full reflective scenario are shown in Figure 7.6, stacked horizontally, preserving the order of microphones within the arrays. Thus, the reader can notice several clusters of 5 adjacent points of similar color (intensity) corresponding to the arrivals at the array's sensors. Thanks to the usage of linear arrays, this visualization allowed us to identify both TOAs and their labeling.

[41] The echogram is defined either as the absolute value or as the squared value of the RIRs.

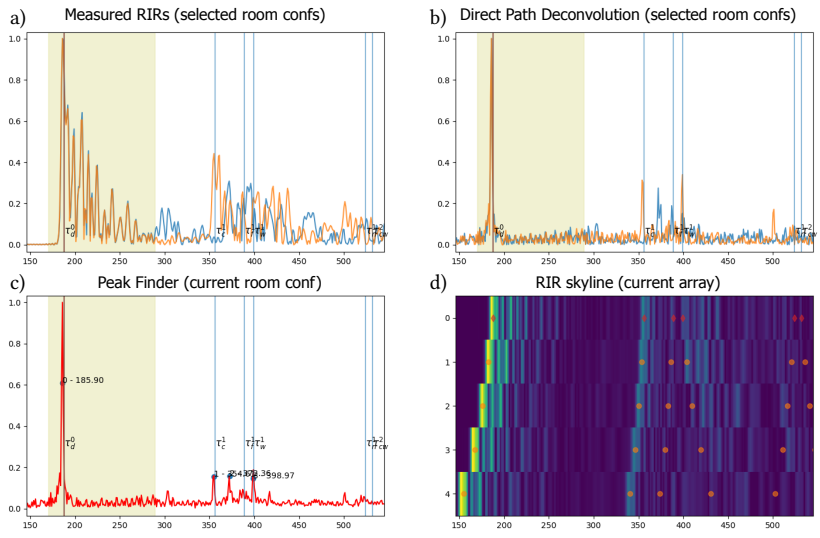


FIGURE 7.5: Detail of the GUI used to manually annotate the RIRs. For a given source and microphone, a) and b) shows 2 RIR for 2 different room walls configuration (blue and orange) before and after the direct path deconvolution respectively. c) shows the results of the peak finder of the equalized RIR and d) is a zoom on the RIR skyline (See Figure 7.6).

- DIRECT PATH DECONVOLUTION was used for compensating the frequency response of the source loudspeaker and microphone [Antonacci et al. 2012; Eaton et al. 2016]. In particular, the direct path of the RIR was manually isolated

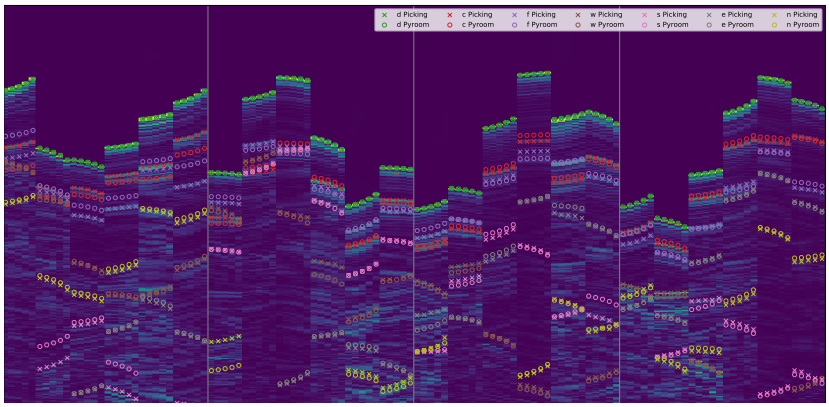


FIGURE 7.6: RIR `Skyline` annotated with observed peaks (×) together with their geometrically-expected position (○) computed with `Pyroomacoustic` simulator. As specified in the legend, different colors are used to indicate the room facets responsible for the reflection: direct path (`d`), ceiling (`c`), floor (`f`), west wall (`w`), . . . , north wall (`n`).

and used as an equalization filter for enhancing early reflections from their superimposition and from background noise before proceed with peak picking. Each RIR was equalized with its relative direct path. As depicted in Figure 7.5, in some situation this process was necessary for correctly identifying the underlying TOAs' peaks.

- DIFFERENT WALL COMBINATIONS for the same geometry influenced the peaks' predominance in the RIR, hence hence facilitating its echo annotation. An example of RIRs corresponding to 2 different surface configurations is shown in Figure 7.5: the reader can notice how the peak prominence change for the different configurations.

- THE INTERPOLATION-BASED PEAK FINDER[42] was used on the normalized echograms $\bar{\eta}_n(l)$ to sightly compensate the sampling process. In [Remaggi et al. 2016] a method that automatically extract peaks in RIRs is proposed. However, in practice, the manual peak finding was found easier and more robust.

[42] In this work, peaks are found using the Python `peakutils` library.

7.3.3 *Limitations of current annotation*

As stated in [Defrance et al. 2008b], we want to emphasize that annotating the correct TOAs of echoes and even the direct path in "clean" real RIRs is far from straightforward. The peaks can be blurred out by the loudspeaker characteristics or the concurrency of multiple reflections. However as showed in Figure 7.6, the proposed annotation was found to be sufficiently consistent both in the geometric and the echo in the echo space. Thus, no further refinement was done. This database can be used as a first basis to develop better AER methods which could be used to iteratively improve the annotation, for instance including 2nd order reflections.

7.4 THE `dEchorate` PACKAGE

The dataset comes with both data and code to parse and process them. The data are presented in 2 modalities: the `raw` data, that is, the collection of recorded wave files, are organized in folders and can be retrieved by querying a simple database table; the `processed` data, which comprise the estimated RIRs and the geometrical and signal annotations, are organized in tensors directly importable in Matlab or Python (*e.g.* all the RIRs are stored in a tensor

scr	mic	signal	floor	...	filename
1	1	chirp	False	...	2020-01-22__22-50-36.wav
1	1	speech	False	...	2020-01-22__22-59-36.wav
⋮	⋮	⋮	⋮	⋮	⋮

FIGURE 7.7: Sample view of the database table to retrieve the raw wave file and its attributes.

of dimension $L \times I \times J \times D$, respectively corresponding to the RIR length in samples, the number of microphones, of sources and of room configurations). Together with the data a Python package is available at the same website. This includes wrappers, GUI, examples as well as the code to reproduce this paper. In particular, all the scripts used for estimating the RIRs and annotating them are available and can be used to further improve and enrich the annotation or as baselines for future works.

7.5 CONCLUSIONS

This work introduced a new database of Room Impulse Response (RIR) featuring accurate annotation of early echoes and microphone positions. These data can be used to test methods in the room geometry estimation pipeline and in echo-aware audio signal processing. We will show some application in SE and RooGE in **??**.

Part IV

ECHO-AWARE APPLICATION

8

Application of Acoustic Echoes

- Synopsis Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

8.1 Overview

8.1.1 *Literature review: an acoustic perspective*

Bibliography with respect to sound propagation

- Ignored
- Anechoic Phat
- Fully modeled
- Early echoes

8.1.2 *Literature review: an algorithmic perspective*

[**subsec:application:algos**] Bibliography with respect to learning and knowledge approaches

- Separation vs. Enhancement

- end2end vs. 2step
  end2end: from data to (feature to) target
  2-step: (from data to features) + features to target

- Knowledge-based vs. Learning-based

  - Bottom-up vs Top-down information processing

- Knowledge-based: specialized signal processing and mathematical algorithms informed by knowledge;
- Learning-based: machine learning usually trained in supervised fashion.

▶ SUPERVISED VS. UNSUPERVISED

▶ MACHINE LEARNING VS. DEEP LEARNING

8.2 AUDIO SOURCE SEPARATION

Audio source separation refers to the process of extracting acoustic signals from mixtures featuring target and interfering sounds. The ability of focusing on a selected source only, while filtering out the rest, is known as *the cocktail party effect* [Cherry 1953; Bee and Micheyl 2008] The scenario where no prior knowledge is available about both the sources and the mixing process, is usually referred to as *blind sound separation*. When it is applied to approximate human hearing, such as using only stereophonic mixture and considering the human auditory system, it is known as Computational Auditory Scene Analysis (CASA).

[Remaggi thesis] Two of the mostly investigated areas study either musical or speech recordings [Vincent et al., 2012]. Although music source separation represents a key part of the audio community [Ewert et al., 2014], the focus of this thesis is on speech source separation. This is of interest for several audio applications, such as: speech enhancement [Mohammadiha et al., 2013], crosstalk cancellation [Akeroyd et al., 2007], hearing aids [Healy et al., 2013], and automatic speech recognition [Li et al., 2014]

Echoes have been used previously to enhance various audio processing tasks.

It was shown that they improve indoor beamforming [Dokmanić et al. 2015a; Scheibler et al. 2015; Scheibler 2017], aid in sound source localization [Ribeiro et al. 2010a], and enable low-resource microphone array self-localization [Dokmanić et al. 2016].

8.2.1 *Speech Source Separation*

[Remaggi thesis] During the last twenty years, speech separation has gained quite of attention. Some of the proposed methods achieved source separation exploiting the availability of a single microphone [Jang and Lee, 2003, Radfar and Dansereau, 2007, Schmidt and Olsson, 2006]. However, they were limited by the amount of information utilised. Therefore, other methods attempted the separation process by employing multichannel microphone arrays. These methods are classically categorised into three main groups, depending on the type of approach undertaken [Vincent et al., 2012]: the beamformers [Araki et al., 2003, Coleman et al., 2015a, VanVeen and Buckley, 1988]; the independent component analysis (ICA) based [Bell and Sejnowski, 1995, Cardoso, 1998, Makino et al., 2007]; the time-frequency (TF) mask based [Alinaghi et al., 2014,

Deleforge et al., 2015, Mandel et al., 2010, Sawada et al., 2011, Yilmaz and Rickard, 2004]. A visualisation of these three

8.2.2 *Speech Source Localization*

8.2.3 *Spatial Filtering*

8.2.4 *Room Geometry Estimation*

9

Sound Source Separation & Separake

Source separation, Echoes, Room Geometry, NMF, Multi-channel Processing.

▶ Synopsis In this chapter echoes are used for boosting the performances of state-of-the-art approaches in Audio Source Separation. At first, we describe the existing methods, which typically that either ignore the acoustic propagation, nor attempt to estimate it fully. Instead, this works investigate whether sound separation can benefit of the knowledge of acoustic echoes derived from know the locations of a few virtual microphones. The improvements are show for two standard algorithms based on non-negative matrix factorization: one that uses only magnitudes of the transfer functions, and one that also uses the phases. The experimental part shows that the proposed approach based on a few echoes beats its vanilla variant, and that with magnitude information only, echoes enable separation where it was previously impossible.

Keywords: Blind Channel Identification, Super Resolution, Sparsity, Acoustic Impulse Response.
Resources:

- Paper
- Code
- Slides

The material presented in the chapter results from a collaboration with Robin Scheibler and Ivan Dokmanić and wes previously published in [Scheibler et al. 2018d]. This chapter recall the main findings of the paper bringing additional insight in the literature and in the proposed model, written according to the used notation. The personal contribution to this collaboration, done in the early months of the Ph. D., was implementations the Expectation Maximization (EM)-NMF method accounting the echoes and pre-trained dictionary in Python.

Scheibler et al., "Separake: Source separation with a little help from echoes"

9.1 Literature review in Echo-aware Audio Source Separation

The are many approaches in audio source separation. In this chapter we will consider only the one based on time-frequency masking and we do not consider the one based on spatial filtering. The latters will be reviewed in a dedicated section in **??**.

According to the definition given in Chapter 3, audio source separation algorithms can be grouped according to how they model with sound propagation in the mixing process:

- those that ignore it [Le Roux et al. 2015] (instantaneous mixing process);
- those that assume a single anechoic path [Rickard 2007; Nesta and Omologo 2012] (anechoic mixing process);
- those that model the RTFs entirely [Ozerov and Févotte 2010; Duong et al. 2010; Nugraha et al. 2016; Li et al. 2019] (convolutive mixing process);

- and those that attempt to separately estimate the contribution of the early echoes and the contribution of the late tail [Leglaive et al. 2015].

Currently, in the literature, only few works can be found that incorporate the knowledge of echoes into sound source separation. In the work [Huang et al. 2005], the authors proposed a decomposition of the source separation problem into different procedures. First they estimate the RIRs by extending the SIMO-BCE framework form Multiple Input Multiple Output (MIMO) systems. Here the RIR are modeled as FIR filter following the multipath echo model. Secondly, the estimated filters are used to build the demixing matrix, thus, used to separate the sources with an inverse-filtering approach. However, this method exhibits an high computational cost, which was addressed later in [Rotili et al. 2010]. Nevertheless these approaches were found not robust of low SNR condition.
Alternatively, in [Asaei et al. 2014], proposed an geometry-based approach embedded in a sparse optimization framework: first, by localizing the image sources and estimating the room geometry, the supports of the RIRs' early contributions are estimated; then, after computing the coefficient of the RIRs element in a convex optimization framework, the individual speech signals are separated with either inverse-filtering or sparse recovery. The performance of this approach rely on the RIR and geometry estimation steps, which is very sensitive to the challenging acoustic condition, e. g. low SNR or high RT_{60}.
Instead, the work [Leglaive et al. 2015] proposes to tackle the convolutive model by imposing a probabilistic prior on the early part of the RIRs, namely, modeled as an autoregressive process in the frequency domain. Later, the same authors extended this work in [Leglaive et al. 2016] accounting for both early and late part of the mixing filters.

▶ The propose of this work is yet different than that presented above. First, rather than fitting the echo model as in [Leglaive et al. 2015; Leglaive et al. 2016], nor estimating the mixing filters as in [Huang et al. 2005; Asaei et al. 2014], we aim to show that separation in the presence of known echoes is better than separation without echoes .
Second, we conduct this investigation in the context of source separation with non-negative source models.
Third, we propose to solve the problem from the point of view of *image microphones*. The image microphone model is equivalent of the Image Source Method (ISM) [Allen and Berkley 1979], where virtual receiver are placed outside of the room (See Figure 9.1). Even if the ISM is more common and implemented in practice in acoustic simulators, the two models are strictly equivalent. Therefore, the assumption behind these models are easy to satisfy in living rooms and conference rooms, but the corresponding model incurs a significant mismatch with respect to the complete reverberation (See Chapter 2). This approach is based on the acoustic rake receivers previously proposed in [Dokmanić et al. 2015a] and thus dubbed as Sound Separation by Raking Echoes (SEPARAKE).

