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, France
Simon DOCLO	Full professor	Carl von Ossietzky Universität, Oldenburg, Germany

Composition du Jury :

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

Abstract

Résumé en français

Acknowledgements

Contents

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
$\mathbf{x}^{\mathrm{T}}$	transpose of the vector $\mathbf{x}$
$\mathbf{x}^{\mathrm{H}}$	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

"*Only echoes answer me.*"
—Anton Chekhov, Swan Song

"*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"

▶ 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.

1.1 The Problems

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*.

1.1.1 *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

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.

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.

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.

1.1.2 *Audio scene analysis*

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.

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*

A historical example are the calculation of the Earth circunference by Eratosthenes in III century bc, and the calculations of Adams and LeVerrier which led to the discovery of Neptune from the perturbed trajectory of Uranus.

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

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

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

[3] This framework was originally proposed by Tikhonov.

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

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), Acoustic Echo Retrieval (AER) must be done.

[5] Machine Learing allows now for end2end approaches

1.3 THESIS ORGANIZATION AND RELATED CONTRIBUTION

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

- ROOM ACOUSTIC MEETS SIGNAL PROCESSING

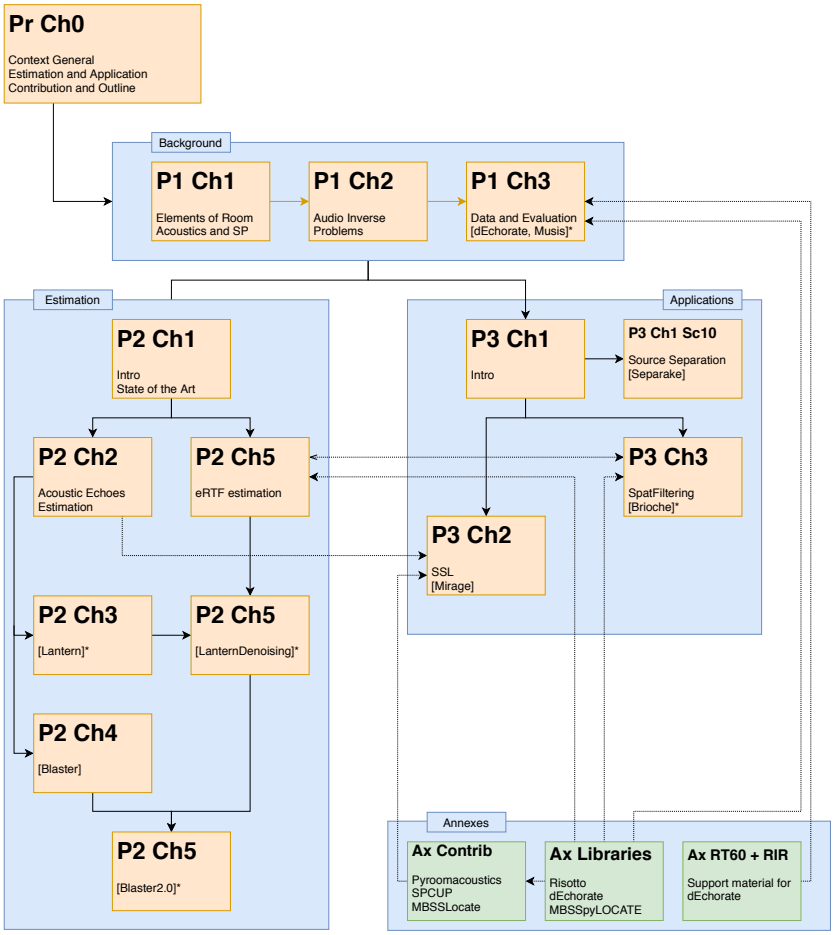


FIGURE 1.1: Schematic rganization of the thesis, dependecies between chapters linked to author contributions.

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.

▸ 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, it 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.

?? 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 (2019). "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 Dokmanic (2018). "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.

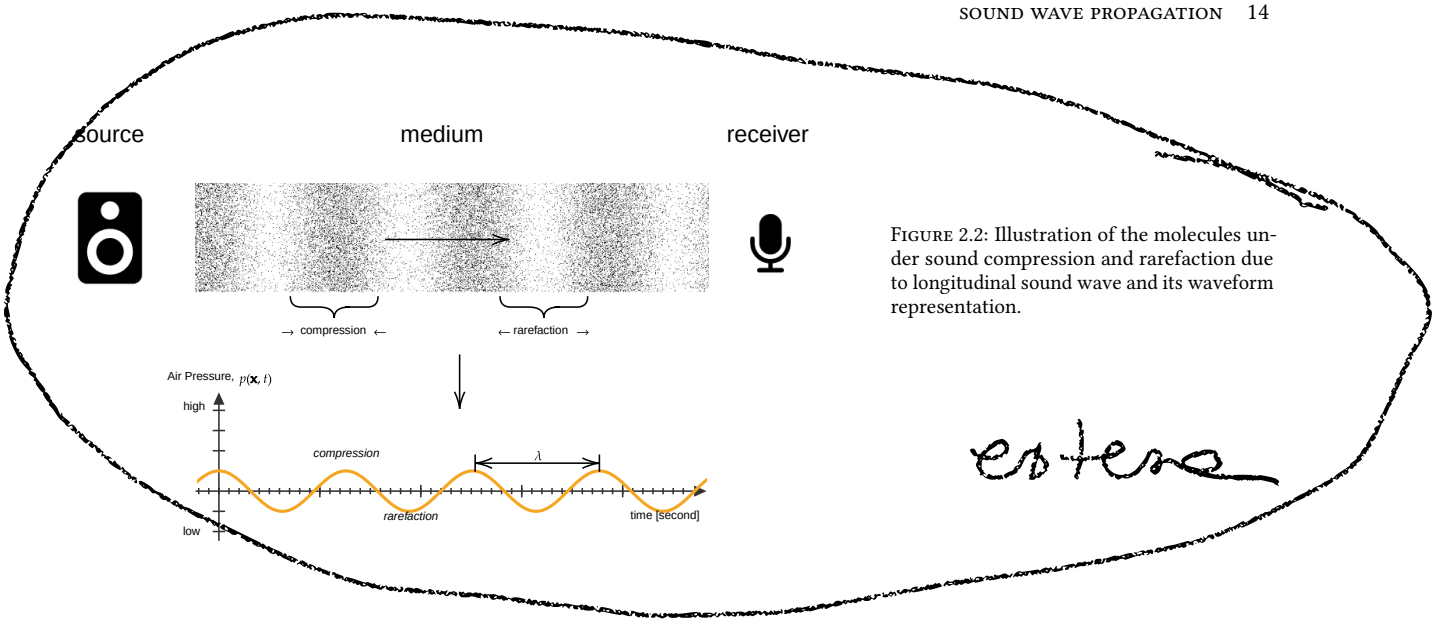


FIGURE 2.2: Illustration of the molecules under sound compression and rarefaction due to longitudinal sound wave and its waveform representation.

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

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].

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*.

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

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

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

where $\mathcal{V}_s$ denotes the source volume, and $\mathrm{d}\underline{\mathbf{s}} = \mathrm{d}x_s \, \mathrm{d}y_s \, \mathrm{d}z_s$ the differential volume element at position $\underline{\mathbf{s}}$. If one ignores the space integral, one can see

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*.

