Hunting Echoes for Auditory Scene Analysis

Hunting Echoes for Auditory Scene Analysis

Dissertation Thesis

Diego Di Carlo

August 21, 2020

Submitted in partial fulfillment of the requirements
for the degree of Doktor der Naturwissenschaften

to the

Faculty of Mathematics
at Ruhr-Universität Bochum

1st Reviewer Prof. Dr. Gregor Leander
2nd Reviewer Prof. Dr. Alexander May

Imprint

Hunting Echoes for Audtory Scene Analysis

Published by the Ruhr-Universität Bochum, Bochum, Germany.

Colophon

This thesis was typeset using LaTeX and the `memoir` documentclass. It is based on Aaron Turon's thesis *Understanding and expressing scalable concurrency*[1], itself a mixture of `classicthesis`[2] by André Miede and `tufte-latex`[3], based on Edward Tufte's *Beautiful Evidence*.

The bibliography was processed by Biblatex. All graphics and plots are made with PGF/Ti*k*Z.

The body text is set 10/14pt (long primer) on a 26pc measure. The margin text is set 8/9pt (brevier) on a 12pc measure. Matthew Carter's Charter acts as both the text and display typeface. Monospaced text uses Jim Lyles's `Bitstream Vera Mono` ("Bera Mono").

[1] `https://people.mpi-sws.org/~turon/turon-thesis.pdf`

[2] `https://bitbucket.org/amiede/classicthesis/`

[3] `https://github.com/Tufte-LaTeX/tufte-latex`

Pleasure to me is wonder—the unexplored, the unexpected,
the thing that is hidden and the changeless thing
that lurks behind superficial mutability.
— Howard Phillips Lovercraft

Abstract

Block ciphers form, without doubt, the backbone of today's encrypted communication and are thus justifiably the workhorses of cryptography. While efficiency of modern designs improved ever since the development of the DES and AES, the case with the corresponding security arguments differs. The thesis at hand aims at two main points, both in the direction of improving security analysis of block ciphers.

Part I studies a new notion for the better understanding of a special type of cryptanalysis and proposes a new block cipher instance. This instance comes with a tight bound on any differential, to the best of our knowledge the first such block cipher.

Part II turns to automated methods in design and analysis of block ciphers. Our main contribution here is an algorithm to propagate subspaces through encryption rounds, together with two applications: an algorithmic security argument against a new type of cryptanalysis and an idea towards the automation of key recovery attacks.

Résumé en français

Block ciphers form, without doubt, the backbone of today's encrypted communication and are thus justifiably the workhorses of cryptography. While efficiency of modern designs improved ever since the development of the DES and AES, the case with the corresponding security arguments differs. The thesis at hand aims at two main points, both in the direction of improving security analysis of block ciphers.

Part I studies a new notion for the better understanding of a special type of cryptanalysis and proposes a new block cipher instance. This instance comes with a tight bound on any differential, to the best of our knowledge the first such block cipher.

Part II turns to automated methods in design and analysis of block ciphers. Our main contribution here is an algorithm to propagate subspaces through encryption rounds, together with two applications: an algorithmic security argument against a new type of cryptanalysis and an idea towards the automation of key recovery attacks.

Acknowledgements

Block ciphers form, without doubt, the backbone of today's encrypted communication and are thus justifiably the workhorses of cryptography. While efficiency of modern designs improved ever since the development of the DES and AES, the case with the corresponding security arguments differs. The thesis at hand aims at two main points, both in the direction of improving security analysis of block ciphers.

Part I studies a new notion for the better understanding of a special type of cryptanalysis and proposes a new block cipher instance. This instance comes with a tight bound on any differential, to the best of our knowledge the first such block cipher.

Part II turns to automated methods in design and analysis of block ciphers. Our main contribution here is an algorithm to propagate subspaces through encryption rounds, together with two applications: an algorithmic security argument against a new type of cryptanalysis and an idea towards the automation of key recovery attacks.

Contents

Glossary

*

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".

CASA Computational Auditory Scene Analysis 38

SOTA State of the Art 21

GA Geometrical (room) acoustics 18

FEM Finite Element Method 21

BEM Boundary Element Method 21

FDTD Finite-Difference-Time-Domain 21

DWM Digital Waveguide Mesh 21

ISM Image Source Method 20

RIR Room Impulse Response xiii

ReIR Relative Impulse Response 38

ATF Acoustic Transfer Function 19

AIR Acoustic Impulse Response 19

TF Time-Frequency 23

SE Speech Enhancement 5

SSL Sound Source Localization 5

RooGE Room Geometry Estimation 5

AER Acoustic Echo Retrieval 5

FT Fourier Transform 33

DFT Discrete Fourier Transform 33

DTFT Discrete-Time Fourier Transform 35

STFT Short Time Fourier Transform 35

FFT Fast Fourier Transform 36

RTF Relative Transfer Function 37

ILD Interchannel Level Difference 38

IPD Interchannel Phase Difference 38

TDOA Time Difference of Arrival 38

AWGN Additive White Gaussian Noise 32

AER Acoustic Echo Retrieval 5

Notations

Linear Algebra

x	scalar
$\mathbf{x}$	vector
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 ($\mathbf{x}$)
$\imath$	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 $x_i(t)$
$\mathbf{x}_i$	3D position of the microphone i recording $x_i(t)$
$\mathbf{s}_j$	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

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

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"

▶ ASDLorem 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.

1.1 The Problem

In the context of audio signal processing, algorithms can be grouped according to how they deal with sound propagation.

1.2 My Thesis

The goal of this dissertation is to improve the above state of affairs along two axes: First, by deepening our understanding of sophisticated scalable algorithms, isolating their essence and thus reducing the barrier to building new ones. Second, by developing new ways ofexpressing scalable algorithms that are abstract, declarative, and user-extensible. To that end, the dissertation demonstrates two claims:

- Scalable algorithms can be understood through linked protocols governing each part of their state, which enables verification that is local in space, time, and thread execution.