The considered setup is illustrated in Figure 9.1. We assume that the array is placed close to a wall or a corner. This is useful for the following reasons:

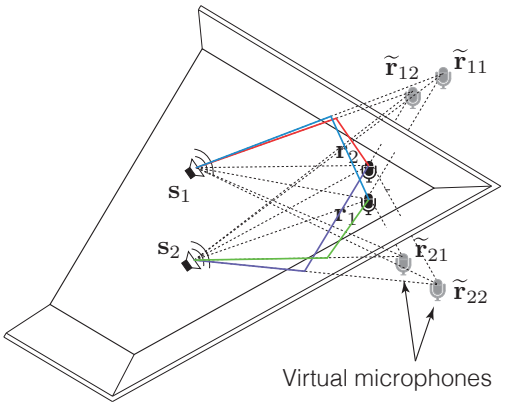


FIGURE 9.1: Typical setup with two speakers recorded by two microphones. The illustration shows the virtual microphone model (grey microphones) with direct sound path (dotted lines) and resulting first-order echoes (colored lines).

first, it makes echoes from the nearby walls significantly stronger than all other echoes; second, it ensures that the resulting image array (real and image microphones) is compact, allowing to assume the far field regime.

▶ TRASLATING ECHOES INTO IMAGE ARRAYS provides an interesting geometrical interpretation in light of beamforming theory. Real and virtual microphones form dipoles with diverse frequency-dependent directivity patterns. By integrating more and more virtual microphones, the directivity patterns change and higher spatial selectivity can be achieved [Dokmanić et al. 2015b]. This effect is shown in § 9.1. Therefore, the goal of this work is to design audio source separation algorithms which benefit from this known spatial diversity.

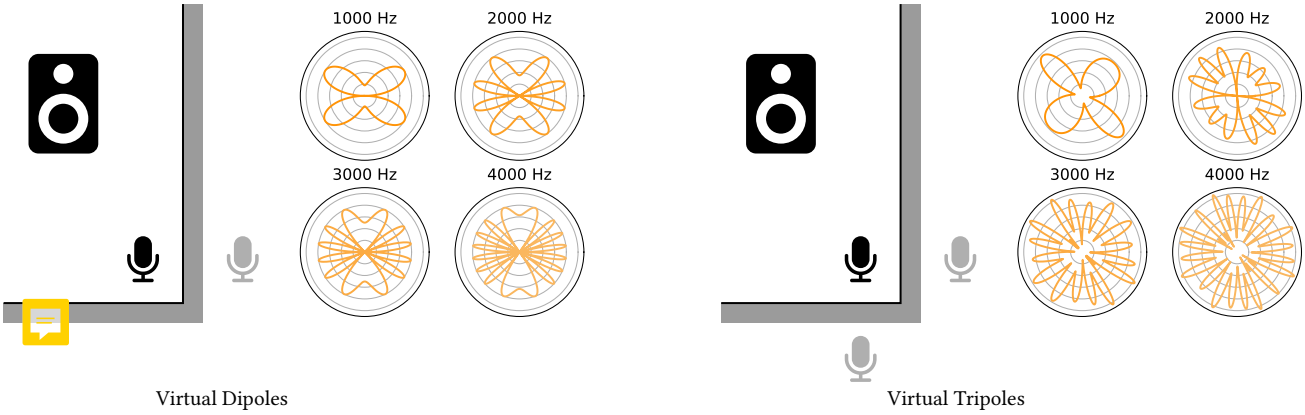


FIGURE 9.2: Frequency-dependent directivity pattern

9.2 Modeling

Recalling the echo model for the RIRs, and assuming R echoes per source are known, the approximate Room Transfer Function (RTF) from source j to microphone i writes

$$\tilde{H}_{ij}(f) = \sum_{r=0}^{R} \alpha_{ij}^{(r)} \mathrm{e}^{-\mathrm{i}2\pi f \tau_{ij}^{(r)}}. \tag{9.1}$$

The far field assumption implies that only the relative arrival times are known, so we can arbitrarily fix the delay of the direct path to zero. In addition, we assume all walls to be spectrally flat in the frequency range of interest and that $\alpha_{ij}^{(r)}$ are known up to a scaling (i.e. $\alpha_{ij}^{(0)} = 1$). In this work the echoes properties are assumed to be known.

Assuming the narrowband approximation, the mixing process can be modeled ~~model~~ as in § 3.2.5. Therefore, the Short Time Fourier Transform (STFT) of the i-th microphone signal reads

$$X_i[k,l] = \sum_{j=1}^{J} H_{ij}[k] S_j[k,l] + N_i[k,l] \tag{9.2}$$

with $k \in [0, \dots, F]$ and $l \in [0, \dots, T]$ being the frequency and frame index, $H_{ij}[k]$ is the DFT approximating the RTF of (9.1), $X_j[k,l]$ the STFT of the j-th source signal, and $N_i[k,l]$ a term including noise and model mismatch. It is convenient to group the microphone observations in vector-matrix form,

$$\boldsymbol{X}[k,l] = \mathbf{H}[k]\boldsymbol{S}[k,l] + \boldsymbol{N}[k,l], \tag{9.3}$$

where $\boldsymbol{X}[k,l], \boldsymbol{N}[k,l] \in \mathbb{C}^{I\times 1}$, $\boldsymbol{S}[k,l] \in \mathbb{C}^{J\times 1}$ and $\mathbf{H}[k,l] \in \mathbb{C}^{I\times J}$.

Let the squared magnitude of the spectrogram of the j-th source be $\mathbf{P}_j = \left[|S_j|^2\right]_{kl} \in \mathbb{R}^{F\times T}$. As depicted in Figure 9.3, the spectrogram can be modeled as the product of 2 non-negative matrices:

$$\mathbf{P}_j = \mathbf{D}_j \mathbf{Z}_j, \tag{9.4}$$

where $\mathbf{D}_j$ is the non-negative *dictionary* whose contains the spectral can be interpreted as spectral templates of the source, and the latent variables $\mathbf{Z}_j$, called *activations*, indicates when and how this templates are activated.

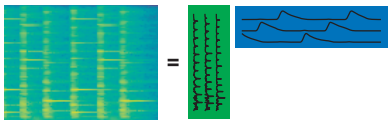


Figure 9.3: Spectrogram of a sound source signal decomposed into dictionary and activation

▶ ~~The~~ NMF-based Audio Source Separation can then be cast as an inference problem in which we maximize the likelihood of the observed $\boldsymbol{X}$ over all possible non-negative factorizations (9.4). This normally involves learning the channels, namely the frequency-domain mixing matrices $\mathbf{H}$. Instead of learning them, we build the channels based on the prior knowledge of the earliest few echoes.

9.3 Source Separation by NMF

In this work we consider ~~the~~ two standard, well-understood multi-channel source separation algorithms which, by default, estimate the channels together

with sources' dictionaries and activations. The first algorithm is ~~the~~ Nonnegative Matrix Factorization (NMF) via Multiplicative Updates (MU) and consider only the magnitudes of the transfer functions. The second one is ~~the~~ multichannel NMF via Expectation Maximization (EM), which instead explicitly model the phases of the mixing filters. In this work, we considered only the (over)determined case ($J \leq I$). In the following we briefly describe the idea behind the two algorithms. We reminds to the work of [Ozerov and Févotte 2010] for further details.

9.3.1 *NMF using Multiplicative Updates (MU-NMF)*

MU for NMF only involve the magnitudes only and the updates rules are guaranteed non-negative as long as the initialization is. This model have been originally proposed by in [Lee and Seung 2001], however we will consider its formulation as it appear in [Ozerov and Févotte 2010]. The observed multi-channel squared magnitude spectra $\mathbf{V}_i = \left[|X_i[k,l]|^2\right]_{kl}$ and their nonnegative factorizations,

$$\widehat{\mathbf{V}}_i = \sum_{j=1}^{J} \operatorname{diag}(\mathbf{Q}_{ij})\mathbf{D}_j\mathbf{Z}_j, \quad i = 1, \ldots, I \tag{9.5}$$

where $\mathbf{Q}_{ij} = \left[|H_{ij}[k]|^2\right]_k$ is the vector of squared magnitudes of the approximate RTF between microphone i and source j.

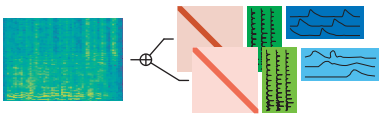


FIGURE 9.4: Schematics of the signal model used for MU-NMF.

- THE MU COST FUNCTION ~~is~~ minimize the *Itakura-Saito divergence* [Févotte and Idier 2011] between the observed spectrogram ($\mathbf{V}_i[k,l]$ and the model $\widehat{\mathbf{V}}_i[k,l]$, that is,

$$\mathcal{C}_{\text{MU}}(\Theta_{\text{MU}}) = \sum_{jkl} \mathcal{D}_{\text{IS}}(\mathbf{V}_i[k,l]|\widehat{\mathbf{V}}_i[k,l]) + \gamma \sum_j \|\mathbf{Z}_j\|_1, \tag{9.6}$$

where $\mathcal{D}_{\text{IS}}(v|\hat{v}) = \frac{v}{\hat{v}} - \log\frac{v}{\hat{v}} - 1$ and $\Theta_{\text{MU}} = \{\mathbf{Q}_{ij}, \{\mathbf{D}_j, \mathbf{Z}_j\}_j\}_{ij}$ is the set of parameters. We add an ℓ_1-penalty term to promote sparsity in the activations due to the potentially large size of the dictionary [Sun and Mysore 2013].

- THE MU ~~UPDATING~~ RULES for each scalar parameter of interest θ are obtained by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w. r. t. this parameter, namely,

$$\theta \leftarrow \theta \frac{[\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_-}{[\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_+}$$

where $\mathcal{C}_{\text{MU}}(\Theta_{\text{MU}}) = [\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_+ - [\nabla_\theta \mathcal{C}_{\text{MU}}(\Theta_{\text{MU}})]_-$ and the summands are both nonnegative. By adapting the original MU rule derivations from Ozerov and Févotte, we obtain:

$$\mathbf{Q}_{ij} \leftarrow \mathbf{Q}_{ij} \odot \frac{\left[\widehat{\mathbf{V}}_j^{-2} \odot \mathbf{V}_j \odot (\mathbf{Z}_j\mathbf{D}_j)\right]\mathbf{1}_{1\times T}}{\left[\widehat{\mathbf{V}}_j^{-1} \odot (\mathbf{Z}_j\mathbf{D}_j)\right]\mathbf{1}_{1\times T}} \tag{9.7}$$

$$\mathbf{Z}_j \leftarrow \mathbf{Z}_j \odot \frac{\sum_i (\operatorname{diag}(\mathbf{Q}_{ij})\mathbf{D}_j)^\intercal \left(\mathbf{V}_j \odot \widehat{\mathbf{V}}_j^{-2}\right)}{\sum_i (\operatorname{diag}(\mathbf{Q}_{ij})\mathbf{D}_j)^\intercal \widehat{\mathbf{V}}_j^{-1} + \gamma}, \tag{9.8}$$

$$\mathbf{D}_j \leftarrow \mathbf{D}_j \odot \frac{\sum_i \operatorname{diag}(\mathbf{Q}_{ij})^\intercal \left(\mathbf{V}_j \odot \widehat{\mathbf{V}}_j^{-2}\right)\mathbf{Z}_j^\intercal}{\sum_i \operatorname{diag}(\mathbf{Q}_{ij})^\intercal \widehat{\mathbf{V}}_j^{-1}\mathbf{Z}_j^\intercal}, \tag{9.9}$$

where multiplication $\odot$, power, and division are element-wise and $\mathbf{1}_{1\times T}$ is a N-vector of ones,.

9.3.2 *NMF using Expectation Maximization (EM-NMF)*

Unlike the MU algorithm that independently maximizes the log-likelihood of spectral magnitudes, the EM-NMF maximizes the joint log-likelihood over all complex-valued channels [Ozerov and Févotte 2010]. Hence, the model takes explicitly into account observed phases. In this approach, each source j is modeled as complex Gaussian in the form of

$$S_j[k,l] \sim \mathcal{N}_c(0, (\mathbf{D}_j\mathbf{Z}_j)_{kl}), \tag{9.10}$$

and the magnitude spectrum $\mathbf{P}_j$ of (9.4) can be understood as the variance of source j.
Under this model, and assuming uncorrelated noise, the microphone signals also follow a complex Gaussian distribution with covariance matrix

$$\mathbf{\Sigma_X}[k,l] = \mathbf{H}[k]\mathbf{\Sigma_S}[k,l]\mathbf{H}^{\mathsf{H}}[k] + \mathbf{\Sigma_N}[k,l], \tag{9.11}$$

where $\mathbf{\Sigma_S}$ and $\mathbf{\Sigma_N}$ are the covariance matrix of the sources and noise, respectively.

- THE EM COST FUNCTION correspond to the negative log-likelihood of the observed signal, that is,

$$\mathcal{C}_{\text{EM}}(\Theta_{\text{EM}}) = \sum_{kl} \operatorname{trace}\Big(\boldsymbol{X}[k,l]\boldsymbol{X}[k,l]^{\mathsf{H}}\mathbf{\Sigma}_{\boldsymbol{X}}^{-1}[k,l]\Big) + \log\det\mathbf{\Sigma_X}[k,l]. \tag{9.12}$$

where the $\Theta_{\text{EM}} = \Big\{\mathbf{H}, \{\mathbf{D}_j, \mathbf{Z}_j\}_j, \mathbf{\Sigma_N}\Big\}$ is the set of parameters.

- THE EM ALGORITHM estimates all the parameters Θ by alternating between the so-called E-step and M-step. In a nutshell, one iteration the E-step consists of computing the *conditional expectation* of the the log likelihood of $\mathcal{C}_{\text{EM}}$ with respect to the current parameter estimates, and the M-step re-estimating the parameters by maximizing the conditional expectation of the log-likelihood of $\mathcal{C}_{\text{EM}}$. This quantity can be efficiently minimized using the EM algorithm proposed in [Ozerov and Févotte 2010]. Moreover, since adding sparsity priors is not straightforward in the EM framework, it was left for future work.