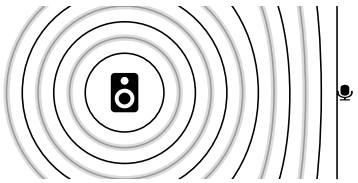


FIGURE 2.3: Visualization of the sound propagation. Since the sensor (i.e. a microphone) is drawn in the far field, the incoming waves can be approximated as plane waves.

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*

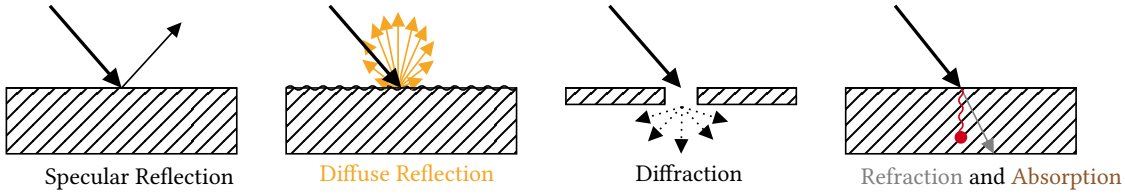


FIGURE 2.4: Different types of sound interact with a surface.

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.

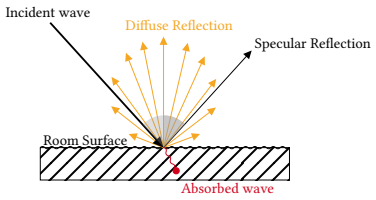


FIGURE 2.5: Specular and diffuse reflection.

" 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]

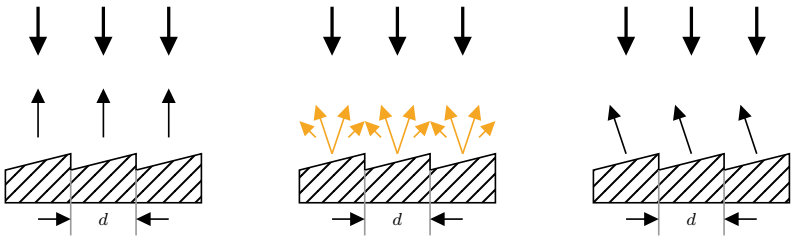
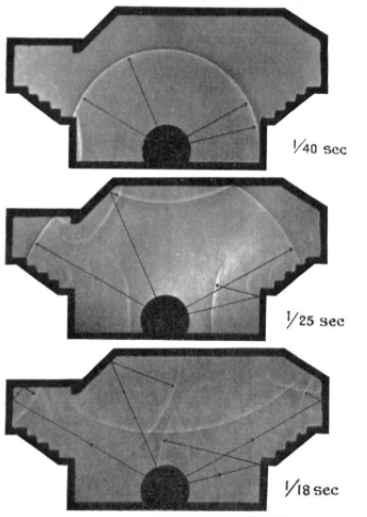


FIGURE 2.6: A reflector having irregularities on its surface with width d much smaller than the sound wavelength λ. Image courtesy of [Kuttruff 2016].

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.

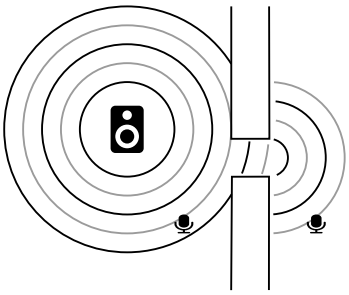


FIGURE 2.8: Schematic representation of sound diffraction. This effect allows to hear "behind walls".

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

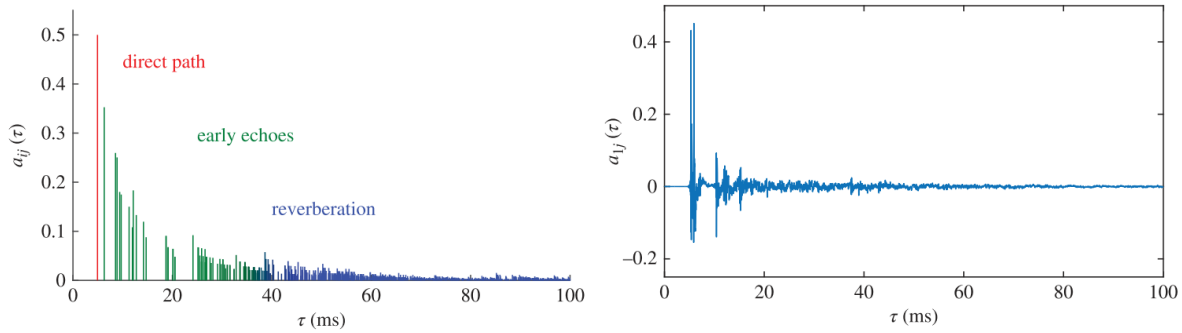


FIGURE 2.9: Schematic illustration of the shape of an RIR and the first 100 ms of a measured one.

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

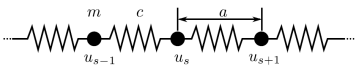


FIGURE 2.10: Example of a mass-spring linear mesh used to simulate a 1D transversal wave.

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.

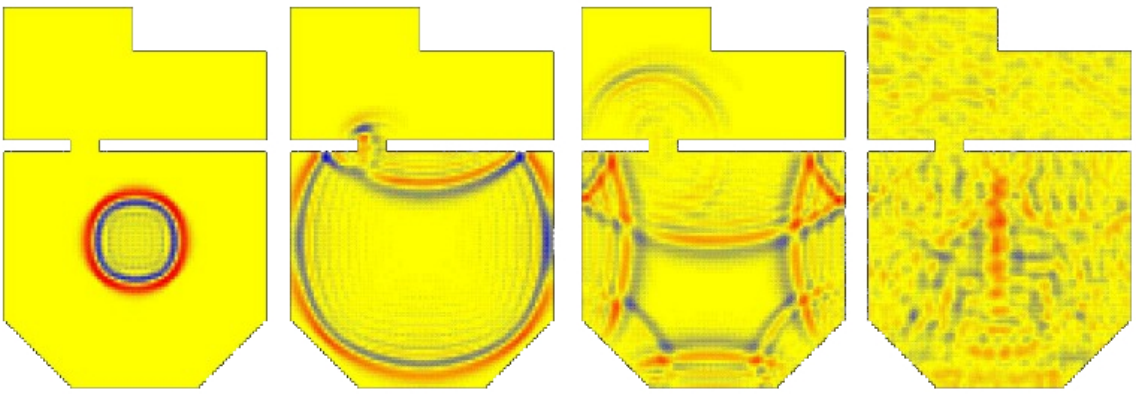


FIGURE 2.11: Simulation of Sound propagation at four consecutive timestamps using the DWM technique. A short, sharp, impulsive sound fired into the larger of two rooms causes a circular wavefront to spread out from the sound source. The the wave is reflected from the walls and part of it passes through 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.