- Scalable algorithms can be expressed through a mixture of shared-state and message-passing combinators, which enables extension by clients without imposing prohibitive overhead.

We elaborate on each claim in turn.

1.2.1 *Audio Signal Processing*

- Motivation
- Definitions, Function, Characteristics
- Current challenges

▶ INVERSE PROBLEM Starting with the effects to discover the causes has concerned physicists for centuries.

While in many ways, mixtures are not different to any other audio signal, two research questions stand out prominently: • Can we obtain the sources sj from the mixture x? • Can we find the number of sources J from x? These two questions are addressed in the scientific fields of sound source separation and source count estimation

Inverse problems appear when we want to see or examine something that we cannot access directly. What we have is an indirect measurement that contains hidden information.

An inverse problem is always a counterpart of a direct problem, as shown in the schematic diagram below. The direct problem is going from object to data, and the inverse problem is about finding the object back from the data.

The assumed few thousand taps. This model was very popular in the early stages of research [48]–[55]. Recently, interest has revived with sparse penalties which account for prior knowledge about the physical properties of AIRs, namely the facts that power concentrates in the direct path and the first early echoes [56]– [60] and that the time envelope decays exponentially [61], but these penalties have not yet been used in a BSS context.

1.2.2 *Echo-aware Processing*

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.

- Motivation
- Definitions, Function, Characteristics
- Current challenges

Auralization is the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modelled space

1.3 AUDIO INVERSE PROBLEMS

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

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

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

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

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:

[1] often referred to as *direct*

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

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 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"

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:

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

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

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 II

1.3.1 *General Processing Scheme*

Digital signal processing (DSP) is the process of analyzing and modifying a signal to optimize or improve its efficiency or performance. It involves applying various mathematical and computational algorithms to analog and digital signals to produce a signal that's of higher quality than the original signal. It is traditional in engineering to represent complex systems as a collection of simpler subsystems, with well-defined tasks, interacting with each other. In signal processing, these subsystems roughly fall into four categories: *representation*, *enhancement*, *estimation*, and *adaptive processing*. Many problems can be decomposed into blocks that belong to one of these categories.

Representation Objects can be represent (described) in many different way. Through different representations, some object *information* becomes more relevant and suitable for certain tasks than other.
Representation can be lossy or lossless, and are generally implemented through (non)linear mapping, such as change of basis or feature. The most famous representation is the Fourier basis.
Depending on the task the representation may be invertible. 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 inverse problems ill-posed and ill-conditioned. Therefore, signal enhancement, that is removing noise, is a necessary step.
Enhancement constitute a huge dome of methods: form simple denoising by averaging of repeated measurement to spectral subtraction to source separation with 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.

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

1.3.2 *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

1.4 My Thesis

1.4.1 *Hunting Acoustic Echoes*

1.4.2 *Echo-aware Auditory Scene Analysis*

1.5 Organization and Contributions

▶ Room Acoustic meets Signal Processing

Chapter 2 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.

▶ Hunting Acoustic Echoes

Chapter 4 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.

► Echo-aware Auditory 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.

Finally, the dissertation concludes with Chapter X, which summarizes the contributions and raises several additional research questions

1.6 This Thesis: 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.

▶ Quick vademecum for the readers:

Kuttruff, *Room acoustics*

- Bibliographic references are denoted as [Kuttruff 2016].
- 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.

Part I

ROOM ACOUSTIC MEETS SIGNAL PROCESSING

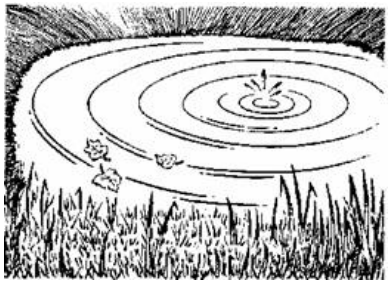
2

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 and PDEs: Kuttruff's *Room Acoustics*, Pierce's *Acoustics: an introduction to its physical principles and applications*, Duffy's *Green's Functions with Applications*. Notions on acoustic simulators are extracted from the Habet's tutorial paper *Room impulse response generator*. More technical details are reported in Appendix **??**.

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. [SE: non andrei a capo] 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.

[SE: perche' a wave? l'hai nominata prima: direi tipo che il sound propagates through a medium. e poi wave l'hai gia' messo in corsivo prima] [A *wave* is a disturbance that → 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 and composition. 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

[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.

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.

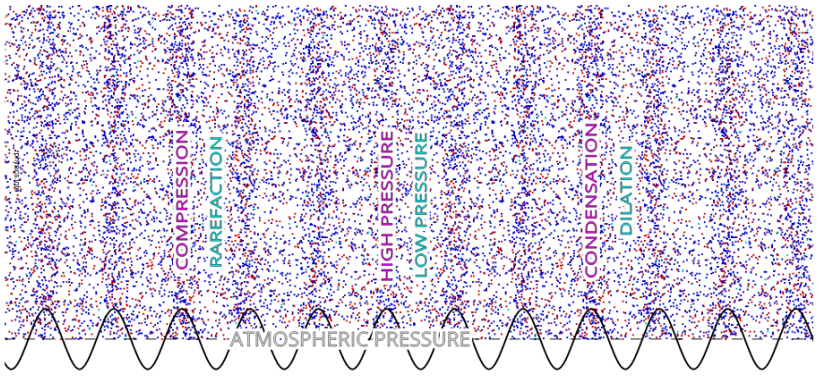


FIGURE 2.2: snapshot of a longitudinal wave in air

We think of this process in the light of the classic *source-medium-receiver* model of communication theory. The *source* is anything that emits or expends energy (waves)[SE: non mi piace che dici energy, di direttamente wave cosi' sei corente con i vocaboli e non aggiungi termini nuovi: non hai mai parlato di energia][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 $\mathbf{x}$ [m] and time t [s] [SE: non mi piace che metti le unita' di misura]

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

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 $[\frac{\text{m}}{\text{s}}]$.