9.4 ECHO-AWARE SOURCE SEPARATION

To evaluate the usefulness of echoes in source separation, we modified ~~the~~ the multi-channel NMF framework of Ozerov and Févotte [Ozerov and Févotte 2010]. The knowledge of the echoes in embedded in ~~the embedded in~~ the model by approximating the entries of mixing matrix with (??), that is,

$$\begin{aligned} H_{ij}[k] &= \sum_{r=0}^{R} \alpha_{ij}^{(r)} \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_{ij}^{(r)}}, \\ \mathbf{H}[k] &= [H_{ij}[k]]_{ij}, \end{aligned} \tag{9.13}$$

where $f_k = kF_s/F$ are the discretized frequencies in Hz corresponding to the k-th bin in the DFT.

Futhermore, the early-echo channel model is kept fixed throughout the iterations. Moreover, instead of updating both sources' dictionaries and activations, we adapted pre-trained dictionaries to better guide the source separation.

- PRE-TRAINED DICTIONARIES are a typical way to inform the NMF algorithm, and sometimes referred to as *supervised NMF*. The idea is run NMF-based source separation on a set of training and collect the atoms of the estimated non-negative matrices[Schmidt and Olsson 2006]. At ~~the~~ test phase, this atoms are used as basis vectors for the dictionary matrix (i. e., $\mathbf{D}$) and can be used as a good initialization point or kept fixed in the algorithm[43]. This can be seen as an instance of the problem of *dictionary learning* which exist also in many other research field. For audio source separation, this ideas has been studied extensively since promising results were obtained, even in single channel scenarios [Smaragdis et al. 2009]. Ad discussed later in § 9.5.2, in this work we will use two different dictionaries: one *universal*, and the other *speaker-specific*.

[43] In the context of NMF-based music transcription applied to piano music, the dictionary can be the collection of spectral templates, each of which is associated to a piano notes [Müller 2015]

- NEGLECTING THE REVERBERATION (or working in the anechoic regime) leads to a constant $\mathbf{Q}_{ij}$ for all j and i. A consequence is that the MU-NMF framework breaks down with a universal dictionary, namely, $\mathbf{D} = \mathbf{D}_j \ \forall j$. Indeed, (9.5) becomes the same for all i,

$$\widehat{\mathbf{V}}_i = \sum_j \mathbf{D}\mathbf{Z}_j = \mathbf{D} \sum_j \mathbf{Z}_j,$$

so even with the correct atoms identified, we can assign them to any source without changing the value of the cost function. Therefore, anechoic multichannel separation with a universal dictionary cannot work well. This intuitive reasoning is corroborated by numerical experiments in Section 9.5.4. The problem is overcome by the EM-NMF algorithm which keeps the channel phase and is thus able to exploit the phase diversity across the array. Of course, as showed in this work, it is also overcome by using echoes.

9.5 Numerical Experiments

We test our hypotheses through computer simulations. In the following, we describe the simulation setup, dictionary learning protocols, and we discuss the results.

9.5.1 *Setup*

An array of three microphones arranged on the corners of an equilateral triangle with edge length $0.3\,\mathrm{m}$ is placed in the corner of a 3D room with 7 walls. We select 40 sources at random locations at a distance ranging from $2.5\,\mathrm{m}$ to $4\,\mathrm{m}$ from the microphone array. Pairs of sources are chosen so that they are at least $1\,\mathrm{m}$ apart. The floor plan and the locations of microphones are depicted in Figure 9.5. The scenario is repeated for every two active sources out of the 780 possible pairs.

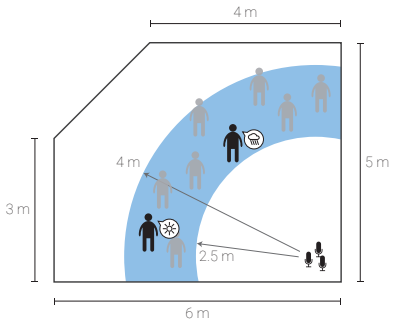


FIGURE 9.5: The simulated scenario.

The sound propagation between sources and microphones is simulated using the image source model implemented in `pyroomacoustics` Python package [Scheibler et al. 2018a]. The wall absorption factor is set to 0.4, leading

			Number of echoes R						
	anechoic	learn	0	1	2	3	4	5	6
$\gamma =$	10	10^{-1}	10	10^{-3}	0	0	0	0	0

TABLE 9.1: Value of the regularization parameter γ used with the universal dictionary.

to a RT_{60} of approximately 100 ms. The sampling frequency is set to 16 kHz, STFT frame size to 2048 samples with 50% overlap between frames, and we use a cosine window for analysis and synthesis. Partial RTFs are then built from the R nearest image microphones. The global delay is discarded, and only the relative amplitudes between echoes are kept.

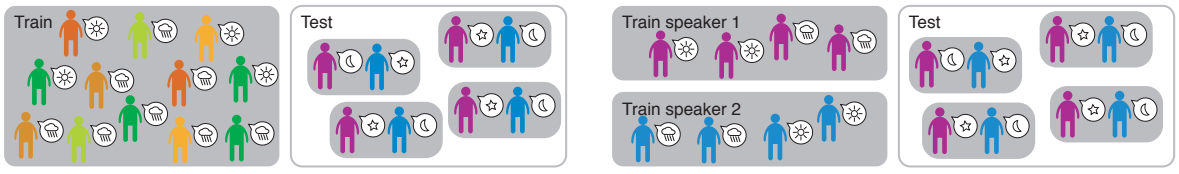
With this setup, we perform three different experiments. In the first one, we evaluate MU-NMF with a universal dictionary. In the other two, we evaluate the performance of MU-NMF and EM-NMF with speaker-specific dictionaries. We vary R from 1 to 6 and use three baseline scenarios:

1. *anechoic*: Anechoic conditions, no model mismatch.
2. *learn*: The RTFs are learned from the data along the activations as originally proposed [Ozerov and Févotte 2010].
3. *no echoes*: Reverberation is present but ignored (i.e. $R = 0$).

With the universal dictionary, the large number of latent variables warrants the introduction of sparsity-inducing regularization. The value of the regularization parameter γ was chosen by a grid search on a holdout set with the signal-to-distortion ratio (SDR) as the figure of merit [Vincent et al. 2007] (See Table 9.1).

9.5.2 *Dictionary Training, Test Set*

First, we introduce a dictionary learned from available training data. We explore both speaker-specific and universal dictionaries [Sun and Mysore 2013]. Speaker-specific dictionaries can be beneficial when speakers are known in advance. Universal dictionary is more versatile but gives a weaker regularization prior.



(a) Universal dictionary

(b) Speaker-specific dictionary

- UNIVERSAL DICTIONARY: Following the methodology of [Sun and Mysore 2013] we select 25 male and 25 female speakers and use all available training sentences to form the universal dictionary $\mathbf{D} = [\mathbf{D}^{\mathtt{M}}_1 \cdots \mathbf{D}^{\mathtt{M}}_{25}\, \mathbf{D}^{\mathtt{F}}_1 \cdots \mathbf{D}^{\mathtt{F}}_{25}]$. The test signals were selected from speakers *and* utterances outside the training set. The number of latent variables per speaker is 10 so that with STFT frame size of 2048 we have $\mathbf{D} \in \mathbb{R}^{1025\times 500}$.

- SPEAKER-SPECIFIC DICTIONARY: Two dictionaries were trained on one male and one female speaker. One utterance per speaker was excluded to be used for testing. The number of latent variables per speaker was set to 20.

All dictionaries were trained on samples from the TIMIT corpus [Garofolo et al. 1993] using the NMF solver in `scikit-learn` Python package [Pedregosa et al. 2011].

9.5.3 *Implementation:*

Authors of [Ozerov and Févotte 2010] provide a Matlab implementation[44] of MU-NMF and EM-NMF methods for stereo separation. We ported their code to Python and extended it to arbitrary number of input channels[45]. However this software features some ad-hoc decisions which do not fit our scenario. Thus, we provide a Python3 adaptation with the following modifications.

- First the original code was restricted to the 2-channel case, i.e. $I = 2$. Thus, in order to embrace the specifics of our scenario and for sake of generalization, we extend it to the multi-channel case, that is $\forall I \geq 1$.
- the MU-NMF was modified to handle sparsity constraint as described in 9.3.1.
- since EM method degenerates where zero-valued entries are present in the dictionary matrix, $\mathbf{D}$, all these entries are initially set to a small constant value of 10^{-6}.
- the code was further modified to deal with fixed dictionary and channel models matrices, which are normalized in order to avoid indeterminacy issues [Ozerov and Févotte 2010].

Now to conclude with, no *simulated annealing* strategies are not used in the final experiments. In fact in some preliminary and informal investigations we noticed that this yields to better results then using annealing. In the experiments, the number of iterations for MU-NMF (EM-NMF) was set to 200 (300).

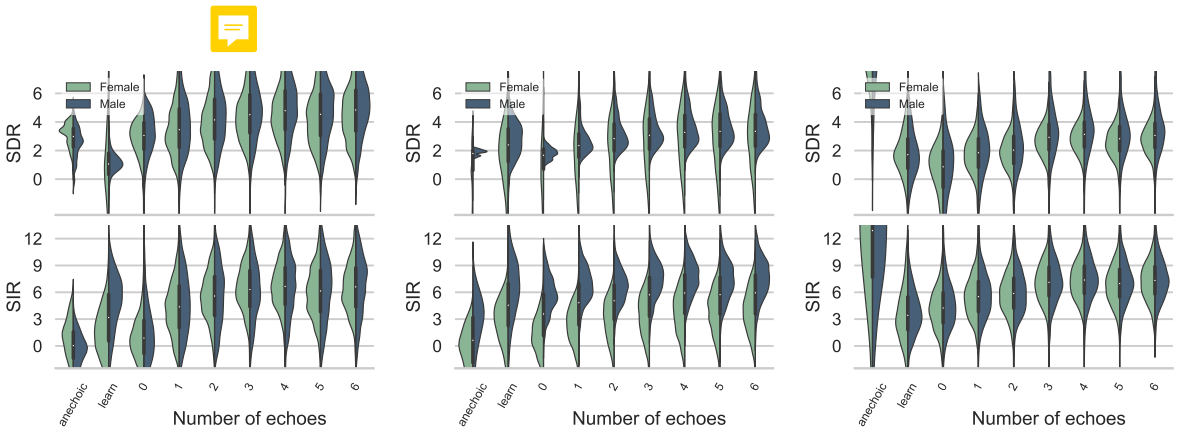
[44] Multichannel nonnegative matrix factorization toolbox (in Matlab)

[45] Our implementation and all experimental code are publicly available in line with the philosophy of reproducible research.

9.5.4 *Results*

We evaluate the performance in terms of signal-to-distortion ratio (SDR) and source-to-interference ratio (SIR) as defined in [Vincent et al. 2007]. We compute these metrics using the `mir_eval` toolbox [Raffel et al. 2014].

The distributions of SDR and SIR for separation using MU-NMF and a universal dictionary are shown in Figure 9.6a, with a summary in Figure 9.7. We use the median performance to compare the results from different algorithms. First, we confirm that separation fails for flat RTFs (*anechoic* and $R = 0$) with SIR at around 0 dB. Learning the RTFs performs somewhat better in terms of SIR than in terms of SDR, though both are low. Introducing approximate RTFs dramatically improves performance: the proposed approach outperforms the learned approach even with a single echo. With up to six echoes, gains



(a) MU-NMF, Universal dictionary (b) MU-NMF, Speaker-specific dictionary (c) EM-NMF, Speaker-specific dictionary

FIGURE 9.6: Distribution of SDR and SIR for male and female speakers as a function of the number of echoes included in modeling, and comparison with the three baselines.

are +2 dB SDR and +5 dB SIR. Interestingly, with more than one echo, non-negativity and echo priors already sufficient for achieving good separation, overlooking the ℓ_1 regularization.

Separation with speaker-dependent dictionaries is less challenging since we have a stronger prior. Accordingly, as shown in Figures 9.6b and 9.7, MU-NMF now achieves a certain degree of separation even without the channel information. The gains from using echoes are smaller, though one echo is still sufficient to match the median performance of learned RTFs. Using an echo, however, results in a smaller variance, while adding more echoes further improves SDR (SIR) by up to +2 dB (+3 dB).

In the same scenario, EM-NMF (Figure 9.6c) has near-perfect performance on anechoic signals which is expected as the problem is overdetermined. For MU, a single echo suffices to reach the performance of learned RTFs and further improve it. Moreover, echoes significantly improve separation quality as illustrated by up to 3 dB improvement over *learn*. It is interesting to note that in all experiments the first three echoes near-saturate the metrics. This is good news since higher order echoes are hard to estimate.

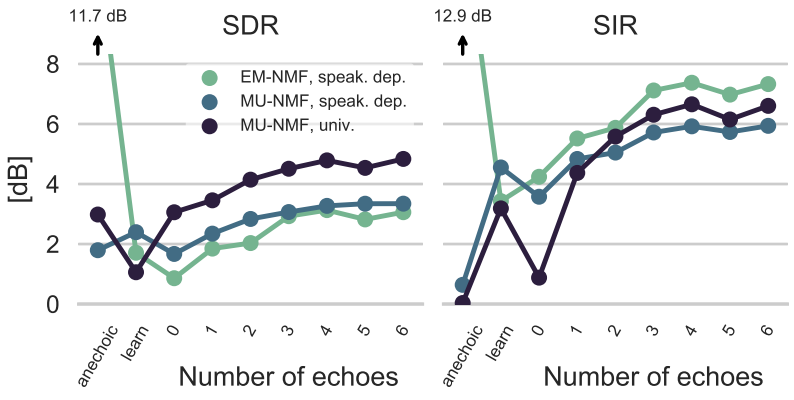


FIGURE 9.7: Summary of the median SDR and SIR for the different algorithms evaluated.

9.6 Conclusion

In this work, we investigated the role of early echoes for the problem of sound source separation. Unlike earlier work, instead of fitting echo model or trying to estimate the blindly the acoustic channels, we investigate the potential of including the properties of known echoes in well established NMF-based source separation algorithms. In particulars, we modified the MU approach — which consider only spectral magnitudes— and the EM — which accounts for complex spectra — by integrating a simple echo model. Despite its simplicity, such echo model lend itself to a interesting interpretation by revising the ISM model: to each echo corresponds an image microphones (instead of image source as in ISM). It follows that real and image microphone can be considered as microphones arrays with specific directivity pattern.
Numerical results shows that echoes seem to play an essential role in magnitude-only algorithms, like the MU-NMF. In general, the showed that using knowledge of a few echoes significantly improve results with respect to an anechoic model. This improvement is measured by the standard metrics even when compared to approaches that learn the transfer functions. To conclude with, does echoes helps sound source separation? The answer is *yes*.