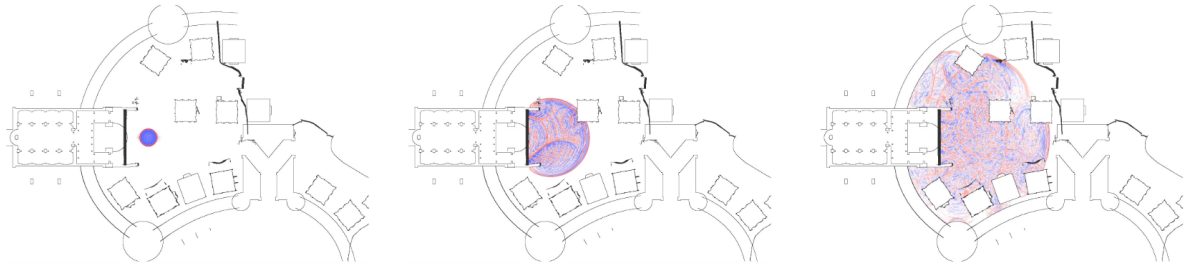


FIGURE 2.12: Sound propagation at three consecutive timestamps using the FDTD-based *Triton* simulator from Microsoft

▶ 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-

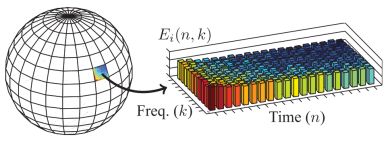


FIGURE 2.13: Directional-time-frequency Energy map resulting form the diffuse rain algorithm [Schröder et al. 2007]. For each direction, that is receiver's spherical bin, a time-frequency histogram collects the energy of incoming rays. Image courtesy of [Schimmel et al. 2009]

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.

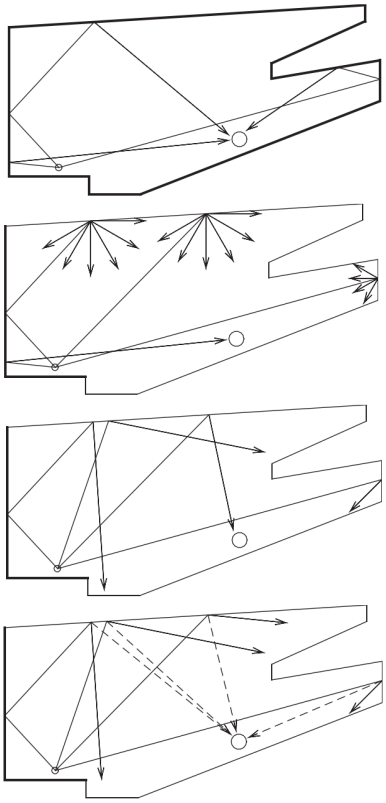


FIGURE 2.14: Visualization of ray-tracing method. From top to bottom: first the method will eventually find specular reflection; then diffuse reflections can be modeled either by splitting a ray into several new rays or a single random one. In the diffuse rain technique a shadow-ray is cast from each 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.

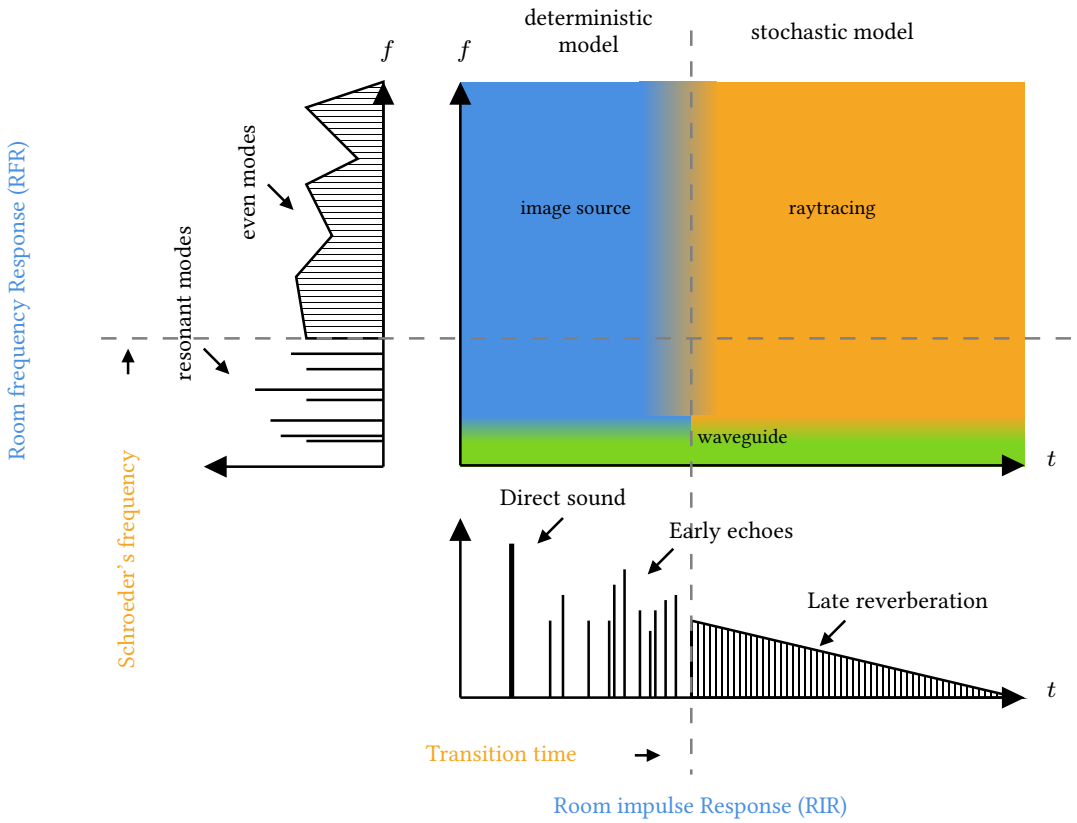


FIGURE 2.15: Time-Frequency regions of the RIR associated to the method that better simulate them. Image adapted from [Thomas 2017; Badeau 2019].

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.

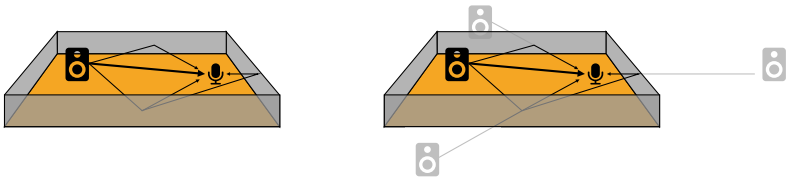


FIGURE 2.16: 3-D representation of the Image Source Method (ISM) and its propagation paths for selected echoes.

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.

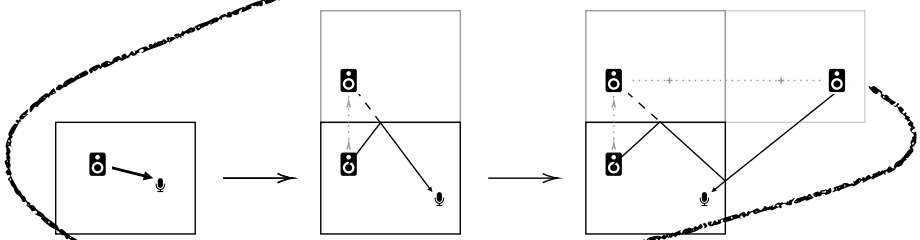


FIGURE 2.17: From left to right, path involving the direct path, one reflection obtained using first-order image, and two reflections obtained using two images. Image inspired from [Habets 2006].

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 **??**, 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.