Despite its complicated formulation, the wave equation is linear leading to the following implications:

- 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.

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*[9],
- the *conservation of mass*, and

[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.

[9] In fluidodynamics, it comes with the name of the Euler's equation. [SE: inutile]

- the *polytropic process relation*[10].

[10] meaning that the medium is an ideal gas undergoing a reversible adiabatic process. [SE: questa la metterei nel testo: non e' quella che ti dice che il gas deve essere omogeneo? alloara l'however dopo avrbbe senso]

[SE: howver a cosa?] 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[wave] equation 2.2 is expressed in the space-time domain $(\mathbf{x}, t)$. By applying the temporal Fourier transform, we obtain the *Helmholtz equation*, i. e.

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

where $k = \frac{2\pi f}{c}$ is known as *wave number* $[\mathrm{m}^{-1}]$, [that → and] relates the frequency f [Hz] [and → to] the propagation velocity c.

Both the wave[equation] 2.2 and the Helmholtz's equation 2.3 are source-independent, namely no source is present in the medium. Therefore they are [called → said to be] *homogeneous* as the right-hand term is zero. [SE: Non andrei a capo] 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 distribution function $s(t, \mathbf{x})$, the wave equation can be written

The minus sign is a conventions. [SE: inutile]

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

Then, the corresponding Helmholtz's equation writes

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

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

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

The function $H(f, \mathbf{x} \mid \mathbf{s})$ that satisfy Eq. (2.6) is called the *Green's function* and is associated to Eq. (2.3), of which it is also a solution.
In the next subsection, we will see that the function H can be interpreted as the free-field *Transfer Function* between the source at $\mathbf{s}$ and the receiver at $\mathbf{x}$.

2.1.2 ... *and its green solution*

THE 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.

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].

They can be seen as a concept analogous to
impulse responses[11] *in signal processing.[SE: citazione! analogo dove? in fisica? in acustica]*

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

Under this light, the physic so-far can be rewritten [in → using] the vocabulary of the communication theory, namely *input*, *filter* and *output*.

If one ignores the space integral, one can see the close relation with a transfer function. [SE: nel testo]

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

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

where $\mathcal{V}_s$ denotes the source volume, and $\mathrm{d}\mathbf{s} = \mathrm{d}x_s \; \mathrm{d}y_s \; \mathrm{d}z_s$ the differential volume element at position $\mathbf{s}$.
The requested sound pressure $p(\mathbf{x}, t)$ 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

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

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

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

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

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 $q = \|\mathbf{x}-\mathbf{s}\|$ from the source to the microphone.

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*.

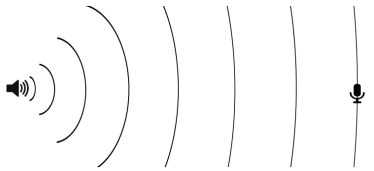


FIGURE 2.3: Visualization of the sound propagation. Since the sensor (i.e. a microphone) is drawn in the far field, the incoming waves can be approximated as plane waves.

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.

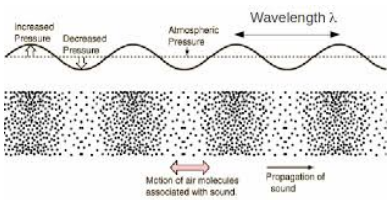


FIGURE 2.4: wavelength

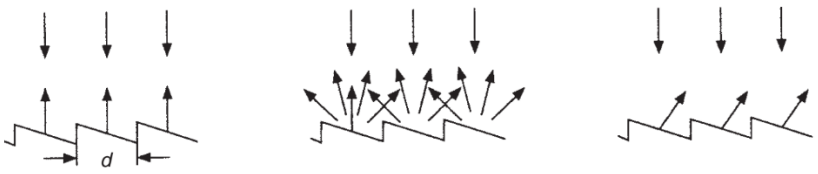


FIGURE 2.7: A reflector having irregularities on its surface with width d much smaller than the sound wavelength λ. Image courtesy of [Kuttruff 2016].

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 impedance*, which describes their acoustic inertia as well as the energy dissipation. The remaining contribution may continue to propagate resulting in the refraction phenomenon[12].

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*[13] 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 2.4, 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 2.7

- $\lambda \gg d$, 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.

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

[12] This is more commonly observed when light passes thought different medium, like a prism.

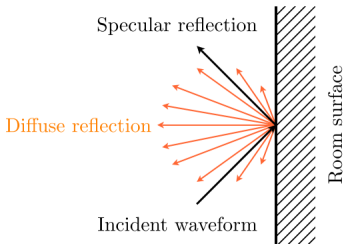
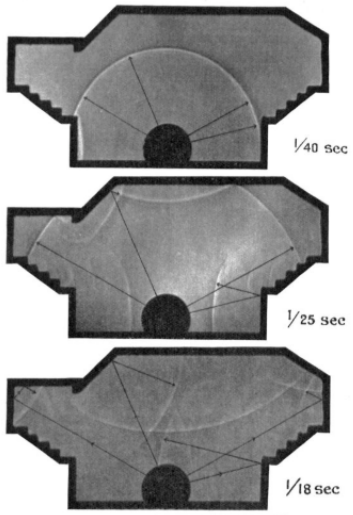


FIGURE 2.5: Specular vs diffuse reflection

[13] for instance near- vs. far-field

" 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]



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]

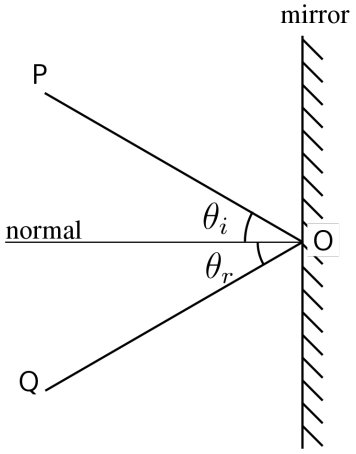


FIGURE 2.8: Schematic representation of specular reflection.