▶ Future work on echo-aware source separation includes:

- integrating the blind estimation of the echoes properties, e. g. using the proposed algorithm `Blaster`.
- including the late reverberation part in the mixing matrices;
- experiments with more microphone, room configurations, more source on real data, e. g. using the one offered by the `dEchorate` dataset.

10

Sound Source Localization & Mirage

Sound Source Localization, Image Microphones, TDOA Estimation, Supervised Learning.

▸ Synopsis It is commonly observed that acoustic echoes hurt performance of sound source localization (SSL) methods. We introduce the concept of microphone array augmentation with echoes (MIRAGE) and show how estimation of early-echo characteristics can in fact benefit SSL. We propose a learning-based scheme for echo estimation combined with a physics-based scheme for echo aggregation. In a simple scenario involving 2 microphones close to a reflective surface and one source, we show using simulated data that the proposed approach performs similarly to a correlation-based method in azimuth estimation while retrieving elevation as well from 2 microphones only, an impossible task in anechoic settings.

10.1 Introduction

Sound source localization (SSL) consists in determining the position of a sound source from microphone signals in 3D space. In polar coordinates, most existing methods focus on estimating the directional of arrival, namely, azimuth and elevation angles. Though this task is performed routinely by humans, it still challenges today's computational methods, in particular in the presence of reverberation or interfering sources (see [**rascon2017localization**] and [**Argentieri2015**] for a review). Computational approaches consist in two components. First, extracting features from audio data that are as independent as possible from the source's content while preserving spatial information. Second, mapping these features to the source position. Two lines of research have been investigated to obtain such mappings: physics-based and learning-based approaches.

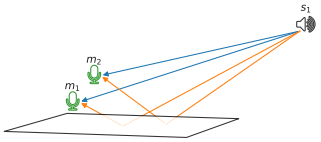


Figure 10.1: Typical setup with one source source recorded by two microphones. The illustration shows direct sound path (blue lines) and resulting first-order echoes (orange lines).

▸ Physics-based approaches rely on a simplified sound propagation model [**rascon2017localization**; **Knapp1976**; **DiBiase2001**; **Lebarbenchon2018**]. The free-field model is by far the most widely used one and assumes a single direct sound path from the source to each microphone. When the source is placed far enough, this yields a closed-form mapping from the sound's time-difference-of-arrival (TDOA) in a microphone pair and the source's azimuth angle in this pair. If multiple microphone pairs are available and form a non-linear array, their TDOAs can be aggregated to obtain 2D directions of arrival [**DiBiase2001**]. These methods strongly suffer in environments where the

free-field assumption is violated, *e.g.*, in the presence of strong acoustic echoes and reverberation [**Scheuing2006**].

- LEARNING-BASED APPROACHES use an annotated training dataset to implicitly learn a mapping from audio features to source positions [**deleforge2015acoustic**; **Vesperini2016**; **Adavanne2017**; **Perotin2018**; Gaultier et al. 2017]. Such data can be obtained from real recordings [**deleforge2015acoustic**] or using physics-based simulators [**Vesperini2016**; **Adavanne2017**; **Perotin2018**; Gaultier et al. 2017]. These methods were showed to overcome some limitations of the free-field model, but are usually trained for specific microphone arrays and fail whenever test conditions strongly mismatch training conditions.

  Gaultier et al., "VAST: The virtual acoustic space traveler dataset"

  Most sound source localization methods, including the above listed, regard reverberation and in particular acoustic echoes as a nuisance. In contrast, some recent work that we refer to as *echo-aware* methods have showed that the knowledge of early acoustic echoes could be used to reconstruct the geometry of an audio scene [**Nakashima2010**; **An2018**; Dokmanić et al. 2013] or to improve performance of signal enhancement methods [**flanagan1993spatially**; **Scheibler2017**; Dokmanić et al. 2015b]. In [**Nakashima2010**], some ad-hoc reflectors are used as artificial *pinnae* to estimate elevation based on a simple reflection model. In [**An2018**], cameras, depth sensors and laser sensors are used to identify reflectors and build a corresponding acoustic model that helps SSL.

  Dokmanić et al., "Acoustic echoes reveal room shape"

  Dokmanić et al., "Raking the cocktail party"

- IN THIS WORK, we combine ideas from physics-based, learning-based and echo-aware approaches to introduce the framework of microphone array augmentation with echoes (MIRAGE) for SSL. We consider a simple yet common scenario to illustrate this idea: two microphones, one source and a nearby reflective surface, as illustrated in Fig. **??**. This may occur, for instance, when the sensors are placed on a table such as in voice-based assistant devices or next to a wall. The reflective surface is assumed to be the most reflective and closest one to the microphones in the environment, hence generating the strongest and earliest echo in each microphone. Under this *close-surface* model, we ask the following questions:

  1. Can early echoes be estimated from two-microphone recordings of an unknown source?
  2. Can they be used to estimate both the azimuth and elevation angles of the source, an *impossible* task in free field conditions?

  We propose to use a deep neural network (DNN) trained on a simulated close-surface dataset to estimate early echoes properties from audio features. The MIRAGE framework then exploits these estimated properties by expressing them as TDOAs in the *virtual 4-microphone array* formed by the true microphone pair and its image with respect to the reflective surface. We show that the proposed framework approximately estimates echo properties, perform similarly to a correlation-based method in azimuth estimation for the considered scenario and estimates *impossible* elevation angles with good accuracy in noiseless settings using two microphones only.

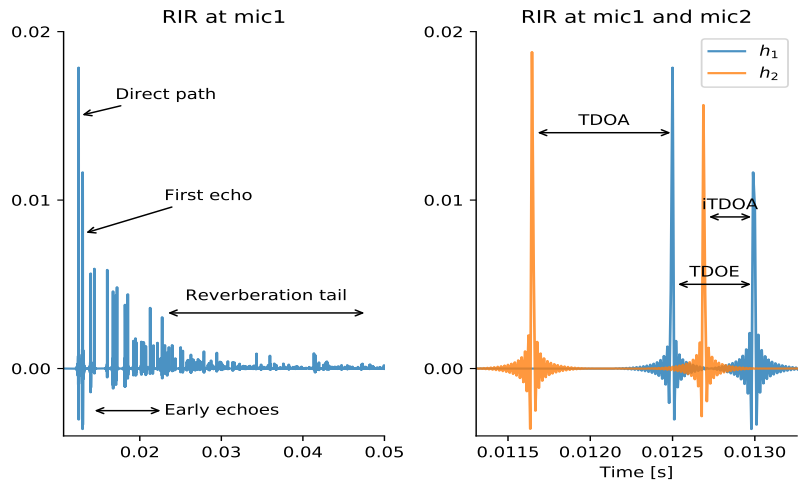


FIGURE 10.2: Left, a typical simulated RIR with annotated components. Right, superposition of two RIRs and visualization of time difference of arrival between direct paths (TDOA), first echoes (iTDOA) and direct path and first echo (TDOE).

10.2 BACKGROUND IN MICROPHONE ARRAY SSL

In this section, we briefly review some necessary background in microphone array SSL. Let us assume a microphone array of I sensors is placed inside a room and records the sound emitted by one static point sound source. In all generality, the relationship between the signal $m_i(t)$ recorded by the sensor placed at fixed position $\mathbf{m}_i$ and the signal $s(t)$ emitted by the source at fixed position $\mathbf{s}$ is defined by:

$$m_i(t) = (h_i * s)(t) \; + \; n_i(t), \tag{10.1}$$

where the convolution with room impulse response (RIR) $h_i(t)$ embodies the fact that sensor i receives a so-called spatial image of the source and n_i denotes possible measurement noise. The RIR depends on the spatial parameters of the scene: microphone positions, source position w.r.t the room, as well as the room acoustic properties (size, absorption and diffuseness of the wall materials.)

RIRs can be typically modelled as the sum of the direct path and multiple reflections of the sound. This can boil down to modelling h_i as a Dirac impulse at time τ_i accounting for the time delay from the source to microphone i, plus an error term. In the frequency domain, this leads to:

$$H_i(f) = \alpha_i(f)\, e^{-2\pi f \tau_i} \; + \; \varepsilon_i(f), \tag{10.2}$$

where the error term $\varepsilon_i(t)$ collects echoes, the reverberation tail, diffusion, and noise. The term $\alpha_i(f)$ captures the air attenuation phenomenon. A time-domain example of RIR is shown in Fig. **??** (left).

10.2.1 *2-channel 1D-SSL*

Let us first consider the stereo case ($I = 2$). Under the far-field assumption, traditional SSL methods use the time difference of arrival (TDOA), $\tau \triangleq \tau_2 - \tau_1$, as a proxy for the estimation of the angle of arrival (AOA), since:

$$\text{AOA} = \arccos\left(c\,\tau \;/\; d\right), \tag{10.3}$$

where c is the speed of sound and d the inter-microphone distance. SSL then reduces to estimating the TDOA, which can be done by cross-correlation-based methods such as the widely used and well performing generalized cross-correlation with phase transform (GCC-PHAT) method [**Knapp1976**; **Blandin2012**]. Given short-time Fourier transforms M_1 and M_2 of the two microphones signals, the GCC-PHAT *angular spectrum* is defined as:

$$\Psi_{\text{GCC}}(\tau) = \sum_{f,n} \frac{M_1(f,n)M_2^*(f,n)}{\mid M_1(f,n)M_2^*(f,n) \mid} e^{-2\pi f \tau}. \tag{10.4}$$

Then, the TDOA estimate is given by $\hat{\tau} = \arg\max_{\tau} \Psi_{\text{GCC}}(\tau)$. Note that Ψ_{GCC} can also be expressed directly as a function of the AOA using (10.3), hence the term *angular spectrum*. This method was showed to be state-of-the-art in a large benchmark [**Blandin2012**].

10.2.2 *Multichannel 2D-SSL*

When more microphones are available and the array is not linear, 2D-SSL can be envisioned. A possible approach is to use 1D-SSL on all pairs and combine their results, a principle which was successfully applied in the steered response power with phase transform (SRP-PHAT) method [**DiBiase2001**]. SRP-PHAT exploits the geometry of the microphone array and the estimated TDOAs from microphone pairs to return the DOA. In a nutshell, this algorithm aims to estimate a global angular spectrum $\Psi_{\text{SRP}}(\theta, \phi)$ which will exhibit a local maximum in the direction of the active source. First, a global grid of possible DOAs is defined according to a desired resolution and computational load. Second, for each pair of microphones, a local set of AOAs is defined and a TDOA-based algorithm (e.g. GCC-PHAT) is used to compute the associated local angular spectrum. Finally all the local contributions (a collection of local $\Psi_{\text{GCC}}(\tau)$) are geometrical aggregated and interpolated back to the global DOA grid to form $\Psi_{\text{SRP}}(\theta, \phi)$, and the DOA maximizing Ψ is used as estimate.

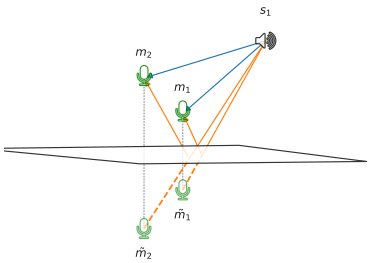


FIGURE 10.3: Illustration of the images $\tilde{m}_1$ and $\tilde{m}_2$ of microphones m_1 and m_2 in the presence of a reflective surface and a source. Blue lines correspond to direct paths, orange lines correspond to echo paths.

10.3 MIRAGE: MICROPHONE ARRAY AUGMENTATION WITH ECHOES

We now introduce the proposed concept of microphone array augmentation with echoes (MIRAGE). Let us first expand formula (10.2) to account for more echoes:

$$H_i(f) = \sum_{k=0}^{K} \alpha_i^k(f)\, e^{-2\pi f \tau_i^k} \; + \; \varepsilon_i(f) \tag{10.5}$$

where the sum now comprises the direct path ($k = 0$) and the K earliest reflections ($K = 1$ in this paper) and ε_i collects the remaining RIR components. Here, $\alpha_i^k(f)$ accounts for both air attenuation and wall absorption phenomena. In the remainder of this paper, we make the approximation of frequency-independent α_i^k. Eq. Eq. (10.5) then corresponds to the well known image-source (IS) model, where reflections are treated as mirror images of the true source with respect to reflective surfaces, emitting the same signal. We will employ here a less common but equivalent interpretation of IS, namely, the image-microphone (IM) model. As illustrated in Fig. Figure 10.3, virtual microphones are mirror images of the true microphones with respect to reflective surfaces. In this view, the echoic signal received at a true microphone is

the sum of the anechoic signals received at this microphone and its images. If we consider the virtual array consisting of both true and image microphones, multiple microphone pairs are now available. For each of them, it is then possible to define a corresponding time difference of arrival. Among them, we will refer to the one between the two real microphones as TDOA, the one between the two image microphones as image TDOA (iTDOA) and the one between the first microphone and its image as time difference of echoes (TDOE). We have:

$$\text{TDOA} = (\|\mathbf{m}_2 - \mathbf{s}\| - \|\mathbf{m}_1 - \mathbf{s}\|)/c = \tau_2^0 - \tau_1^0, \tag{10.6}$$

$$\text{iTDOA} = (\|\tilde{\mathbf{m}}_2 - \mathbf{s}\| - \|\tilde{\mathbf{m}}_1 - \mathbf{s}\|)/c = \tau_2^1 - \tau_1^1, \tag{10.7}$$

$$\text{TDOE} = (\|\tilde{\mathbf{m}}_1 - \mathbf{s}\| - \|\mathbf{m}_1 - \mathbf{s}\|)/c = \tau_1^1 - \tau_1^0, \tag{10.8}$$

where $\tilde{\mathbf{m}}_i$ denotes the image of position $\mathbf{m}_i$. These three quantities are directly connected to RIRs, as illustrated in Fig. Figure 10.2(right). Let $V = \{\text{TDOA}, \text{iTDOA}, \text{TDOE}\} \in \mathbb{R}^3$. Following the 2D-SSL scheme described in Sec. § 10.2.2 and given the virtual microphone-array geometry (which depends on the relative position of microphones to the surface), V could in principle be used to estimate the 2D directional of arrival of the source. In the next section, we present a learning-based method to estimate V using audio features obtained from only two microphones.