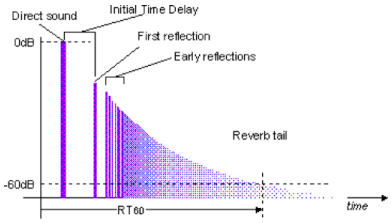


FIGURE 2.18: illustration of the Reverberation Time ($\mathtt{RT}_{60}$) definition. It. Image courtesy of wikipedia.

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.

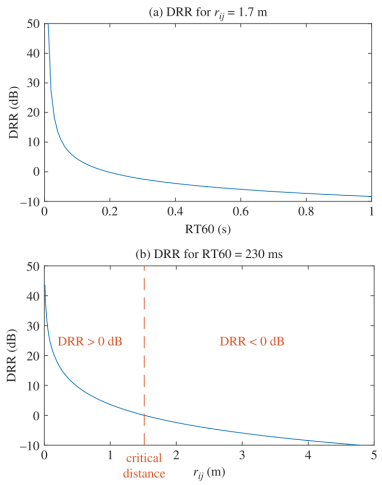


FIGURE 2.19: DRR as a function of the RT60 and the source distance rij based on Eyring's formula (Gustafsson et al., 2003). These curves assume that there is no obstacle between the source and the microphone, so that the direct path exists. The room dimensions are the same as in Figure 3.1.

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.

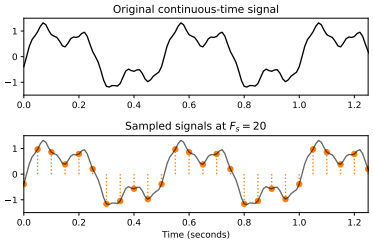


Figure 3.1: Continuous-time signal and its sampled version.

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 $\tilde{\boldsymbol{x}}(t) = [\tilde{x}_1, \ldots, \tilde{x}_I]^{\mathrm{T}} \in \mathbb{R}^I$.

⇄ 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

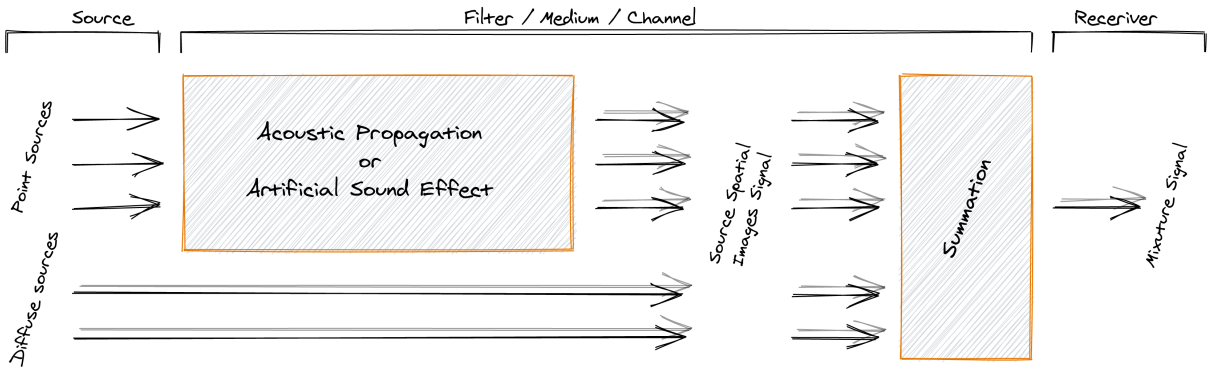


FIGURE 3.2: General mixing process, illustrated in the case of $J = 3$ sources, including three point sources and one diffuse source, and $I = 2$ channels.

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}.

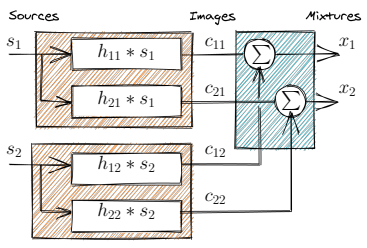


FIGURE 3.3: Graphical representation of the mixing model 3.5 for 2 sources and 2 microphones.

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

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

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

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:

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

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.

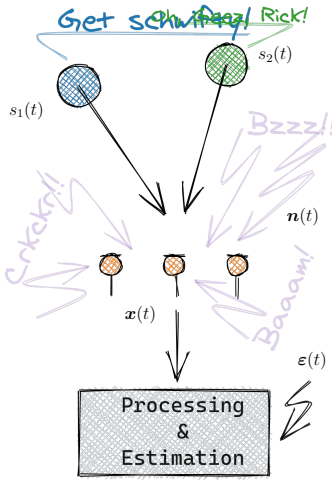


FIGURE 3.4: Graphical representation of the mixing model (3.5): $s_2(t)$ is the *interferer*, $\boldsymbol{n}(t)$ contributes to the *diffuse noise field*, and $\boldsymbol{\varepsilon}(t)$ model acquisition and modeling errors.

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

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

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

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

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.

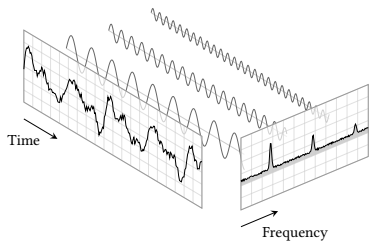


FIGURE 3.5: A signals resolved into its Fourier series: a linear combination of sines and cosines represented as peaks in the frequency domain.

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, 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 room transfer function of Eq. (3.27). Thus we can write,

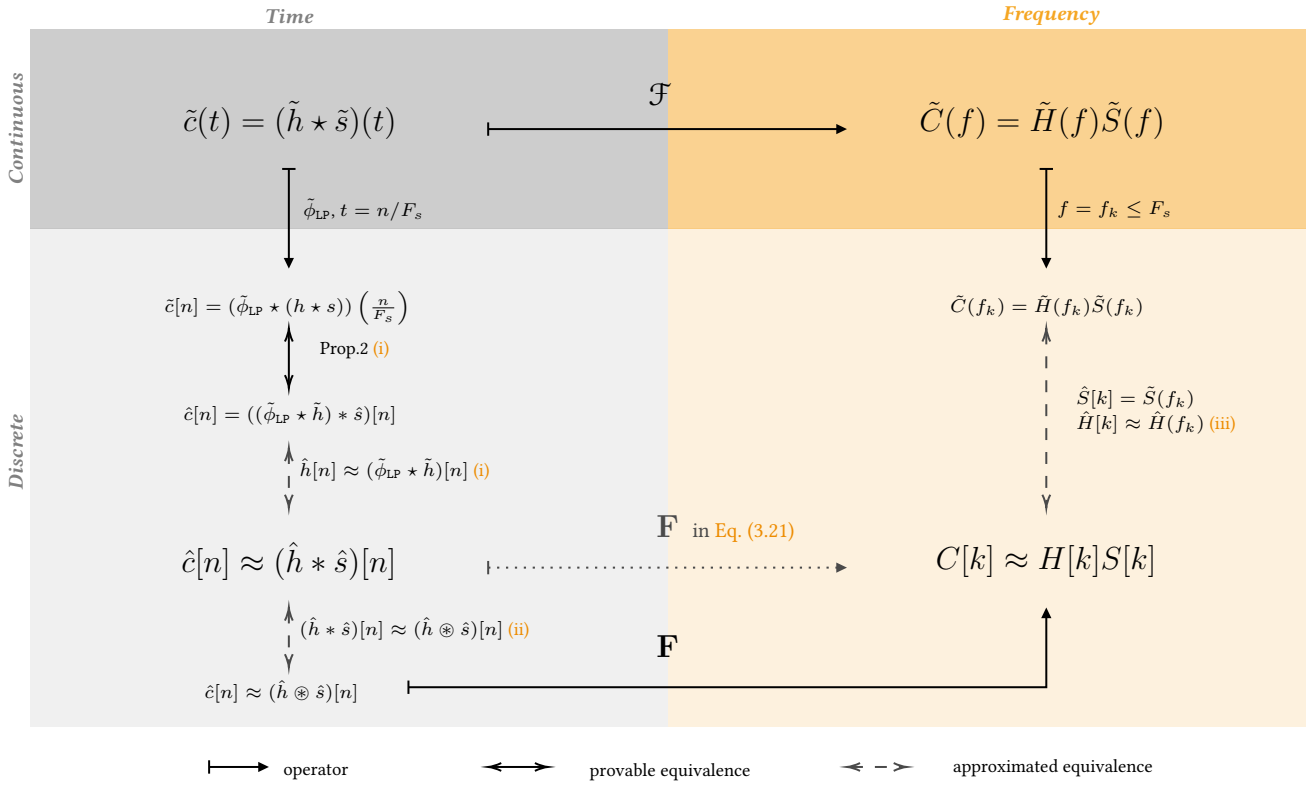
[23] The finite-time linear convolution for two vectors $\hat{u} \in \mathbb{R}^L$ and $\hat{v} \in \mathbb{R}^D$ is $(\hat{u} * \hat{v})[n] = \sum_{l=0}^{L-1} \hat{u}[l]\hat{v}[L-1+n-j]$ for $n = 0, \cdots, D-L$.