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.

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 the audio signal processing. It comes from the following approximations [Savioja and Svensson 2015]: (*i*) The energy or intensity of the plane wave[14], 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 assumption are motivated by the difficulty of measuring the acoustic impedance and the possibility to compute an equivalent coefficient a posteriori

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.

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

▶ ECHOES ARE SPECULAR REFLECTIONS which stand out in terms of energy strength or timing. Originally this term used to refer to sound reflections

The word echo derives from the Greek "echos", litterarly "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.

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.

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 2.9 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.

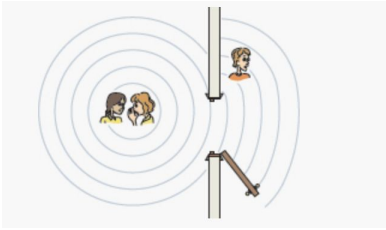


FIGURE 2.9: Schematic representation of sound diffraction. This effect allows to hear "behind walls".

2.3 ROOM ACOUSTICS AND ROOM IMPULSE RESPONSE

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

room *is a physical enclosure containing the medium and has boundaries limit the sound propagation.*

MATHEMATICALLY the sound propagation is described by the wave equation (2.2). By solving it, the Acoustic Impulse Response (AIR)[15] from a source to a microphone can be obtained. In the context of room acoustics, it is commonly referred to as Room Impulse Response (RIR), usually to put attention of on the geometric relation between reflections and the geometry of the scene. In this thesis the two terms will be used indistinctly.

[15] Acoustic Transfer Function (ATF) in the Fourier transform of the AIR

2.3.1 *The room impulse response*

It is a fundamental concept of this dissertation and it is where physical room acoustic (Green's function/Solution of wave equation) and indoor audio signal processing meets. From now on, we well adopt an signal processing perspective and

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

Figure 2.10 provides a schematic illustration of the shape of a RIR in comparison with measured one.

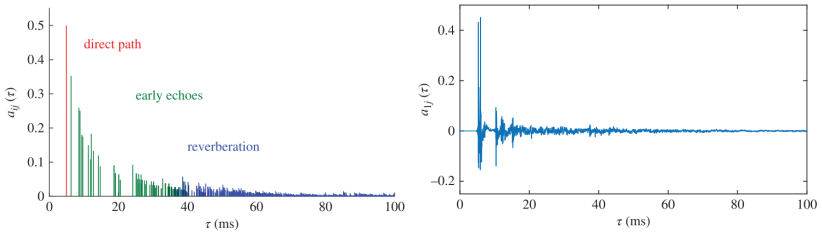


FIGURE 2.10: Schematic illustration of the shape of an RIR and the first 100 ms of a measured one.

RIRs usually exhibits common structure. Based on the consideration in § 2.2, they are commonly divided into three components [Kuttruff 2016]:

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

where

Direct path $h^d(t)$ is the line-of-sight contribution of the sound wave. This term coincides with the spike modeled by the free-field propagation[16].

[16] Cf. The free-field Green's function, i. e. Eq. (2.9))

Echoes or Early Refelction 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. This first reflections are typically specular and are well modeled in general by the Image Source Method (ISM)[17].

[17] Cf. § 2.3.3

Later 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.

This 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 with, let $s(t)$ be the source signal, sound received is

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

where the symbol $*$ is the convolution operator.

A part for certain simple scenarios, computing RIRs in closed forms is a cumbersome task. Therefore numerical solver or approximation model are used instead.

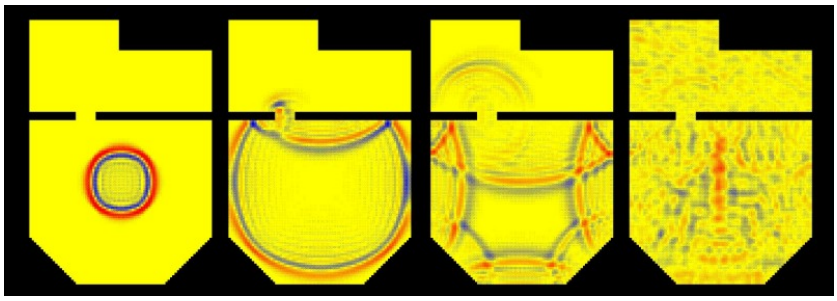


FIGURE 2.12: Simulation of Sound propagation at four consecutive timestamps using the DWM technique. A short, sharp, impulsive sound fired into the larger of two rooms causes a circular wavefront to spread out from the sound source. The the wave is reflected from the walls and part of it passes through 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.

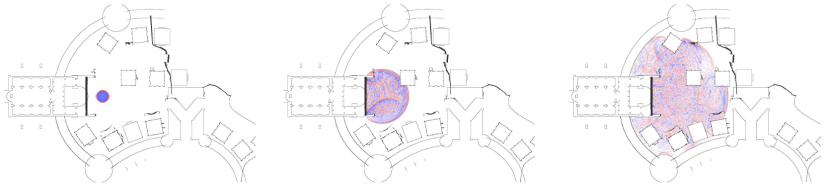


FIGURE 2.13: Sound propagation at three consecutive timestamps using the FDTD-based *Triton* simulator from Microsoft

2.3.2 *Simulating room acoustics*

[18] There are two main categories: geometric and wave-based methods [Habets 2006; Savioja and Svensson 2015; Thomas 2017].