10.4 Implementation and Results

To the best of the authors' knowledge, no reference implementation of algorithms for 2D-SSL using only 2 microphones is available to date. To check the validity of TDOA estimation, it is compared to GCC-PHAT using the true microphones (see Sec. § 10.2.1). For training and validation of the DNN we generate many random shoe-box room configurations using the software presented in [**Schimmel2009**]. This software implements both the image-method for simulating reflections and a ray-tracing algorithm for diffusion. Room widths are uniformly drawn at random in $[3, 9]$ m, heights in $[2, 4]$ m. Random source/microphones positions and absorption coefficients for the 6 surfaces are used, respecting the close-surface scenario. In particular, the microphones are at most 30 cm from the close-surface, placed 10 cm from each other, the absorption coefficients of the other walls are uniformly sampled in $(0.5, 1)$ and the one of the close-surface is in $(0, 0.5)$. The same realistic diffusion profile [Gaultier et al. 2017] is used for all surfaces. Around $90,000$ audio scenes are generated this way, yielding reverberation times (RT_{60}) between 20 ms and 250 ms.

Gaultier et al., "VAST: The virtual acoustic space traveler dataset"

For training and validation, the RIRs are convolved with 1 sec of white-noise (wn) with no additional noise. All signals and RIRs are sampled at 16 kHz. The STFT is performed on 1024 point with 50% overlap. Finally the features are computed as in (6.1) yielding a vector of size $D = 1534$ for each observation. While we validate the DNN on a portion of the dataset in a *holdout* fashion, the test is conducted on 200 new RIRs convolved with both wn and speech (sp) utterances. This set is generated similarly to the training and validation sets. Moreover the recordings are perturbed by external white

noise at 10 dB SNR (wn+n, sp+n). The speech signals are normalized speech utterances of various lengths (from 1 s to 6 s), randomly selected from the TIMIT corpus. A free and open-source Matlab implementation of SRP-PHAT[1] is used to aggregate local angular spectra obtained from the DNN's output. A sphere sampling with 0.5° resolution and coordinates $\theta \in [-179, 180]$ and $\phi \in [0, 90]$ is used for the DOA search.

[1] http://bass-db.gforge.inria.fr/bss_locate/

		nRMSE			ACCURACY	
	Input	TDOA	iTDOA	TDOE	$\theta < 10^\circ$	$\theta < 20^\circ$
MIRAGE	wn	0.18	0.28	0.25	4.10 (77)	5.97 (97)
MIRAGE	wn+n	0.68	0.69	0.89	5.00 (26)	9.89 (54)
MIRAGE	sp	0.31	0.34	0.56	4.83 (63)	7.26 (82)
MIRAGE	sp+n	0.99	0.98	1.48	4.60 (16)	9.88 (35)
GCC-PHAT	wn	0.21	-	-	4.22 (81)	6.19 (97)
GCC-PHAT	wn+n	0.68	-	-	4.03 (65)	5.34 (83)
GCC-PHAT	sp	0.32	-	-	4.08 (82)	5.34 (97)
GCC-PHAT	sp+n	1.38	-	-	4.70 (19)	8.38 (32)

TABLE 10.1: Normalize root mean squared error for TDOA estimation and mean angular error in $^\circ$ (with accuracies (%)) for AOA estimation with 10° and 20° angular tolerance.

DoA		ACCURACY $< 10^\circ$		ACCURACY $< 20^\circ$	
	Input	θ	ϕ	θ	ϕ
MIRAGE	wn	4.5 (59)	3.9 (71)	6.8 (79)	5.9 (88)
MIRAGE	wn+n	4.4 (18)	5.5 (26)	9.4 (35)	11.1 (66)
MIRAGE	sp	4.6 (45)	4.8 (59)	8.1 (71)	7.2 (83)
MIRAGE	sp+n	5.2 (17)	5.9 (12)	10.7 (38)	12.3 (43)

TABLE 10.2: Mean angular error in degree (with accuracies (%)) for 2D SSL (azimuth and elevation) with 10° and 20° tolerance.

TDOA estimation errors using the proposed approach and GCC-PHAT are presented in Table Table 10.1. Training a DNN to estimate TDOAs brings similar performances as GCC-PHAT in terms of nRMSE. Estimation of iTDOA and TDOE seems to be a harder task for the simple DNN we used. Nevertheless, our results confirm the possibility of retrieving early echoes from only two-microphone recordings. When some external noise is added, performance of both methods severely degrades. This is a well-know and expected behaviour for GCC-PHAT. It suggests that noise should be considered in the training phase of MIRAGE. When we compare the performance in terms of AOA, the two methods yield the same accuracy within a 20° threshold, as can be see in Table Table 10.1. When a smaller tolerance is considered, GCC-PHAT outdoes the proposed approach in accuracy, with comparable errors. Again, when adding noise, performance decreases. In Table Table 10.2 the performance of the full 2D-SSL pipeline is showed. Within a tolerance of 20°, the MIRAGE model allows estimation of both azimuth and elevation of the target source. However since in our data the 2 microphones were free to move, the inclinations of the true and image pairs are rarely flat. While this helps elevation estimation, it reduces the accuracy of predicting the right azimuth. While external noise is again decreasing the accuracy dramatically, it is interesting to notice that our DNN model trained and validated with white noise sources somewhat generalizes to speech sources.

10.5 CONCLUSION

In this paper we demonstrated how a simple echo model could allow 2D SSL with only two microphones, using simulated data. Future research will focus on extending this proof-of-concept to real data. The problem of echo-delay

estimation proved to be very challenging, and extensions of the proposed learning scheme will be developed to obtain more reliable estimations of angular spectra. Extensions of the method to better handle various types of noise and emitted signals will also be sought. Finally, applications of the MIRAGE framework to larger microphone arrays, higher order echoes and a variety of tasks beyond SSL will be explored.

11

Application of & dEchorate

11.1 Using the Data

"*Signal, a function that conveys information about a phenomenon.* [. . .] *Consider an acoustic wave, which can convey acoustic or music information.*"
—R. Priemer, *Introductory Signal Processing*

In this section we exemplify the utilization of the database considering three possible use-cases: acoustic echo estimation, echo-aware beamforming and room geometry estimation.

11.1.1 *Acoustic Echo Estimation*

Work in progress: Use BSN, Crocco and BLASTER for echo acoustic echo retrieval. Data: Sym/Real on Dirac/Noise/Speech

11.1.2 *Echo-aware Beamforming*

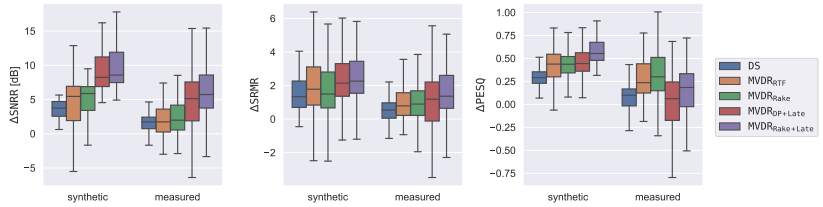


Figure 11.1: Comparison of echo-aware beamforming for the room configuration `011111` ($\texttt{RT}_{60} \approx$ 600 ms) on measured and synthetic data for all combinations of source-array positions in the `dEchorate` dataset.

As mentioned, the knowledge of early echoes should boost spatial filtering performances. However the perfect knowledge of such elements are of difficult estimation. To investigate this, we compare two types of spatial filters on both synthetic and measured data: echo-agnostic and echo-aware beamformers.
The formers do not need any echo-estimation step: they either ignore their contributions, such in the direct-path delay-and-sum beamformer (`DS`) [Van Trees 2004], or they consider coupling filters between pairs of microphones, called Relative Transfer Functions (RTFs) [Gannot et al. 2001][1]. The RTFs can be naturally incorporated in powerful beamforming algorithms achieving speech dereverberation and noise reduction in static [Schwartz et al. 2014] and dynamic scenarios [**Kodrasi2017evd**]. In this work, generalized eigenvector decomposition (GEVD) was used for the RTFs estimation [Doclo and Moonen 2003].
Echo-aware beamformers fall in the category of *rake receivers*, borrowing the idea from telecommunication where an antenna rakes (*i.e.* combines) coherent signals arriving from different propagation paths. In particular, they consider

[1] Note that, as opposed to AER, estimating the RTF is a non-blind problem.

"partial steering vectors" using known R echoes' delays and attenuations [Jan et al. 1995]. This methods have been used for interfer and noise suppression in [Dokmanić et al. 2015] and for noise and reverberation reduction [Javed et al. 2016; Kowalczyk 2019]. Here we assume that echoes are known, computed from the geometry as in Section 7.3.

In addition to the standard DS design, we considered the minimum-variance-distrortionless-response design echo-agnostic $\texttt{MVDR}_{\texttt{RTF}}$ build on RTFs as in [Schwartz et al. 2014] as echo-agnostic method, and the echo-aware beamformers $\texttt{MVDR}_{\texttt{Rake}}$ [Dokmanić et al. 2015] and its extension for dereverberation, the $\texttt{MVDR}_{\texttt{Rake+Late}}$ [Kowalczyk 2019] considers the statistical contribution of the reverberation tail.

The performances of the different designs are compared for enhancing a target speech in 5-channel mixture (that is, one linear array used in the dataset) featuring high reverberation and diffuse babble noise, opportunely scaled to given signal-to-noise ratio ($\texttt{SNR} \in \{0, 10, 20\}$). Using the `dEchorate` data, we considered the room configuration `011111` ($\texttt{RT}_{60} \approx 600$ ms) and all the possible combinations of target/array's positions. Both real and matching synthetic RIRs are used which are then convolved with anechoic speech from the WSJ corpus and corrupted by recorded diffuse noise.

The evaluation is conducted similarly to the one in [Kowalczyk 2019]. We considered the following metrics: the signal-to-noise-plus-reverberation improvment (ΔSNRR) in [dB] computed as difference between the input (SNRR) at the reference microphone and the SNRR at the filter output; the speech-to-reverberation-energy-modulation ratio improvement (ΔSRMR)[Falk et al. 2010] as measure of dereverberation; and the perceptual quality of the signal is evaluated using the PESQ score. As target signal, the clean signal convolved with the early part of the RIR (up to the R-th echo) is considered.
Numerical results are reported in Figure 11.1. The simple DS beamformer is outperformed by the other filters, since more information is used to reduce noise and late reverberation. When using synthetic data, the know echoes perfectly match numerically the components in the simulated RIRs. In this ideal scenario, one can see that the more information used the better the performances: RTF- and Rake- beamformers outperform the simple DS design; and including the late reverberation statistics boosts considerably all the performances. Interestingly RTF-based design performs similarly to the Rake-one. This can be explained by the fact that GEVD method tends to robustly consider the stronger and more stable components of the RTFs, which in reverberant and noisy static scenario's are similar to the earlier portion of the RIRs.
When it comes to measured RIRs, the little errors in echo estimation, due to calibration mismatch, lead to a drop in the performances. This is even more clear when considering the ΔPESQ metrics, as it accounts also for artifacts. Here the echo-agnostic $\texttt{MVDR}_{\texttt{RTF}}$ outperform the other methods.

11.1.3 *Room Geometry Estimation*

The shape of a convex room can be estimated knowing the positions of first-order image sources. Several methods have been proposed which take into account different levels of prior information and noise (see [Remaggi et al. 2016; Crocco et al. 2018] for a review). Nonetheless, when the echoes' TOA and

source id	1		2		3		4	
wall	DE	AE	DE	AE	DE	AE	DE	AE
west	0.74	8.99°	4.59	8.32°	5.89	5.75°	**0.05**	2.40°
east	**0.81**	0.08°	0.9	0.50°	*69.51*	55.70°	0.31	0.21°
south	3.94	16.08°	**0.18**	1.77°	*14.37*	18.55°	0.82	1.65°
north	1.34	0.76°	1.40	8.94°	**0.63**	0.17°	2.08	1.38°
floor	**5.19**	1.76°	7.27	2.66°	7.11	2.02°	5.22	1.90°
ceiling	1.16	0.28°	0.67	0.76°	**0.24**	1.16°	0.48	0.26°

TABLE 11.1: Distance errors (DE) in centimeters and angular errors (AE) in degrees between ground truth and estimated room sides using each of the sound source (#1 to #4) as a probe. For each wall, bold font is used in correspondence with the sources yielding the best DE and AE; while, the italic font highlight the outliers, if present.

their labeling are known for 4 non-coplanar microphones, one can perform this task using simple geometrical reasoning as in [**Dokmanic2013acoustic**]. In fact, the 3D coordinates of each image source can be retrieved solving a multilateration problem [Beck et al. 2008] and the position and orientation of each wall can be easily derived from the ISM equations as the plane bisecting the line joining the real source position and the position of its corresponding image (see Figure 11.2)

In `dEchorate` the annotation of all the first order images for all the sound sources are available. Table 11.1 shows the results of the estimation of the wall positions in terms of distance error (in centimeters) and surface orientation error (in degrees) using the four direct sources and all the 30 microphones, namely the 6 arrays). The room facets are estimated using each of the source as a probe. Despite few outliers, the majority of the facets are estimated correctly in terms of their placement and orientation with respect to the coordinate system computed in Section 7.3: for instance, for the source #4, all 6 surfaces were localized with less than 6 cm and 2.5° errors. Small errors are due to concurrency of multiple factors, such as tiny offsets in the annotation and the ideal shoebox approximation. In the real recording room, some gaps were present between revolving panels in the walls. In addition it is possible that for some source-receiver pairs the far-field assumption is not verified, causing the inaccuracy of *reverting* the ISM. Finally, the 2 outliers for the source #3 are due to a wrong annotation caused by source directivity and mis-classification. When a wall is "behind" the source, the energy of the related 1st reflection is very small and might not appear in the RIRs. This happened for the eastern wall and a second order image was taken instead. Secondly, the contribution of multiple reflections arriving at the same time can results in large late spikes in estimated RIRs. This effect is particularly amplified when the microphone and loudspeakers exhibit long impulse responses. As a consequence, some spikes can be miss-classified. This happened for the southern-wall were again a second-order image was taken instead. Nevertheless, this second type of errors can be manually corrected and the annotations updated.