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

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 $\hat{h}[n] \approx (\tilde{\phi}_{\text{LP}} * \tilde{h})[n]$, 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

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

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 **??**.

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

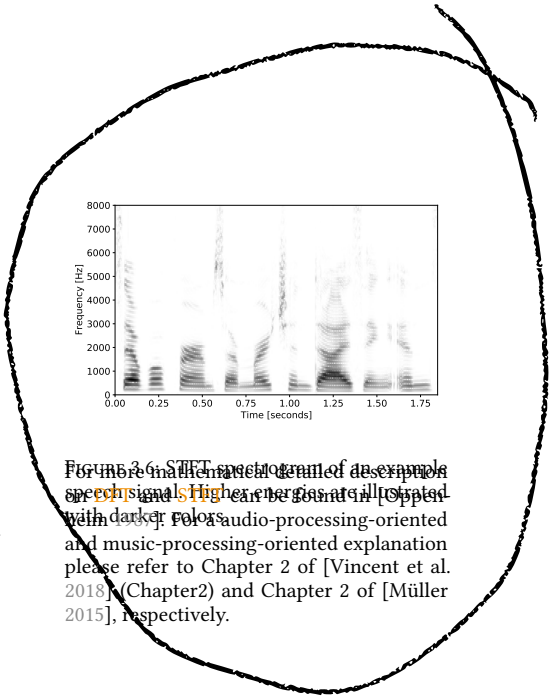
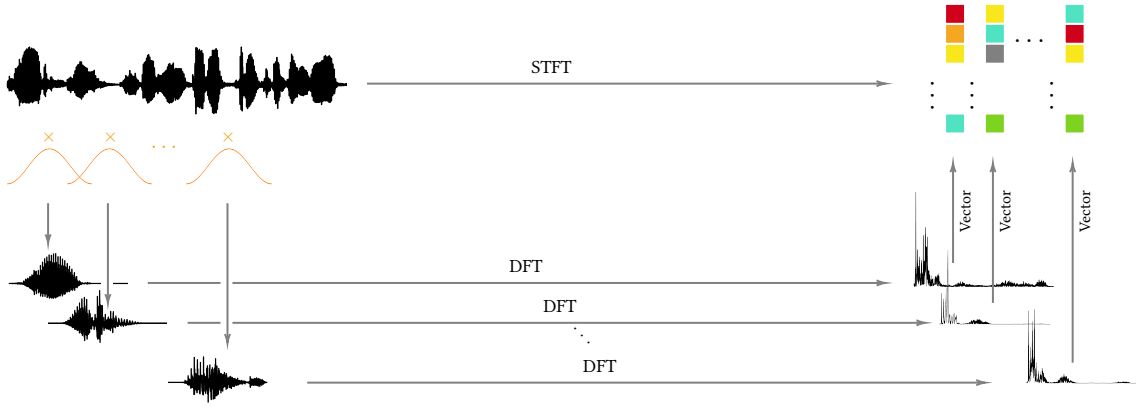


FIGURE 3.6: STFT spectrogram of an example speech signal. Higher energies are illustrated with darker colors.

For more mathematical detailed description on DFT and STFT can be found in [Oppenheim 1989]. For a 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 is simple to invert;
- 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 $\hat{\boldsymbol{H}}_j[k] = \left[\hat{h}_{1j}[k], \cdots, \hat{h}_{Ij}[k]\right]^{\mathrm{T}}$ is the $I \times 1$ vector of the room transfer functions 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 $\boldsymbol{U}[l,k] = \boldsymbol{N}[l,k] + \boldsymbol{\varepsilon}(l,k)$ 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 (RTF) 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 RTF needs not be so.

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

- the RTF associated to the reference channel ($i = 1$) is equal to 1 for each frequency bin k.
- The problem of estimating the RTF 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 RTFs 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.

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

- A RIR can be seen as a special case of RTF 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].
- As discussed below, RTFs simplify to special steering vectors in free- and far-field conditions, which have interesting geometrical properties.

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

In the general case of multiple microphone arrays ($I > 2$) and multiple sources, the vector of RTFs $\boldsymbol{G}_j[k] = \left[\hat{G}_{1j}[k], \cdots, \hat{G}_{Ij}[k]\right]^{\mathrm{T}}$ 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 **??** as they are fundamental quantities for sound source localization.

▶ IN THE CONTEXT OF SPATIAL AUDITORY PERCEPTION and Computational Auditory Scene Analysis (CASA), the RTF is related to the *interchannel cues*[28]. In fact, the RTFs 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.