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

wave-based aims at solving the wave equation numerically, while

geometric methods make some simplifying assumption about the wave propagation: they typically ignore the *wave* behavior of the sound, choosing much lighter models such as *ray*s or *particle*s.

▶ WAVE-BASED METHODS

These are iterative methods that divide the 3D bounded enclosure into a grid of interconnected nodes [19]. For instance, the Finite Element Method (FEM) divide the space into small volume elements smaller of the sound wavelengths, while Boundary Element Method (BEM) divide only the boundaries of the space are divided 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.

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

The Finite-Difference-Time-Domain (FDTD) method replace 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.

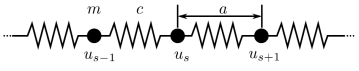


FIGURE 2.11: Example of mass-spring linear mesh used to simulate a 1D transversal wave.

THE MAIN DRAWBACK OF THESE METHODS is discretisation problem: less dense grid may simplify too much the simulation, while denser grid increase

the computational load. Moreover, they require delicate definitions of the boundaries condition at the physical lever, like knowing complex impedances, parameters not always available in the literature.

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* [20], namely collection of resonances that exist in a room and characterize it.

As stated in [Välimäki et al. 2016], among the wave-based methods, Digital Waveguide Mesh (DWM) are usually preferred: they run directly in the time domain, requiring typically an easier implementation, and they exhibits a natural huge level of parallelism.

[20] 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.

- GEOMETRIC METODS

  They can be 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].

- STOCHASTICS 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 time-frequency domain [Badeau 2019]. Rather than the detailed room geometry, these methods generally use high-level descriptors[21] 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 qualities of the prior knowledge modeled.

  RAY-TRACING [Kulowski 1985] is one the most common method that fall in this category and very popular in the field of computer graphic for light simulation. 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 all the *diffuse rain algorithm* [Schröder et al. 2007; Heinz 1993] is commonly used in many acoustic simulator. 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 histogram, namely a *directional-time-frequency energy map* of the room's diffuse sound field for a giver receiver location (Cf. Figure 2.15) This map is then used as prior distribution for drawing random set of impulses which are used to form the RIR.

  While neglecting some detailed description of early reflection and room modes, these methods are good to capture and simulate the statistical behavior

[21] such as the amount of reverberation

of the diffuse sound field for low computational cost.

- DETERMINISTIC methods are good to simulate early reflection instead: they accurate traces the exact direction and the timing of the main reflections' paths.

  The most popular is the Allen and Barkley's Image Source Method (ISM) [Allen and Berkley 1979]. Even if the basic idea is rather inutile and simple, the model is able to produce the exact solution to the wave equation for a 3D shoebox with rigid walls. Since it models only specular (perfect) reflections, ignoring diffuse and diffracted components. it only approximate arbitrary enclosures and the late diffuse reflections.

  The naïve 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. Two are the main limitation of this method. First, in a shoebox the complexity of the algorithm is cubic in the order of reflection and for order higher 30 the algorithm become 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 method to model the full impulse response.

- HYBRID METHODS 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 field. And in general geometric methods fails to proper model lower frequencies and room modes. The waveguide method models physical phenomena better than the geometric methods, but is expensive at high frequencies. All these limitations corresponds into three regions in the Time-Frequency (TF) representation of the RIR. As depicted in Figure 2.19,

  - 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.

  By combining three methods, accurate broadband impulse responses can be synthesized, but for a much lower computational cost than would be possible with any individual method. 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 CROSSOVER POINT in the time domain is called *transition time* or *mixing time*. It 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 [22].
  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

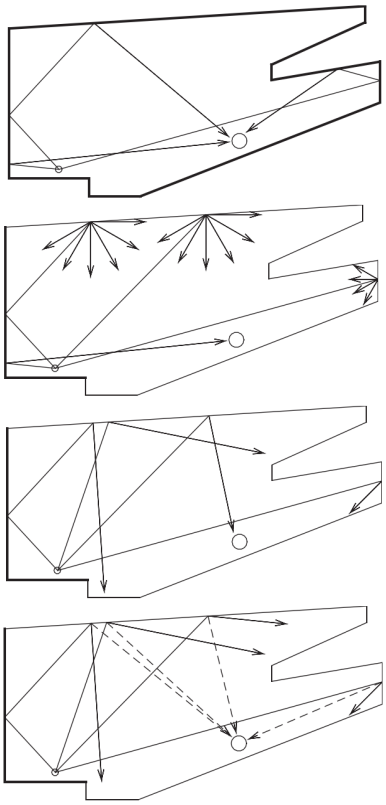


FIGURE 2.14: Visualization of ray-tracing method. From top to bottom: first the method will eventually find specular reflection; then diffuse reflections can be modeled either by splitting a ray into several new rays or a single random one. In the diffuse rain technique a shadow-ray is cast from each diffuse reflection point to the receiver to speed-up convergence of the simulation. Image courtesy of [**Savioja2015goemetric**]

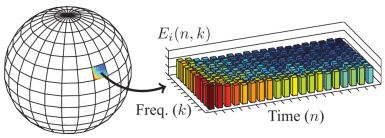


FIGURE 2.15: Directional-time-frequency Energy map resulting form the diffuse rain algorithm[Schröder et al. 2007]. For each direction, that is receiver's spherical bin, a time-frequency histogram collects the energy of incoming rays. Image courtesy of [Schimmel et al. 2009]

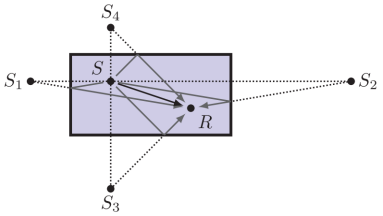
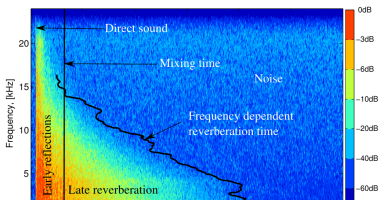


FIGURE 2.16: Virtual image sources correspond to multiple sound propagation paths in the ISM. Image courtesy of [Schimmel et al. 2009]



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 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 application 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 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 produce a solution the Helmholtz's equation for rectangular enclosure with rigid boundaries.