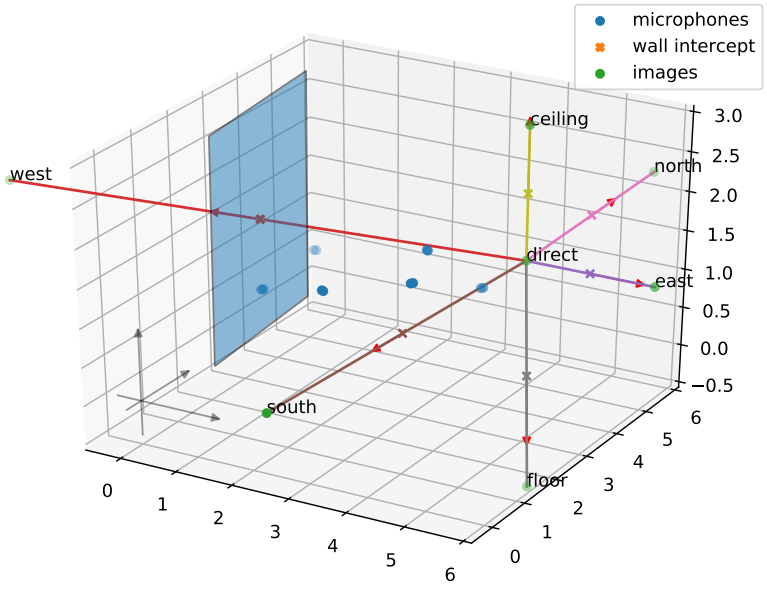


FIGURE 11.2: Images source estimation and reflector estimation for one of the sound sources in the dataset.

11.2 CONCLUSIONS AND PERSPECTIVES

This paper introduced a new database of room impulse responses featuring accurate annotation of early echoes and microphone positions. These data can be used to test methods in the room geometry estimation pipeline and in echo-aware audio signal processing. In particular, robustness of these methods can be validated against different levels of RT_{60}, SNR or even early echo density.

Part V

EPILOGUE

BIBLIOGRAPHY

Bibliography

Abed-Meraim, Karim, Philippe Loubaton, and Eric Moulines (1997). "A subspace algorithm for certain blind identification problems". In: *IEEE transactions on information theory* 43.2, pp. 499–511 (cit. on p. 54).

Affes, Sofiène and Yves Grenier (1997). "A signal subspace tracking algorithm for microphone array processing of speech". In: *IEEE Transactions on Speech and Audio Processing* 5.5, pp. 425–437. ISSN: 10636676. DOI: 10.1109/89.622565 (cit. on p. 5).

Ahmad, Rehan, Andy WH Khong, and Patrick A Naylor (2006). "Proportionate frequency domain adaptive algorithms for blind channel identification". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 5. IEEE, pp. V–V (cit. on p. 58).

Aissa-El-Bey, Abdeldjalil and Karim Abed-Meraim (2008). "Blind SIMO channel identification using a sparsity criterion". In: *2008 IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*. IEEE, pp. 271–275 (cit. on pp. 54, 64).

Al-Karawi, Khamis A and Duraid Y Mohammed (2019). "Early reflection detection using autocorrelation to improve robustness of speaker verification in reverberant conditions". In: *International Journal of Speech Technology* 22.4, pp. 1077–1084 (cit. on p. 52).

Allen, Jont B and David A Berkley (1979). "Image method for efficiently simulating small-room acoustics". In: *The Journal of the Acoustical Society of America* 65.4, pp. 943–950 (cit. on pp. 24–26, 70, 90).

Annibale, P., F. Antonacci, P. Bestagini, A. Brutti, A. Canclini, L. Cristoforetti, E. Habets, W. Kellermann, K. Kowalczyk, A. Lombard, E. Mabande, D. Markovic, P. Naylor, M. Omologo, R. Rabenstein, A. Sarti, P. Svaizer, and M. Thomas (2011). "The SCENIC project: Environment-aware sound sensing and rendering". In: *Procedia Computer Science* 7, pp. 150–152. ISSN: 18770509. DOI: 10.1016/j.procs.2011.09.039. URL: http://dx.doi.org/10.1016/j.procs.2011.09.039 (cit. on p. 5).

Annibale, Paolo, Jason Filos, Patrick A Naylor, and Rudolf Rabenstein (2012). "Geometric inference of the room geometry under temperature variations". In: *2012 5th International Symposium on Communications, Control and Signal Processing*. IEEE, pp. 1–4 (cit. on pp. 49, 50).

Antonacci, Fabio, Augusto Sarti, and Stefano Tubaro (2010). "Geometric reconstruction of the environment from its response to multiple acoustic emissions". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 2822–2825 (cit. on p. 51).

Antonacci, Fabio, Jason Filos, Mark RP Thomas, Emanuël AP Habets, Augusto Sarti, Patrick A Naylor, and Stefano Tubaro (2012). "Inference of room geometry from acoustic impulse responses". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.10, pp. 2683–2695 (cit. on pp. 51, 54, 81).

Aoshima, Nobuharu (1981). "Computer-generated pulse signal applied for sound measurement". In: *The Journal of the Acoustical Society of America* 69.5, pp. 1484–1488 (cit. on p. 48).

Asaei, Afsaneh, Mohammad Golbabaee, Herve Bourlard, and Volkan Cevher (2014). "Structured sparsity models for reverberant speech separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.3, pp. 620–633 (cit. on p. 90).

Baba, Youssef El, Andreas Walther, and Emanuël A.P. Habets (2018). "3D room geometry inference based on room impulse response stacks". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 26.5, pp. 857–872. ISSN: 23299290. DOI: 10.1109/TASLP.2017.2784298 (cit. on p. 81).

Badeau, Roland (2019). "Common mathematical framework for stochastic reverberation models". In: *The Journal of the Acoustical Society of America* 145.4, pp. 2733–2745 (cit. on pp. 23, 25).

Bal, Guillaume (2012). "Introduction to inverse problems". In: *Lecture Notes-Department of Applied Physics and Applied Mathematics, Columbia University, New York* (cit. on pp. 5, 6).

Barron, Michael (1971). "The subjective effects of first reflections in concert halls—the need for lateral reflections". In: *Journal of sound and vibration* 15.4, pp. 475–494 (cit. on p. 28).

Beck, Amir, Petre Stoica, and Jian Li (2008). "Exact and approximate solutions of source localization problems". In: *IEEE Transactions on Signal Processing* 56.5, pp. 1770–1778. ISSN: 1053587X. DOI: 10.1109/TSP.2007.909342 (cit. on pp. 80, 109).

Bee, Mark A and Christophe Micheyl (2008). "The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it?" In: *Journal of comparative psychology* 122.3, p. 235 (cit. on p. 87).

Bello, Juan Pablo, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler (2005). "A tutorial on onset detection in music signals". In: *IEEE Transactions on speech and audio processing* 13.5, pp. 1035–1047 (cit. on pp. 49, 50).

Bertin, Nancy, Ewen Camberlein, Romain Lebarbenchon, Emmanuel Vincent, Sunit Sivasankaran, Irina Illina, and Frédéric Bimbot (2019). "VoiceHome-2, an extended corpus for multichannel speech processing in real homes". In: *Speech Communication* 106, pp. 68–78 (cit. on p. 57).

Betlehem, Terence, Paul D Teal, and Yusuke Hioka (2012). "Efficient crosstalk canceler design with impulse response shortening filters". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 393–396 (cit. on p. 48).

Böck, Sebastian, Florian Krebs, and Markus Schedl (2012). "Evaluating the Online Capabilities of Onset Detection Methods." In: *ISMIR*, pp. 49–54 (cit. on pp. 59, 69).

Boutin, Mireille and Gregor Kemper (2020). "A drone can hear the shape of a room". In: *SIAM Journal on Applied Algebra and Geometry* 4.1, pp. 123–140 (cit. on p. 5).

Bredies, Kristian and Marcello Carioni (2020). "Sparsity of solutions for variational inverse problems with finite-dimensional data". In: *Calculus of Variations and Partial Differential Equations* 59.1, p. 14 (cit. on p. 67).

Candès, Emmanuel J and Carlos Fernandez-Granda (2014). "Towards a mathematical theory of super-resolution". In: *Communications on pure and applied Mathematics* 67.6, pp. 906–956 (cit. on pp. 62, 67).

Chen, Jingdong, Jacob Benesty, and Yiteng Arden Huang (2006). "Time delay estimation in room acoustic environments: an overview". In: *EURASIP Journal on Advances in Signal Processing* 2006.1, p. 026503 (cit. on p. 61).

Cheng, Tian, Matthias Mauch, Emmanouil Benetos, Simon Dixon, et al. (2016). "An attack/decay model for piano transcription". In: ISMIR (cit. on pp. 49, 50).

Cherry, Colin (1953). "Cocktail party problem". In: *Journal of the Acoustical Society of America* 25, pp. 975–979 (cit. on p. 87).

Chi, Yuejie, Louis L Scharf, Ali Pezeshki, and A Robert Calderbank (2011). "Sensitivity to basis mismatch in compressed sensing". In: *IEEE Transactions on Signal Processing* 59.5, pp. 2182–2195 (cit. on p. 55).

Čmejla, Jaroslav, Tomáš Kounovskỳ, Sharon Gannot, Zbyněk Koldovskỳ, and Pinchas Tandeitnik (2019). "MIRaGe: Multichannel Database Of Room Impulse Responses Measured On High-Resolution Cube-Shaped Grid In Multiple Acoustic Conditions". In: *arXiv preprint arXiv:1907.12421* (cit. on pp. 55, 57, 77).

Condat, Laurent and Akira Hirabayashi (2013). "Robust spike train recovery from noisy data by structured low rank approximation". In: *Int. Conf. Sampl. Theory Appl.(SAMPTA), Bremen, Germany* (cit. on pp. 50, 56).

— (2015). "Cadzow denoising upgraded: A new projection method for the recovery of Dirac pulses from noisy linear measurements". In: (cit. on pp. 46, 56).

Crocco, Marco and Alessio Del Bue (2015). "Room impulse response estimation by iterative weighted l 1-norm". In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1895–1899 (cit. on pp. 54, 64, 69).

— (2016a). "Estimation of TDOA for room reflections by iterative weighted l 1 constraint". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3201–3205 (cit. on pp. 54, 55, 62, 64, 69).

— (2016b). "Estimation of TDOA for room reflections by iterative weighted l1 constraint". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2016-May. 2. IEEE, pp. 3201–3205. ISBN: 9781479999880. DOI: 10.1109/ICASSP.2016.7472268 (cit. on p. 80).

Crocco, Marco, Alessio Del Bue, Matteo Bustreo, and Vittorio Murino (2012). "A closed form solution to the microphone position self-calibration problem". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2597–2600 (cit. on p. 55).

Crocco, Marco, Andrea Trucco, Vittorio Murino, and Alessio Del Bue (2014). "Towards fully uncalibrated room reconstruction with sound". In: *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 910–914 (cit. on pp. 51, 52).

Crocco, Marco, Andrea Trucco, and Alessio Del Bue (2017). "Uncalibrated 3D room geometry estimation from sound impulse responses". In: *Journal of the Franklin Institute* 354.18, pp. 8678–8709 (cit. on pp. 5, 49, 51, 54, 55, 57).

— (2018). "Room Reflector Estimation from Sound by Greedy Iterative Approach". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2018-April, pp. 6877–6881. ISBN: 9781538646588. DOI: 10.1109/ICASSP.2018.8461640 (cit. on p. 108).

Davis, AH and N Fleming (1926). "Sound pulse photography as applied to the study of architectural acoustics". In: *Journal of Scientific Instruments* 3.12, p. 393 (cit. on p. 19).

De Castro, Yohann and Fabrice Gamboa (2012). "Exact reconstruction using Beurling minimal extrapolation". In: *Journal of Mathematical Analysis and applications* 395.1, pp. 336–354 (cit. on p. 67).

Defrance, Guillaume, Laurent Daudet, and Jean-Dominique Polack (2008a). "Detecting arrivals within room impulse responses using matching pursuit". In: *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland*. Vol. 10. Citeseer, pp. 307–316 (cit. on pp. 49, 50).

— (2008b). "Finding the onset of a room impulse response: Straightforward?" In: *The Journal of the Acoustical Society of America* 124.4, EL248–EL254 (cit. on pp. 49, 55, 82).

Deleforge, Antoine, Diego Di Carlo, Martin Strauss, Romain Serizel, and Lucio Marcenaro (2019). "Audio-Based Search and Rescue With a Drone: Highlights From the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]". In: *IEEE Signal Processing Magazine* 36.5, pp. 138–144 (cit. on p. 10).

Denoyelle, Quentin, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies (2019). "The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy". In: *Inverse Problems* 36.1, p. 014001 (cit. on pp. 62, 68).

Di Carlo, Diego and Antoine Deleforge. *HRI-JF collaboration - Final Phase II Deliverable*. Tech. rep. Inria Nancy - Grand Est (cit. on p. 72).

Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019a). "Mirage: 2D Source Localization Using Microphone Pair Augmentation with Echoes". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2019-May, pp. 775–779. ISBN: 9781479981311. DOI: 10.1109/ICASSP.2019.8683534 (cit. on p. 5).

— (2019b). "Mirage: 2d source localization using microphone pair augmentation with echoes". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 775–779 (cit. on pp. 10, 71, 72).

Di Carlo, Diego, Clement Elvira, Antoine Deleforge, Nancy Bertin, and Rémi Gribonval (2020). "Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 156–160 (cit. on pp. 10, 55, 61).

Di Carlo, Diego, Pinchas Tandeitnik, Sharon Gannot, Antoine Deleforge, and Nancy Bertin (2021). "dEchorate: a calibrated Room Impulse Response database for acoustic echo retrieval". In: *Workin progess* (cit. on p. 10).

DiBiase, Joseph H, Harvey F Silverman, and Michael S Brandstein (2001). "Robust localization in reverberant rooms". In: *Microphone Arrays*. Springer, pp. 157–180 (cit. on pp. 52, 56).

Doclo, Simon and Marc Moonen (2003). "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments". In: *EURASIP Journal on Advances in Signal Processing* 2003.11, p. 495250 (cit. on p. 107).

Dokmanić, Ivan, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli (2013). "Acoustic echoes reveal room shape". In: *Proceedings of the National Academy of Sciences* 110.30, pp. 12186–12191 (cit. on pp. 51, 57, 101).