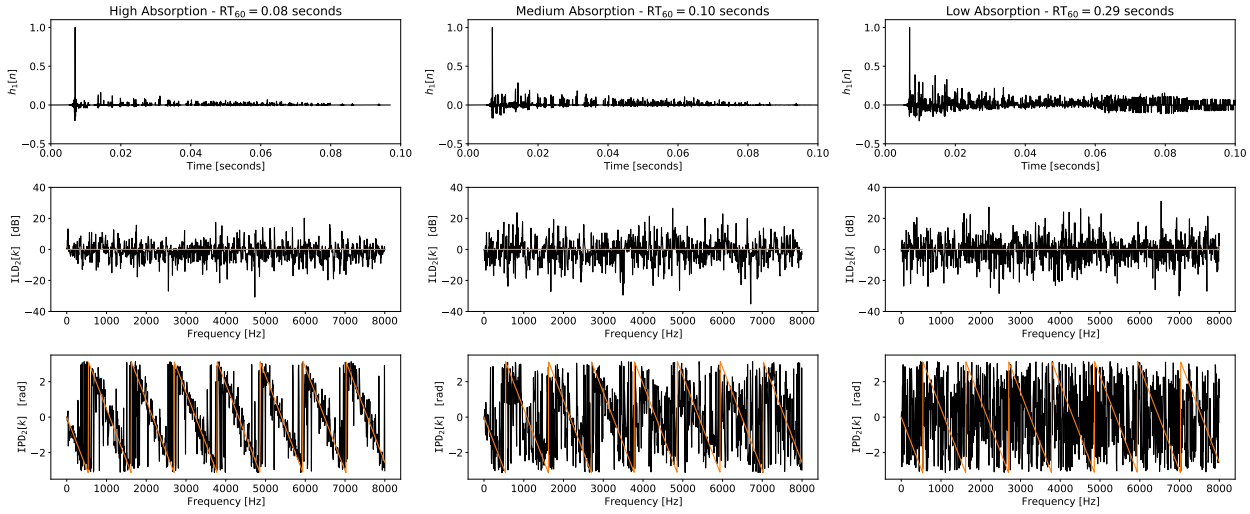


FIGURE 3.7: RIR, ILD and IPD corresponding to the pair of synthetic impulse responses of Figure 2.9 for different absorption conditions. Orange lines denote the theoretical far- and free- field ILD and IPD as defined by the relative steering vectors of Eq. (3.30)

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, it 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 (a. k. a. delays, Time of Arrival (TOA) or location) $\{\tau_i^{(r)}\}_{i,r}$ and attenuations (or gains) $\{\alpha_i^{(r)}\}_{i,r}$ of Eq. (4.1).
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 a instance of *acoustic channel estimation* and *shaping*, and *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 when sound 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 work assumes active scenarios, namely, they use one or more loudspeaker to probe the environment and one or more microphones to record the propagated probe sound. Therefore, the assume that the source reference signal is known. They falls into the big categories of *deconvolution problem* since a "clean" reference signal is used to *deconvolve* the observed one. Two are the main advantages of these approaches. First, with proper probe signal, a good estimation of the RIR can be achieved. Second, this methods can be used on single-channel recordings.
Instead, *passive* approaches uses passive sensors to record the sound field. To decouple environment from source signal, they rely either on prior knowledge on 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 and a great deal of efforts have been devoted to the these problems and it is still active research topic in signal processing, notably due to its fundamental ill-posedness. Moreover, other obvious advantage is that these approach are non-intrusive since only already existing sounds are used in the estimation.

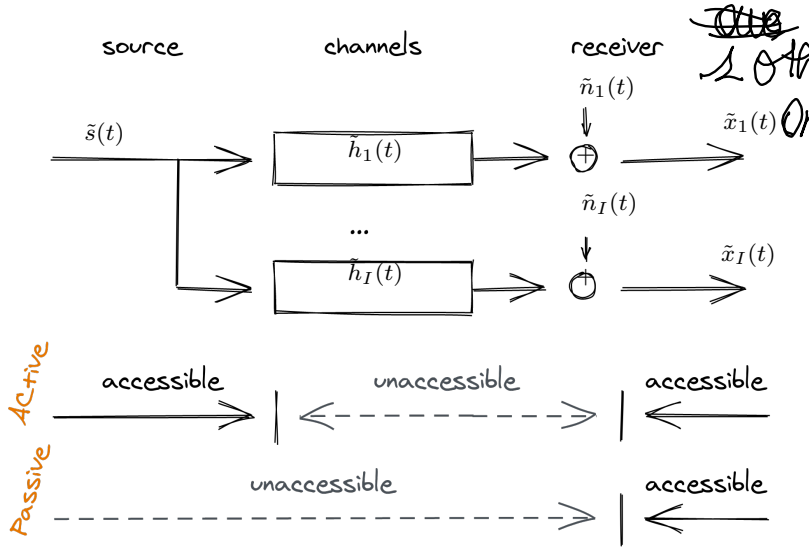


FIGURE 4.1: Schematic of active and passive approaches.

⇄ 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 and echo extraction in their early part. 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*. Other methods estimate the RIRs partially using assumptions derived by the application. It is the case of *impulse response shaping* or *shortening*. In context of room acoustics, they aim 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 estimate echo properties directly the parameter space of echos' TOAs and amplitudes.

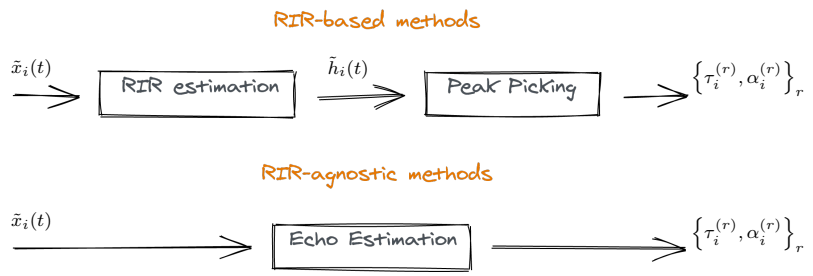


FIGURE 4.2: Schematic of RIR-based and RIR-agnostic approaches.

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 depends on the type of transmitted signal. When the transmitted signal is arbitrary, several methods were developed to measure real RIRs and can be found the field of acoustic measurements. 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 speaker and microphone and computing the 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. The last property allows the user not to 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 technique cannot be used. In this scenario, 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 case of 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 [**boooh**].

▸ Echo Retrieval from RIR

As discussed through 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, so, to better visualize them, the *echogram* [Kuttruff 2016], namely $h = |h|$, or the energy envelop [Schroeder 1979] are used instead[29]. 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 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 way 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 [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 techniques used in music onset detection and music transcription, using edge-detection wavelet filters [Bello et al. 2005], identifying attach-decay patterns [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 identified 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 identified first prominent reflection through deconvolution was used in [Annibale et al. 2012]. This technique is also referred 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, recent works [Ristić et al. 2013; Pavlović et al. 2016] use (multi-)fractal analysis to detect echoes in the TF domain. Alternatively, two recent works operate in the cepstral domain [Ferguson et al. 2019; Jia et al. 2017]. 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 and they were in the early stage of signal processing for source-filter identification. 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 coupe with the pathological issues of this approach, off-grid framework can be used, e. g. [Condat and Hirabayashi 2013]. This approach can be related to other classical Maximum Likelihood (ML) estimation problem, which consist in selecting the model which is most likely to explain the observed noisy data. In this category fall classical spectral estimation technique, 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. This problem can be reformulated as a *matrix denoising* problem, from which the echoes location and amplitudes can be retrieved in closed-from. Although this method reach statistical optimality in 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 on clean RIR estimate. 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 SNR 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 this issues is to overestimate the echoes in the RIR by including some false-positive observation and further prune them using echo labeling.

- TOAs Disambiguation or *Echo labeling* is the task of assigning acoustic echoes to different image sources or reflectors. Many methods are 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 this works focuses on actively localizing (angle of arrival of) multiple sources while discarding reflection, rather than localizing the actual image sources.
  The other disambiguation schemes are typically used in 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 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 distance between each microphone are known with precision, which may not be available in practice.
  In the works [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 non-convex optimization methods based on geometrical reasoning and the Hough transform [30], they are able to disambiguate TOAs and reconstruct reflectors position and inclination. However, they require a very specific acquisition setup and use the optimization which are sensible to local minima.
  In general, all the above methods do not have specific strategies to cope with missing or spurious estimate of echoes given by malfunctioning of the peak finder or by selection of peaks corresponding to high order reflections and in some specific case manual annotation is used.
  Another interest approach is presented in [El Baba et al. 2017], which exploits

[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. This tranformor is typically used in computer vision as feature extractor to detect lines and edges in pictures.

the geometrical 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 draws a specific pattern which can be identified and labeled more easily. This approach avoid the combinatorial search, but still require specific setup for measuring the RIRs.