- ON A SINGLE REFLECTOR, 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. As show in Figure 2.21, 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 has been "mirrored" in the surface. In this way, the each wavefront that arrives to the receiver from each reflection off the walls as the direct path received from an equivalent (or image) source.

  The ISM makes use of the following assumptions:

  - sound source and receiver as points in a rectangular cavity
  - 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
  - Rays are perfectly reflected at boundaries

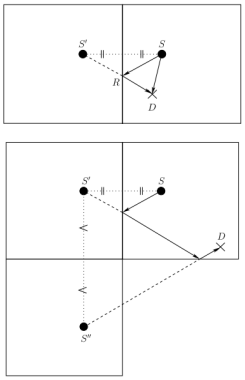


FIGURE 2.20: Path involving one reflection obtained using first-order image (top) and two reflections obtained using two images. It. Image courtesy of [Habets 2006].

- FINALLY 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 in position $\mathbf{s}$ and the receiver in $\mathbf{x}$ reads

$$h_{\text{ISM}}(t, \mathbf{x} \mid \mathbf{s}) = \sum_{r=0}^{R} \frac{1}{4\pi \|\mathbf{x} - \mathbf{s}_r\|} \delta\left(t - \frac{\|\mathbf{x} - \mathbf{s}_r\|}{c}\right) \tag{2.15}$$

where $\mathbf{s}_r$ is the r-th image of the source.

The above equation assume perfect rigid and reflective wall. In order to easily incorporate frequency-dependent acoustic impedances (and absorption coefficient) of real surfaces, the Fourier transform of Eq. (2.16) is consider instead, where each reflection term addendum is appropriately scaled

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

where α_r is the damping coefficient related to the r-th image which in general consider the all the absorption coefficient of the considered surfaces.

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. [23]

[23] ***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.

- Early Reflections and Echoes are reflections which are by nature highly correlated with to the direct sound. They convey a sense of geometry which modify 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 $5\,\text{ms}$ to $40\,\text{ms}$. However, the perceived spatial location carried by the first-arriving sound is preserved suppress 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 optical 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)[24], i. e. the ratio between the direct path ration and the remain portion of the RIR. Regarding the depth perception, early reflection are the main responsible. In the context of virtual reality, correct modeling of these quantities is essentials in order to maintain a coherent depth impression [Kearney et al. 2012].

[24] Cf. § 2.4.4

- 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 depend 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

> mixing time *is the instant that divides the early reflections from the late reverberation in a RIR,*

And it is represented in Equation 2.47 by the symbol Tm. Due to this, it is an parameters important also in the context of RIRs synthesis as it defines cross-over point for room acoustics simulator using hybrid methods [Savioja and Svensson 2015][25].

[25] Cf. § 2.3.2

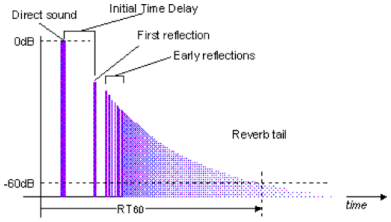


FIGURE 2.21: illustration of the Reverberation Time ($\mathtt{RT}_{60}$) definition. It. Image courtesy of wikipedia.

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. A practical approach is presented in **??**.

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

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

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

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 2.22 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.

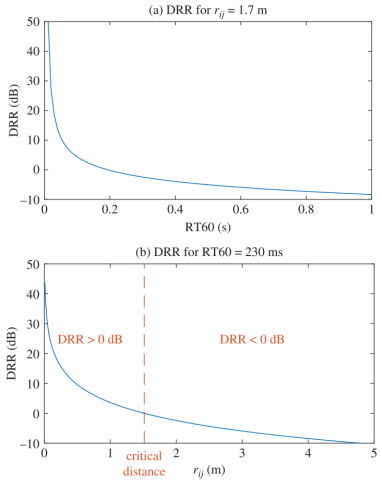


FIGURE 2.22: DRR as a function of the RT60 and the source distance rij based on Eyring's formula (Gustafsson et al., 2003). These curves assume that there is no obstacle between the source and the microphone, so that the direct path exists. The room dimensions are the same as in Figure 3.1.

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 move from the physics to signal processing. At first this chapter formalized fundamental concepts of audio signal processing such as signal, mixture and noise § 3.1 in the time domain. Since most of the estimation and processing will be conducted in the frequency domain, § 3.2 presents this representation where the relation between audio signal and sound propagation can be easily written. Finally, after assuming the narrowband approximation, in § 3.3 some important models for the RIR are described.
The material presented in this chapter is extracted from the book [Vincent et al. 2018] while echo model its consideration derives from the work [Di Carlo et al. 2020].