Dokmanić, Ivan, Robin Scheibler, and Martin Vetterli (2015). "Raking the Cocktail Party". In: *IEEE Journal on Selected Topics in Signal Processing* 9.5, pp. 825–836. ISSN: 19324553. DOI: 10.1109/JSTSP.2015.2415761 (cit. on pp. 5, 108).

Dokmanić, Ivan, Robin Scheibler, and Martin Vetterli (2015a). "Raking the Cocktail Party". In: *IEEE J. Sel. Top. Signal Process.* 9.5, pp. 825–836 (cit. on pp. 87, 90).

— (2015b). "Raking the cocktail party". In: *IEEE journal of selected topics in signal processing* 9.5, pp. 825–836 (cit. on pp. 54, 91, 101).

Dokmanić, Ivan, Juri Ranieri, and Martin Vetterli (2015). "Relax and unfold: Microphone localization with Euclidean distance matrices". In: *European Signal Processing Conference, (EUSIPCO)*, pp. 265–269. ISBN: 9780992862633. DOI: 10.1109/EUSIPCO.2015.7362386 (cit. on p. 80).

Dokmanić, Ivan, Laurent Daudet, and Martin Vetterli (2016). "From acoustic room reconstruction to SLAM". In: *Proc. IEEE ICASSP*. Shanghai, CHN, pp. 6345–6349 (cit. on p. 87).

Duffy, Dean G (2015). *Green's functions with applications*. CRC Press (cit. on pp. 14, 16).

Dunn, Chris and Malcolm J Hawksford (1993). "Distortion immunity of MLS-derived impulse response measurements". In: *Journal of the Audio Engineering Society* 41.5, pp. 314–335 (cit. on p. 48).

Duong, Ngoc QK, Emmanuel Vincent, and Rémi Gribonval (2010). "Under-determined reverberant audio source separation using a full-rank spatial covariance model". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, pp. 1830–1840 (cit. on pp. 53, 89).

Eaton, James, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor (Oct. 2016). "Estimation of Room Acoustic Parameters: The ACE Challenge". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 24, pp. 1681–1693. ISSN: 23299290. DOI: 10.1109/TASLP.2016.2577502 (cit. on p. 81).

El Baba, Youssef, Andreas Walther, and Emanuël AP Habets (2017). "Time of arrival disambiguation using the linear Radon transform". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 106–110 (cit. on pp. 51, 52).

Falk, Tiago H, Chenxi Zheng, and Wai-Yip Chan (2010). "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, pp. 1766–1774 (cit. on p. 108).

Farina, Angelo (2000). "Simultaneous measurement of impulse response and distortion with a swept-sine technique". In: *Audio Engineering Society Convention 108*. Audio Engineering Society (cit. on p. 49).

— (2007). "Advancements in impulse response measurements by sine sweeps". In: *Audio Engineering Society Convention 122*. Audio Engineering Society (cit. on pp. 49, 79).

Ferguson, Eric L, Stefan B Williams, and Craig T Jin (2019). "Improved multipath time delay estimation using cepstrum subtraction". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 551–555 (cit. on p. 50).

Févotte, C and J Idier (2011). "Algorithms for nonnegative matrix factorization with the β-divergence". In: *Neural computation* 23.9, pp. 2421–2456 (cit. on p. 93).

Filos, Jason, Antonio Canclini, Mark RP Thomas, Fabio Antonacci, Augusto Sarti, and Patrick A Naylor (2011). "Robust inference of room geometry from acoustic measurements using the Hough transform". In: *2011 19th European Signal Processing Conference*. IEEE, pp. 161–165 (cit. on p. 51).

Fourier, Jean Baptiste Joseph (1822). *Théorie analytique de la chaleur*. F. Didot (cit. on p. 34).

Gannot, Sharon, David Burshtein, and Ehud Weinstein (2001). "Signal enhancement using beamforming and nonstationarity with applications to speech". In: *IEEE Transactions on Signal Processing* 49.8, pp. 1614–1626 (cit. on pp. 5, 41, 55, 107).

Garofolo, John S, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue (1993). "TIMIT acoustic-phonetic continuous speech corpus". In: *Linguistic data consortium* 10.5, p. 0 (cit. on p. 97).

Gaultier, Clément, Saurabh Kataria, and Antoine Deleforge (2017). "VAST: The virtual acoustic space traveler dataset". In: *Lecture Notes in Computer Science*. Vol. 10169 LNCS, pp. 68–79. ISBN: 9783319535463. DOI: 10.1007/978-3-319-53547-0{_}7 (cit. on pp. 74, 101, 104).

Genovese, Andrea F, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J Tashev (2019). "Blind room volume estimation from single-channel noisy speech". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 231–235 (cit. on pp. 56, 57).

Gilloire, Andre and Martin Vetterli (1992). "Adaptive filtering in sub-bands with critical sampling: analysis, experiments, and application to acoustic echo cancellation". In: *IEEE transactions on signal processing* 40.ARTICLE, pp. 1862–1875 (cit. on p. 40).

Griesinger, David (1997). "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces". In: *Acta Acustica united with Acustica* 83.4, pp. 721–731 (cit. on p. 28).

Guillemain, Philippe and Richard Kronland-Martinet (1996). "Characterization of acoustic signals through continuous linear time-frequency representations". In: *Proceedings of the IEEE* 84.4, pp. 561–585 (cit. on p. 50).

Habets, Emanuel AP (2006). "Room impulse response generator". In: *Technische Universiteit Eindhoven, Tech. Rep* 2.2.4, p. 1 (cit. on p. 26).

Habets, Emanuël AP and Sharon Gannot (2007). "Generating sensor signals in isotropic noise fields". In: *The Journal of the Acoustical Society of America* 122.6, pp. 3464–3470 (cit. on p. 33).

Hadad, Elior, Florian Heese, Peter Vary, and Sharon Gannot (2014). "Multichannel audio database in various acoustic environments". In: *2014 14th International Workshop on Acoustic Signal Enhancement, IWAENC 2014*, pp. 313–317. ISBN: 9781479968084. DOI: 10.1109/IWAENC.2014.6954309 (cit. on p. 76).

Heinz, Renate (1993). "Binaural room simulation based on an image source model with addition of statistical methods to include the diffuse sound scattering of walls and to predict the reverberant tail". In: *Applied Acoustics* 38.2-4, pp. 145–159 (cit. on p. 24).

Huang, Yiteng and Jacob Benesty (2003). "A class of frequency-domain adaptive approaches to blind multichannel identification". In: *IEEE Transactions on signal processing* 51.1, pp. 11–24 (cit. on pp. 53, 58).

Huang, Yiteng, Jacob Benesty, and Jingdong Chen (2005). "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment". In: *IEEE Transactions on Speech and Audio Processing* 13.5, pp. 882–895 (cit. on p. 90).

Jager, Ingmar, Richard Heusdens, and Nikolay D Gaubitch (2016). "Room geometry estimation from acoustic echoes using graph-based echo labeling". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 51).

Jan, Ea Ee, Piergiorgio Svaizer, and J. L. Flanagan (1995). "Matched-filter processing of microphone array for spatial volume selectivity". In: *IEEE International Symposium on Circuits and Systems*. Vol. 2. 908. IEEE, pp. 1460–1463. ISBN: 0780325702. DOI: 10.1109/iscas.1995.521409 (cit. on pp. 5, 108).

Javed, Hamza A, Alastair H Moore, and Patrick A Naylor (2016). "Spherical microphone array acoustic rake receivers". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 111–115 (cit. on p. 108).

Jensen, Jesper Rindom, Usama Saqib, and Sharon Gannot (2019). "An EM method for multichannel TOA and DOA estimation of acoustic echoes". In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 120–124 (cit. on pp. 5, 52, 61).

Jia, Hongjian, Xiukun Li, Xiangxia Meng, and Yang Yang (2017). "Extraction of echo characteristics of underwater target based on cepstrum method". In: *Journal of Marine Science and Application* 16.2, pp. 216–224 (cit. on p. 50).

Kearney, Gavin, Marcin Gorzel, Henry Rice, and Frank Boland (2012). "Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields". In: *Acta Acustica united with Acustica* 98.1, pp. 61–71 (cit. on p. 28).

Kelly, Ian J and Francis M Boland (2014). "Detecting arrivals in room impulse responses with dynamic time warping". In: *IEEE/ACM transactions on audio, speech, and language processing* 22.7, pp. 1139–1147 (cit. on pp. 49, 50).

Kitic, Srdan (2015). "Cosparse regularization of physics-driven inverse problems". PhD thesis. Rennes 1 (cit. on pp. 5, 6).

Kodrasi, Ina and Simon Doclo (2017). "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods". In: *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, pp. 116–120 (cit. on p. 55).

Koldovsky, Zbynek and Petr Tichavsky (2015). "Sparse reconstruction of incomplete relative transfer function: Discrete and continuous time domain". In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 394–398 (cit. on p. 55).

Koldovskỳ, Zbyněk, Jiří Málek, and Sharon Gannot (2015). "Spatial source subtraction based on incomplete measurements of relative transfer function". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.8, pp. 1335–1347 (cit. on p. 55).

Kowalczyk, Konrad (2019). "Raking early reflection signals for late reverberation and noise reduction". In: *The Journal of the Acoustical Society of America* 145.3, EL257–EL263. ISSN: 0001-4966. DOI: 10.1121/1.5095535. URL: http://dx.doi.org/10.1121/1.5095535 (cit. on pp. 5, 108).

Kowalczyk, Konrad, Emanuël AP Habets, Walter Kellermann, and Patrick A Naylor (2013). "Blind system identification using sparse learning for TDOA estimation of room reflections". In: *IEEE Signal Processing Letters* 20.7, pp. 653–656 (cit. on pp. 54, 64).

Krokstad, Asbjørn, Staffan Strom, and Svein Sørsdal (1968). "Calculating the acoustical room response by the use of a ray tracing technique". In: *Journal of Sound and Vibration* 8.1, pp. 118–125 (cit. on p. 19).

Kulowski, Andrzej (1985). "Algorithmic representation of the ray tracing technique". In: *Applied Acoustics* 18.6, pp. 449–469 (cit. on p. 23).

Kuster, Martin (2008). "Reliability of estimating the room volume from a single room impulse response". In: *The Journal of the Acoustical Society of America* 124.2, pp. 982–993 (cit. on p. 49).

Kuttruff, Heinrich (2016). *Room acoustics*. CRC Press (cit. on pp. 11, 14, 17–19, 49).

Le Roux, Jonathan, John R Hershey, and Felix Weninger (2015). "Deep NMF for speech separation". In: *Proc. IEEE ICASSP*, pp. 66–70 (cit. on p. 89).

Lebarbenchon, Romain, Ewen Camberlein, Diego Di Carlo, Clément Gaultier, Antoine Deleforge, and Nancy Bertin (2018). "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge". In: *arXiv preprint arXiv:1812.05901* (cit. on p. 10).

Lee, Daniel D. and H. Sebastian Seung (2001). "Algorithms for Non-negative Matrix Factorization". In: *Advances in Neural Information Processing Systems 13*. Ed. by T. K. Leen, T. G. Dietterich, and V. Tresp. MIT Press, pp. 556–562 (cit. on p. 93).

Leglaive, Simon, Roland Badeau, and Gaël Richard (2015). "Multichannel audio source separation with probabilistic reverberation modeling". In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 1–5 (cit. on p. 90).

— (2016). "Multichannel audio source separation with probabilistic reverberation priors". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.12, pp. 2453–2465 (cit. on pp. 5, 53, 54, 90).

— (2018). "Student's t source and mixing models for multichannel audio source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.6, pp. 1154–1168 (cit. on pp. 53, 54).

Li, Xiaofei, Laurent Girin, and Radu Horaud (2019). "Expectation-maximisation for speech source separation using convolutive transfer function". In: *CAAI Transactions on Intelligence Technology* 4.1, pp. 47–53 (cit. on p. 89).

Lin, Yuanqing and Daniel D Lee (2006). "Bayesian regularization and nonnegative deconvolution for room impulse response estimation". In: *IEEE Transactions on Signal Processing* 54.3, pp. 839–847 (cit. on p. 49).

Lin, Yuanqing, Jingdong Chen, Youngmoo Kim, and Daniel D Lee (2007). "Blind sparse-nonnegative (BSN) channel identification for acoustic time-difference-of-arrival estimation". In: *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 106–109 (cit. on pp. 54, 63, 64, 69).

— (2008). "Blind channel identification for speech dereverberation using l1-norm sparse learning". In: *Advances in Neural Information Processing Systems*, pp. 921–928 (cit. on pp. 54, 63–65).

Loutridis, Spyros J (2005). "Decomposition of impulse responses using complex wavelets". In: *Journal of the Audio Engineering Society* 53.9, pp. 796–811 (cit. on p. 50).

Morgan, Dennis R, Jacob Benesty, and M Mohan Sondhi (1998). "On the evaluation of estimated impulse responses". In: *IEEE Signal processing letters* 5.7, pp. 174–176 (cit. on p. 58).

Müller, Meinard (2015). *Fundamentals of Music Processing*. Springer Verlag. ISBN: 978-3-319-21944-8 (cit. on pp. 39, 95).

Naylor, Patrick A, Anastasis Kounoudes, Jon Gudnason, and Mike Brookes (2006). "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1, pp. 34–43 (cit. on p. 49).

Nesta, Francesco and Maurizio Omologo (2012). "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 222–230 (cit. on p. 89).

Nugraha, Aditya Arie, Antoine Liutkus, and Emmanuel Vincent (2016). "Multichannel audio source separation with deep neural networks". In: *IEEE/ACM Trans. Audio, Speech, Language Process.* 24.9, pp. 1652–1664 (cit. on p. 89).

O'Donovan, Adam E, Ramani Duraiswami, and Dmitry N Zotkin (2010). "Automatic matched filter recovery via the audio camera". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 2826–2829 (cit. on pp. 53, 56).

O'Donovan, Adam, Ramani Duraiswami, and Dmitry Zotkin (2008). "Imaging concert hall acoustics using visual and audio cameras". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 5284–5287 (cit. on pp. 53, 56).

Oppenheim, Alan V (1987). *Signals and Systems: An Introduction to Analog and Digital Signal Processing*. MIT Center for Advanced Engineering Study (cit. on p. 39).

Ozerov, Alexey and Cédric Févotte (2009). "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3, pp. 550–563 (cit. on p. 53).

— (2010). "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation". In: *IEEE Trans. Audio, Speech, Language Process.* 18.3, pp. 550–563 (cit. on pp. 89, 93, 94, 96, 97).