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 microphone and source positions. Then, the whole set of extracted peaks are used to estimate the planar reflectors positions which are then used refine the microphone and source localization. Alternating between physical/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, hence, avoiding the RIR estimation and peak finding steps. Here two approach can be identified: ML-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 microphone are used and their relative position is known, the relation between the two quantities can be express in closed-form. The DOAs can be used to reduce the ambiguity of the estimated echoes. This method extends a class of existing methods used in multipath communication system, denoted as Joint Angle and Delay Estimation (JADE) [Vanderveen et al. 1997; Verhaevert et al. 2004].

Alternatively, the echoes contribution can be extracted 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 is used in [Crocco et al. 2014; Al-Karawi and Mohammed 2019].

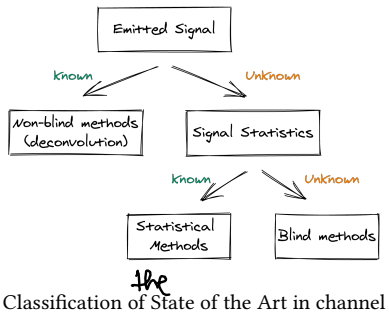
When the array geometry is known, the time lag axes of the cross-correlation function for all the microphones can be mapped to possible 2D direct of arrivals (elevation and azimuth), namely from TOAs to 2D-DOAs. The identification of strong reflection can be then conducted the so-called angular spectrum domain, which consider sound energy as function possible DOAs [DiBiase et al. 2001]. With proper clustering approach, the reflections can be identified, disambiguated and their TOAs computed. This approach is used in [O'Donovan et al. 2008; O'Donovan et al. 2010] and SSL [DiBiase et al. 2001] and it is called *audio camera*. The 2D-polar coordinates can mapped into cartesian ones and the angular spectrum can be superimposed to a panoramic picture of the audio scene taken by the barycenter of the recording arrays. For detailed review for robust cross-correlation methods [Chen et al. 2006].

Cross-correlation and convolution operation are very similar, an mathematically they differ just by the inversion of the 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.

4.3.3 *Passive and RIR-based method*

Passive approaches rely on using 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)). In the particular case of a single source, it is referred to as SIMO Blind Channel Estimation (BCE). Common to all this method is the assumption that RIR 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 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].



Classification of 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 source probability function. Even though they are optimal in ML sense, they optimize non-convex cost functions, typically via Expectation Maximization (EM). In this categories 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. 2018]. This methods are built on the well-studied framework of Multichannel Nonnegative Matrix Factorization (NMF) [Ozerov and Févotte 2009] which lend itself to account for various type of side information. Here the source signals are typically modeled as Gaussian distribution centred in zero and unknown variances. Using pre-trained dictionaries for modeling the variances of the source, they are able to estimated both the acoustic channels and the source contribution. In particular the work [Duong et al. 2010] extends this framework for reverberant recordings using physic-based models for the late reverberations, while [Leglaive et al. 2016] consider explicitly the contribution of early echoes, further improved in [Leglaive et al. 2018].

Even if statistical methods has reported 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 the final estimated RIRs may match the real one in the statistical sense, they lacks of proper definition, 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 2016].

  The former is based on the key idea that the channel (or part of it) vector spans a one-dimensional subspace of (block of of) noiseless observations. This methods have the attractive property that the channel estimates can often be obtained in a closed-form from optimizing a quadratic cost function. However they relay on the they may not be 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 condition 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 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 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. 2010; Dokmanić et al. 2015] and room geometry [Antonacci et al. 2012; Crocco et al. 2017] estimation have proven the effectiveness of this approach. On a similar scheme, in [Kowalczyk et al. 2013], (??) 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 Crocco *et al.* in [Crocco and Del Bue 2015; Crocco and Del Bue 2016], where the anchor constraint is replaced by an *iterative weighted* ℓ_1 equality ~~constraint.~~ to better balance sparsity penalty and the model constraints. These approaches will be further formalized and detailed in ??. 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 application.

  An alternative approach is used in [Čmejla et al. 2019], where the RIR estimation problem is treated as special case of RTF estimation ~~problem~~. As mentioned in § 3.3.2, in 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]. Methods for RTF estimation will be detailed in ??. 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 scenario where the user have the possibility to place 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 allows to solve geometrical ambiguities, yielding to closed-form solution, as done in [Crocco et al. 2012].

In general, the main drawbacks of FIR SIMO BCE works is that they rely on on-grid estimation and 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 ~~strictly speaking~~ non-sparse and non-negative ~~in practice~~. 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. Since these methods are fundamentally on-grid, the estimated echo locations integer multiple 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 issue may arise, affecting the estimation. Nevertheless, despite this slight mismatch between theoretical assumptions and real data, for some scenario the position of the estimated peaks by the methods [Crocco and Del Bue 2016] reproduces the positions of the ground truth peaks with remarkable precision as demonstrated 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 method*

Methods in this categories bypass the onerous task of estimating the (full or partial) acoustic channel and, to the best ~~author~~ knowledge, only a few have been identified. As for the active and RIR-agnostic case, the audio cameras 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. ~~This~~ 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 method needs to be used to convert DOAs ~~them~~ back to echoes timing.

Recently a fully blind, passive, off-grid and RIR-agnostic method was proposed by Tukuljac et al. for stereophonic recordings, namely using only 2 microphones. They proposed a method, called Multichannel Annihilation (MULAN),

based on the properties of the *annihilation filter*[31], the work of [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 in 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 [Peic Tukuljac 2020] using the generalization of Cadzow denoising framework [Condat and Hirabayashi 2015]. This method is show to outperform conventional methods 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 is not being tested on challenging real scenario featuring external noise and partial knowledge on the number of echos

[31] For a sequence of Fourier coefficients (describing a signal or a filter), its annihilation filter is such that the linear convolution between the sequence and the filter coefficients is identically zero.

4.4 Data and Evaluation

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

4.4.1 *Datasets*

As listed in [Szöke et al. 2019] and in [Genovese et al. 2019], a number of recorded RIRs corpora are available online and for free, 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 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 in audio signal processing. While simulated datasets are more versatile, simple and quicker to obtain, they fail to fully capture the complexity and the richness of real acoustic environments. Due to this, methods trained or validated on them may fail to generalize to real conditions, as will be shown in ??.

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

Database Name	Annotated Pos.	Annotated Echoes	Annotated Rooms	Number of RIRs	Number of Rooms	Number of Mic×Pos.	Number of Src	Key characteristics	Purpose
[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 ??	✓	✓	✓	∼1.8k	1	30	6	Accurate annotation Different Echo-energy	RooGE SE/ASR

TABLE 4.1: Comparison between some existing RIR databases that 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 read is invited to refer to [Szöke et al. 2019; Genovese et al. 2019] for 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.

include acoustic echoes as highly correlated interfering source coming from close reflectors, such as 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. The Table ?? 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 address as FIR SIMO BCE problem, the ground-truth acoustic channels are considered as a discrete vector $h \in \mathbb{R}^L$, 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:

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

  where $|\cdot|$ denotes the absolute value. This metrics is known to be highly sensible to scaling and little translation. For instance, if the $\hat{h}^*$ is a just a shifted and scaled version of the $\hat{h}$, huge RMSE is recorded.