Vincent et al., *Audio source separation and speech enhancement*

3.1 Signal model in the time domain

A raw *audio signal* encodes the variation of pressure over time on the microphone membrane. Mathematically it is denoted as the function

$$\tilde{x}(t) \in \mathbb{R}, \tag{3.1}$$

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

Nowadays signals are typically processed, stored and analyzed as *digital audio signal*. This corresponds to the discrete-time signal $\hat{x}$ obtained by periodically sampling[26] the continuous-time signal $\tilde{x}$ at rate F_s [Hz]. 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 $\tilde{\phi}_{\text{LP}}$[27] with frequency support in $]-F_s/2, F_s/2]$; then its time-support is regularly discretized, $t = n/F_s$ for $n \in \mathbb{Z}$. This will restrict the frequency support of the signal to satisfy the Nyquist–Shannon sampling theorem ~~and~~ in order to avoid aliasing effect. This is expressed by

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

where $\star$ is the continuous-time convolution operator. Moreover, in this thesis we assume that amplitudes of discrete signals are real values, ignoring the *quantization* process.

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 $44.1\,\text{kHz}$ for music distribution on CDs and $8\,\text{kHz}$ for first-generation speech communication. Now multiple of $8\,\text{kHz}$ are typical used in audio processing: $(16, 48, 96, 128\,\text{kHz})$.

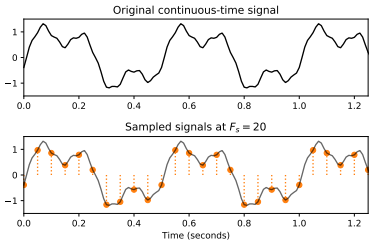


Figure 3.1: Continuous-time signal and its sampled version.

[26] The use of the word *sample* will have different meanings in the context of machine learning, where a sample is an instance of a set instead of a time instant.

[27] The ideal low-pass filter is $\tilde{\phi}_{\text{LP}}(t) = \text{sinc}(t) = \sin(\pi t)/(\pi t)$

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 emit sound. When multiple sources are active at the same time, the sound that reaches our ears or is recorded using a microphone is superimposed or *mixed* into a single sound. This resulting signal is denoted as *mixture*.

[28] Please note, the term *channel* has also different meaning: it indicates the medium in communication (e. g. Channel Estimation) and sometimes one of the dimension of the input in machine learning (e. g. image's channel).

⇄ SINGLE-CHANNEL VS. MULTICHANNEL: The term *channel*[28] is used here to indicate the output of one microphone 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 $\tilde{\boldsymbol{x}}(t) = [\tilde{x_1}(t), \ldots, \tilde{x_I}(t)]^{\mathrm{T}} \in \mathbb{R}^I$.

⇄ POINT VS. DIFFUSE SOURCES: *Point sources* are emitted by a single and well-defined point in ~~the~~ space and their signal is single-channel. Human speakers or the sound emitted by a loudspeaker can, for instance, be approximated to point sources. As opposed to, Wind, traffic noise, or large musical instruments which emit sound in a large region of space are considered *diffuse sources* Their sound cannot be associate to a punctual source, but rather to a distributed collection of them.

⇄ DIRECTIONAL VS. ONMIDIRECTIONAL: An *omnidirectional* source (resp. receiver) emits (resp. pick up) sound equally from all directions, both in time and in frequency. Although this simplify greatly processing models and frameworks, this is not true in real scenario. The physical properties of real sources (resp. receivers) lead to *directivity patterns*, a. k. a. *polarity*, which may be different at different frequencies. In this thesis we will assume always 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 microphone i and each source j have a well defined position in the space, $\mathbf{x}_i$, $\mathbf{s}_j$, respectively.

The mixing process describes then the nature of the mixtures. In order to better formalized it, Sturmel et al. 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* $\tilde{x_j}(t)$ is the possibly non-linear combination of images associated to the source j. Depending on the "contribution" the image describes, the following type of mixture can be defined:

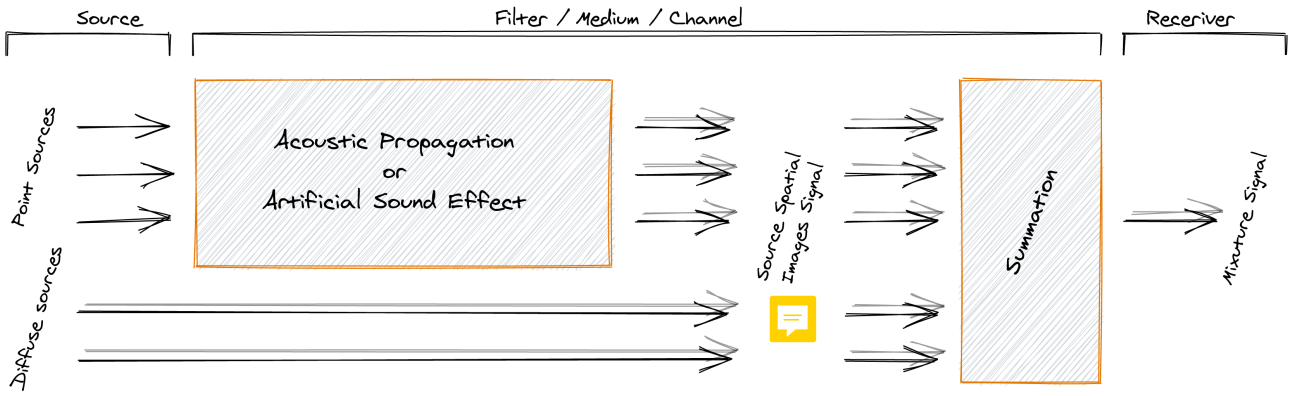


FIGURE 3.2: General mixing process, illustrated in the case of $J = 3$ sources, including three point sources and one diffuse source, and $I = 2$ channels.

⇄ 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.

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}.

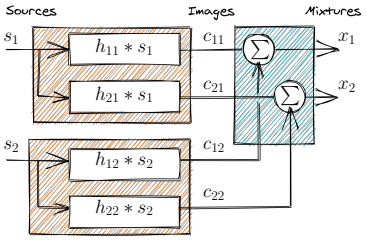


FIGURE 3.3: Graphical representation of the mixing model 3.4 for 2 sources and 2 microphones.