Parhizkar, Reza, Ivan Dokmanić, and Martin Vetterli (2014). "Single-channel indoor microphone localization". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1434–1438 (cit. on p. 51).

Park, Yongsung, Woojae Seong, and Youngmin Choo (2017). "Compressive time delay estimation off the grid". In: *The Journal of the Acoustical Society of America* 141.6, EL585–EL591 (cit. on p. 61).

Paul, Douglas B and Janet M Baker (1992). "The design for the Wall Street Journal-based CSR corpus". In: *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, pp. 357–362 (cit. on p. 79).

Pavlović, Milan, Dragan M Ristić, Irini Reljin, and Miomir Mijić (2016). "Multifractal analysis of visualized room impulse response for detecting early reflections". In: *The Journal of the Acoustical Society of America* 139.5, EL113–EL117 (cit. on p. 50).

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830 (cit. on p. 97).

Pierce, Allan D (2019). *Acoustics: an introduction to its physical principles and applications*. Springer (cit. on pp. 14, 19).

Plinge, Axel, Florian Jacob, Reinhold Haeb-Umbach, and Gernot A. Fink (2016). "Acoustic microphone geometry calibration". In: *IEEE Signal Processing Magazine* July, pp. 14–28 (cit. on p. 80).

Qi, Yuanlei, Feiran Yang, Ming Wu, and Jun Yang (2019). "A Broadband Kalman Filtering Approach to Blind Multichannel Identification". In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 102.6, pp. 788–795 (cit. on p. 54).

Raffel, Colin, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel (2014). "mir_eval: A transparent implementation of common MIR metrics". In: *Proc. ISMIR* (cit. on p. 97).

Remaggi, Luca, Philip JB Jackson, Philip Coleman, and Wenwu Wang (2016). "Acoustic reflector localization: novel image source reversion and direct localization methods". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.2, pp. 296–309 (cit. on pp. 49, 57, 82, 108).

Remaggi, Luca, Philip JB Jackson, and Wenwu Wang (2019). "Modeling the Comb Filter Effect and Interaural Coherence for Binaural Source Separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12, pp. 2263–2277 (cit. on p. 57).

Ribeiro, Flavio, Demba Elimane Ba, and Cha Zhang (2010a). "Turning Enemies Into Friends: Using Reflections to Improve Sound Source Localization". In: *ICME* (cit. on p. 87).

Ribeiro, Flávio, Demba Ba, Cha Zhang, and Dinei Florêncio (2010b). "Turning enemies into friends: Using reflections to improve sound source localization". In: *2010 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 731–736 (cit. on pp. 5, 54).

Rickard, Scott (2007). "The DUET blind source separation algorithm". In: *Blind Speech Separation*, pp. 217–241 (cit. on p. 89).

Ristić, Dragan M, Milan Pavlović, Dragana Šumarac Pavlović, and Irini Reljin (2013). "Detection of early reflections using multifractals". In: *The Journal of the Acoustical Society of America* 133.4, EL235–EL241 (cit. on p. 50).

Rotili, Rudy, Claudio De Simone, Alessandro Perelli, Simone Cifani, and Stefano Squartini (2010). "Joint multichannel blind speech separation and dereverberation: A real-time algorithmic implementation". In: *International Conference on Intelligent Computing*. Springer, pp. 85–93 (cit. on p. 90).

Roy, Robert, Arogyaswami Paulraj, and Thomas Kailath (1986). "ESPRIT–A subspace rotation approach to estimation of parameters of cisoids in noise". In: *IEEE transactions on acoustics, speech, and signal processing* 34.5, pp. 1340–1342 (cit. on p. 50).

Rudin, Walter (1987). "Real and complex analysis (mcgraw-hill international editions: Mathematics series)". In: (cit. on p. 66).

Salvati, Daniele, Carlo Drioli, and Gian Luca Foresti (2016). "Sound source and microphone localization from acoustic impulse responses". In: *IEEE Signal Processing Letters* 23.10, pp. 1459–1463 (cit. on p. 51).

Santamarina, J Carlos and Dante Fratta (2005). "Discrete signals and inverse problems". In: *An Introduction for Engineers and Scientists. UK: Wiley & Sons* (cit. on p. 5).

Saqib, Usama, Sharon Gannot, and Jesper Rindom Jensen (2020). "Estimation of acoustic echoes using expectation-maximization methods". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2020.1, pp. 1–15 (cit. on p. 52).

Sato, Yoichi (1975). "A method of self-recovering equalization for multilevel amplitude-modulation systems". In: *IEEE Transactions on communications* 23.6, pp. 679–682 (cit. on p. 53).

Savioja, Lauri and U Peter Svensson (2015). "Overview of geometrical room acoustic modeling techniques". In: *The Journal of the Acoustical Society of America* 138.2, pp. 708–730 (cit. on pp. 18, 19, 23, 24, 29).

Scheibler, Robin (2017). “Rake, Peel, Sketch: The Signal Processing Pipeline Revisited”. PhD thesis. Switzerland: EPFL (cit. on p. 87).

Scheibler, Robin, Ivan Dokmanić, and Martin Vetterli (2015). “Raking Echoes in the Time Domain”. In: *IEEE ICASSP*. Brisbane (cit. on p. 87).

Scheibler, Robin, Eric Bezzam, and Ivan Dokmanić (2018a). “Pyroomacoustics: A Python package for audio room simulations and array processing algorithms”. In: *Proc. IEEE ICASSP*. accepted. Calgary, CA (cit. on p. 95).

— (2018b). “Pyroomacoustics: A python package for audio room simulation and array processing algorithms”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 351–355 (cit. on p. 70).

Scheibler, Robin, Diego Di Carlos, Antoine Deleforge, and Ivan Dokmanic (2018c). “Separake: Source Separation with a Little Help from Echoes”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2018-April, pp. 6897–6901. ISBN: 9781538646588. DOI: 10.1109/ICASSP.2018.8461345. URL: http://arxiv.org/abs/1711.06805 (cit. on p. 5).

Scheibler, Robin, Diego Di Carlo, Antoine Deleforge, and Ivan Dokmanić (2018d). “Separake: Source separation with a little help from echoes”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6897–6901 (cit. on pp. 10, 53, 71, 89).

Scheuing, Jan and Bin Yang (2006). “Disambiguation of TDOA estimates in multi-path multi-source environments (DATEMM)”. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 4. IEEE, pp. IV–IV (cit. on p. 51).

Schimmel, Steven M, Martin F Muller, and Norbert Dillier (2009). “A fast and accurate “shoebox” room acoustics simulator”. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 241–244 (cit. on pp. 23, 74).

Schmidt, Mikkel N and Rasmus K Olsson (2006). “Single-channel speech separation using sparse non-negative matrix factorization”. In: *Ninth International Conference on Spoken Language Processing* (cit. on p. 95).

Schröder, Dirk, Philipp Dross, and Michael Vorländer (2007). “A fast reverberation estimator for virtual environments”. In: *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society (cit. on pp. 23, 24).

Schroeder, Manfred R (1979). “Integrated-impulse method measuring sound decay without using impulses”. In: *The Journal of the Acoustical Society of America* 66.2, pp. 497–500 (cit. on pp. 48, 49).

Schwartz, Ofer, Sharon Gannot, and Emanuël AP Habets (2014). “Multi-microphone speech dereverberation and noise reduction using relative early transfer functions”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.2, pp. 240–251 (cit. on pp. 107, 108).

Shih, Oliver and Anthony Rowe (2019). “Can a phone hear the shape of a room?” In: *ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, pp. 277–288 (cit. on p. 5).

Smaragdis, Paris, Madhusudana Shashanka, and Bhiksha Raj (2009). “A sparse non-parametric approach for single channel separation of known sounds”. In: *Advances in neural information processing systems*, pp. 1705–1713 (cit. on p. 95).

Sturmel, Nicolas, Antoine Liutkus, Jonathan Pinel, Laurent Girin, Sylvain Marchand, Gaël Richard, Roland Badeau, and Laurent Daudet (2012). “Linear mixing models for active listening of music productions in realistic studio conditions”. In: *Proceedings of the Audio Engineering Society Convention*. 8594. IEEE (cit. on p. 32).

Sun, D L and G J Mysore (2013). “Universal speech models for speaker independent single channel source separation”. In: *IEEE ICASSP*, pp. 141–145 (cit. on pp. 93, 96).

Szöke, Igor, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Honza Černockỳ (2019). “Building and evaluation of a real room impulse response dataset”. In: *IEEE Journal of Selected Topics in Signal Processing* 13.4, pp. 863–876 (cit. on pp. 49, 56, 57).

Tammen, Marvin, Ina Kodrasi, and Simon Doclo (2018). "Iterative Alternating Least-Squares Approach to Jointly Estimate the RETFs and the Diffuse PSD". In: *Speech Communication; 13th ITG-Symposium*. VDE, pp. 1–5 (cit. on p. 58).

Tervo, Sakari (2011). "Localization and tracing of early acoustic reflections". PhD thesis (cit. on p. 53).

Tervo, Sakari and Archontis Politis (2015). "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.10, pp. 1539–1551 (cit. on p. 53).

Tervo, Sakari, Teemu Korhonen, and Tapio Lokki (2011). "Estimation of reflections from impulse responses". In: *Building Acoustics* 18.1-2, pp. 159–173 (cit. on p. 52).

Thomas, Matthew Reuben (2017). "Wayverb: A Graphical Tool for Hybrid Room Acoustics Simulation". PhD thesis. University of Huddersfield (cit. on pp. 22, 25).

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288 (cit. on p. 63).

Tong, Lang and Sylvie Perreau (1998). "Multichannel blind identification: From subspace to maximum likelihood methods". In: *Proceedings of the IEEE* 86.10, pp. 1951–1968 (cit. on pp. 53, 54).

Tong, Lang, Guanghan Xu, and Thomas Kailath (1994). "Blind identification and equalization based on second-order statistics: A time domain approach". In: *IEEE Transactions on information Theory* 40.2, pp. 340–349 (cit. on pp. 54, 64).

Tukuljac, Helena Peic (2020). "Sparse and Parametric Modeling with Applications to Acoustics and Audio". PhD thesis. École polytechnique fédérale de Lausanne (cit. on p. 56).

Tukuljac, Helena Peic, Antoine Deleforge, and Rémi Gribonval (2018). "MULAN: a blind and off-grid method for multichannel echo retrieval". In: *Advances in Neural Information Processing Systems*, pp. 2182–2192 (cit. on pp. 37, 49, 55, 56, 62).

Tuzlukov, Vyacheslav (2018). *Signal processing noise*. CRC Press (cit. on p. 33).

Usher, John (2010). "An improved method to determine the onset timings of reflections in an acoustic impulse response". In: *The Journal of the Acoustical Society of America* 127.4, EL172–EL177 (cit. on pp. 49, 50).

Välimäki, Vesa, Julian Parker, Lauri Savioja, Julius O Smith, and Jonathan Abel (2016). "More than 50 years of artificial reverberation". In: *Audio engineering society conference: 60th international conference: dreams (dereverberation and reverberation of audio, music, and speech)*. Audio Engineering Society (cit. on pp. 22, 28).

Van Trees, Harry L (2004). *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons (cit. on p. 107).

Vanderveen, Michaela C, Constantinos B Papadias, and Arogyaswami Paulraj (1997). "Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array". In: *IEEE Communications letters* 1.1, pp. 12–14 (cit. on p. 52).

Varzandeh, Reza, Maja Taseska, and Emanuël AP Habets (2017). "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation". In: *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, pp. 11–15 (cit. on p. 58).

Venkateswaran, Sriram and Upamanyu Madhow (2012). "Localizing multiple events using times of arrival: a parallelized, hierarchical approach to the association problem". In: *IEEE Transactions on Signal Processing* 60.10, pp. 5464–5477 (cit. on p. 51).

Verhaevert, Jo, Emmanuel Van Lil, and Antoine Van de Capelle (2004). "Direction of arrival (DOA) parameter estimation with the SAGE algorithm". In: *Signal Processing* 84.3, pp. 619–629 (cit. on p. 52).

Vesa, Sampo and Tapio Lokki (2010). "Segmentation and analysis of early reflections from a binaural room impulse response". In: *Helsinki University of Technology: Technical Report TKK-ME-RI, TKK Reports in Media Technology* (cit. on p. 50).

Vincent, Emmanuel, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian P Rosca (2007). "First stereo audio source separation evaluation campaign: data, algorithms and results". In: *International Conference on Independent Component Analysis and Signal Separation*. Springer, pp. 552–559 (cit. on pp. 96, 97).

Vincent, Emmanuel, Tuomas Virtanen, and Sharon Gannot (2018). *Audio source separation and speech enhancement*. John Wiley & Sons (cit. on pp. 30, 39).

Wallach, Hans, Edwin B Newman, and Mark R Rosenzweig (1973). "The precedence effect in sound localization (tutorial reprint)". In: *Journal of the audio engineering society* 21.10, pp. 817–826 (cit. on p. 28).

Watson, LT, JA Ford, and M Bartholomew-Biggs (2001). *Nonlinear Equations and Optimisation*. Vol. 4. Elsevier (cit. on p. 5).

Woodward, Philip M and Ian L Davies (1952). "Information theory and inverse probability in telecommunication". In: *Proceedings of the IEE-Part III: Radio and Communication Engineering* 99.58, pp. 37–44 (cit. on p. 30).

Xu, Guanghan, Hui Liu, Lang Tong, and Thomas Kailath (1995). "A least-squares approach to blind channel identification". In: *IEEE Transactions on signal processing* 43.12, pp. 2982–2993 (cit. on pp. 53, 54, 61).

Zahorik, Pavel (2002). "Direct-to-reverberant energy ratio sensitivity". In: *The Journal of the Acoustical Society of America* 112.5, pp. 2110–2117 (cit. on p. 29).

Zannini, Cecilia Maria, Albenzio Cirillo, Raffaele Parisi, and Aurelio Uncini (2010). "Improved TDOA disambiguation techniques for sound source localization in reverberant environments". In: *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, pp. 2666–2669 (cit. on p. 51).

van denBoomgaard, Rein and Rik van derWeij (2001). "Gaussian convolutions numerical approximations based on interpolation". In: *Scale-Space and Morphology in Computer Vision: Third International Conference, Scale-Space 2001 Vancouver, Canada, July 7–8, 2001 Proceedings 3*. Springer, pp. 205–214 (cit. on p. 37).