- *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:

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

  where $\|\cdot\|_2$ denotes the Euclidean norm. By projecting $\hat{h}$ onto h and defining a projection error, only the intrinsic misalignment of the chan-

nel 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]

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

  As the NPM, this metrics is invariant to possible scaling factors and length difference 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 or 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

$$\mathtt{RMSE}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) \stackrel{\text{def}}{=} \sqrt{\sum_{r=0}^{R} |\hat{\tau}_r - \tau_r|^2} \quad \text{second/samples}, \tag{4.6}$$

  This metric describes the mean error between estimated and reference of echoes' TOAs. Unfortunately, the RMSE is proportional to the size of 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 an 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. Then, the

F-measure, defined as the harmonic mean of precision and recall, is used to summarizes the two values with one value. Depending on the application, precision and recall can have different impact. RooGE methods are more sensible to missing TOAs than their misalignment which can be redefined with geometrical reasoning. Thus they are more incline to prefer recall on precision and sometimes allow for some false-positive which can be pruned using the echo labelling methods. Instead, echo-aware SE methods prefer to accurately select the relevant echoes, thus favoring 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 one needs to be labeled, pruned and matched. As discussed at the end of § 2.3.3, echoes can be sorted differently according their amplitudes, their TOAs or the generation of image-source in the ISM model. 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 evaluation process which is currently unsolved without proper echo labeling step.

Part IV

ECHO-AWARE APPLICATION

BIBLIOGRAPHY

[toc]

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. 52).

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. 55).

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 p. 52).

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. 50).

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. 23–25).

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. 47, 48).

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. 49).

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. 49, 52).

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. 46).

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. 22, 24).

Bal, Guillaume (2012). "Introduction to inverse problems". In: *Lecture Notes-Department of Applied Physics and Applied Mathematics, Columbia University, New York* (cit. on p. 5).

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. 27).

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. 47, 48).

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 pp. 54, 55).

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. 46).

Böck, Sebastian, Florian Krebs, and Markus Schedl (2012). "Evaluating the Online Capabilities of Onset Detection Methods." In: *ISMIR*, pp. 49–54 (cit. on p. 56).

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. 50).

Cheng, Tian, Matthias Mauch, Emmanouil Benetos, Simon Dixon, et al. (2016). "An attack/decay model for piano transcription". In: ISMIR (cit. on pp. 47, 48).

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. 53).

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. 52, 54).

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. 48, 54).

— (2015). "Cadzow denoising upgraded: A new projection method for the recovery of Dirac pulses from noisy linear measurements". In: (cit. on pp. 44, 54).

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 p. 52).

— (2016). "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. 52, 53).

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. 53).

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. 49, 50).

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. 47, 49, 52–55).

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. 18).

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. 47, 48).

— (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. 47, 53).

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. 9).

Di Carlo, Diego, Antoine Deleforge, and Nancy Bertin (2019). "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 p. 9).

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. 9, 53).

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. 9).

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. 50, 53).

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. 49, 54, 55).

Dokmanić, Ivan, Robin Scheibler, and Martin Vetterli (2015). "Raking the cocktail party". In: *IEEE journal of selected topics in signal processing* 9.5, pp. 825–836 (cit. on p. 52).

Duffy, Dean G (2015). *Green's functions with applications*. CRC Press (cit. on pp. 13, 15).

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. 46).

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 p. 51).

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 p. 49).

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. 47).

— (2007). "Advancements in impulse response measurements by sine sweeps". In: *Audio Engineering Society Convention 122*. Audio Engineering Society (cit. on p. 47).

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. 48).

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. 49).

Fourier, Jean Baptiste Joseph (1822). *Théorie analytique de la chaleur*. F. Didot (cit. on p. 33).

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. 40, 52).

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. 54, 55).

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. 39).

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. 27).

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. 48).

Habets, Emanuel AP (2006). "Room impulse response generator". In: *Technische Universiteit Eindhoven, Tech. Rep* 2.2.4, p. 1 (cit. on p. 25).

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. 32).

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. 23).

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. 51, 55).

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. 49).

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 p. 50).

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. 48).

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. 27).

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. 47, 48).

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. 52).

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. 52).

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. 52).

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 p. 52).

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. 18).

Kulowski, Andrzej (1985). "Algorithmic representation of the ray tracing technique". In: *Applied Acoustics* 18.6, pp. 449–469 (cit. on p. 22).

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. 47).

Kuttruff, Heinrich (2016). *Room acoustics*. CRC Press (cit. on pp. 10, 13, 16–18, 47).

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).

Leglaive, Simon, Roland Badeau, and Gaël Richard (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 p. 51).

— (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 p. 51).

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. 47).

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 p. 52).

— (2008). "Blind channel identification for speech dereverberation using l1-norm sparse learning". In: *Advances in Neural Information Processing Systems*, pp. 921–928 (cit. on p. 52).

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. 48).

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. 55).

Müller, Meinard (2015). *Fundamentals of Music Processing*. Springer Verlag. ISBN: 978-3-319-21944-8 (cit. on p. 38).

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. 47).

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. 50, 53).

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. 50, 53).

Oppenheim, Alan V (1987). *Signals and Systems: An Introduction to Analog and Digital Signal Processing*. MIT Center for Advanced Engineering Study (cit. on p. 38).

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. 51).

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. 49).

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. 48).

Peic Tukuljac, Helena (2020). *Sparse and Parametric Modeling with Applications to Acoustics and Audio*. Tech. rep. EPFL (cit. on p. 54).

Pierce, Allan D (2019). *Acoustics: an introduction to its physical principles and applications*. Springer (cit. on pp. 13, 18).

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. 52).

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. 47, 54, 55).

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. 55).

Ribeiro, Flávio, Demba Ba, Cha Zhang, and Dinei Florêncio (2010). "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 p. 52).

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. 48).

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. 48).

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. 49).

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. 50).

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. 51).

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. 17, 18, 22, 23, 28).

Scheibler, Robin, Diego Di Carlo, Antoine Deleforge, and Ivan Dokmanic (2018). "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, 51).

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. 49).

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 p. 22).

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. 22, 23).

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. 46, 47).

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. 31).

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. 47, 54, 55).

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. 56).

Thomas, Matthew Reuben (2017). "Wayverb: A Graphical Tool for Hybrid Room Acoustics Simulation". PhD thesis. University of Huddersfield (cit. on pp. 21, 24).

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. 51, 52).

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 p. 52).

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. 36, 47, 53).

Tuzlukov, Vyacheslav (2018). *Signal processing noise*. CRC Press (cit. on p. 32).

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. 47, 48).

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. 21, 27).

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. 50).

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. 56).

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. 49).

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. 50).

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. 48).

Vincent, Emmanuel, Tuomas Virtanen, and Sharon Gannot (2018). *Audio source separation and speech enhancement*. John Wiley & Sons (cit. on pp. 29, 38).

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. 27).

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. 29).

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. 51, 52).

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. 28).

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. 49).

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. 36).