▶ 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:

$$\tilde{c}_{ij}(t) = \left(\tilde{h}_{ij} \star \tilde{s}_j\right)(t) \tag{3.3}$$

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

$$\tilde{\boldsymbol{x}}(t) = \sum_{j=1}^{J} \tilde{\boldsymbol{c}}_j(t). \tag{3.4}$$

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:

$$\tilde{h}_{ij}(t) = \sum_{r=0}^{R} \frac{\alpha_{ij}^r}{4\pi c \tau_{ij}^r}\, \delta\left(t - \tau_{ij}^r\right) + \varepsilon_{ij}(t) \tag{3.5}$$

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.

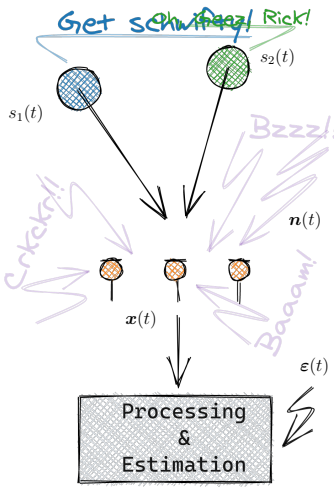


FIGURE 3.4: Graphical representation of the mixing model (3.4): $s_2(t)$ is the *interferer*, $\boldsymbol{n}(t)$ contributes to the *diffuse noise field*, and $\boldsymbol{\varepsilon}(t)$ model acquisition and modeling errors.

3.1.2 *Noise, interferer and errors*

In Eq. (3.4) 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. (3.5) 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.

By making the noisy terms explicit, the mixing model in Eqs. (3.3) and (3.4) writes:

$$\tilde{c}_{ij}(t) = \left(\tilde{h}_{ij} \star \tilde{s}_j\right)(t) + \tilde{\varepsilon}_{ij}(t) \tag{3.6}$$

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

$$\tilde{\boldsymbol{x}}(t) = \sum_{j=1}^{J} \tilde{\boldsymbol{c}}_j(t) + \tilde{\boldsymbol{n}}(t) \tag{3.7}$$

3.2 SIGNAL MODEL IN THE SPECTRAL DOMAIN

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.

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.

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:

$$\tilde{X}(f) = (\mathcal{F}\tilde{x})(f) = \int_{-\infty}^{+\infty} \tilde{x}(t) \mathrm{e}^{-\mathrm{i}2\pi ft} \, \mathrm{d}t, \tag{3.8}$$

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. (3.4) simply as:

$$\tilde{\boldsymbol{x}}(t) = \sum_{j=1}^{J} \tilde{\mathbf{c}}_j(t) \quad \overset{\mathcal{F}}{\to} \quad \tilde{X}(f) = \sum_{j=1}^{J} \tilde{C}_j(f) \tag{3.9}$$

Secondly, by the *convolution theorem*, the source spatial images in Eq. (3.3) writes as:

$$\tilde{c}_{ij}(t) = \left(\tilde{h}_{ij} \star \tilde{s}_j\right)(t) \quad \overset{\mathcal{F}}{\to} \quad \tilde{C}_{ij}(f) = \tilde{H}_{ij}(f)\tilde{S}_j(f). \tag{3.10}$$

As discussed in Chapter 2, the FT of a RIR can be computed exactly in closed-form as

$$\tilde{H}_{ij}(f) = \sum_{r=0}^{R} \frac{\alpha_{ij}^r}{4\pi c \tau_{ij}^r} \mathrm{e}^{-\mathrm{i}2\pi f \tau_{ij}^r}. \tag{3.11}$$

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 have only access to finite and discrete-time microphones signals for which the properties (3.10) is valid with some precautions.

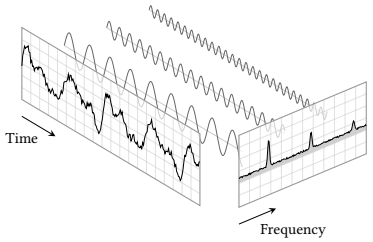


FIGURE 3.5: A signals resolved into its Fourier series: a linear combination of sines and cosines represented as peaks in the frequency domain.

3.2.1 *Discrete frequency domain*

The spectral representation of a discrete- and finite-time signal $\hat{x}[n]$ is given by its (forward) Discrete Fourier Transform (DFT) [29], $\mathbf{F} : \mathbb{R} \mapsto \mathbb{C}$:

$$\hat{X}[k] = (\mathbf{F}\,x)[k] = \sum_{n=0}^{N-1} \hat{x}[n] \mathrm{e}^{-\mathrm{i}2\pi kn/F}. \tag{3.12}$$

where $k \in [0, F-1]$ in the discrete *frequency bin* and F is the total number of bins. The natural frequency f_k in Hz corresponding to the k-th frequency bin can be computed as

$$f_k = \frac{k}{F} F_s. \tag{3.13}$$

The DFT is linear, so the discrete version of Eq. (3.9) becomes

$$\hat{\boldsymbol{x}}[n] = \sum_{j=1}^{J} \hat{\mathbf{c}}_j[n] \quad \overset{\mathbf{F}}{\to} \quad \hat{\boldsymbol{x}}[k] = \sum_{j=1}^{J} \hat{\mathbf{c}}_j[k] \tag{3.14}$$

Secondly, by using naïvely the discrete convolution theorem, one may translate Eq. (3.3) as

$$\hat{c}_{ij}[n] = (\hat{h}_{ij} * \hat{s})[n] \quad \overset{\mathbf{F}}{\to} \quad \hat{C}_{ij}[k] \approx \hat{H}_{ij}[k]\hat{S}[k] \tag{3.15}$$

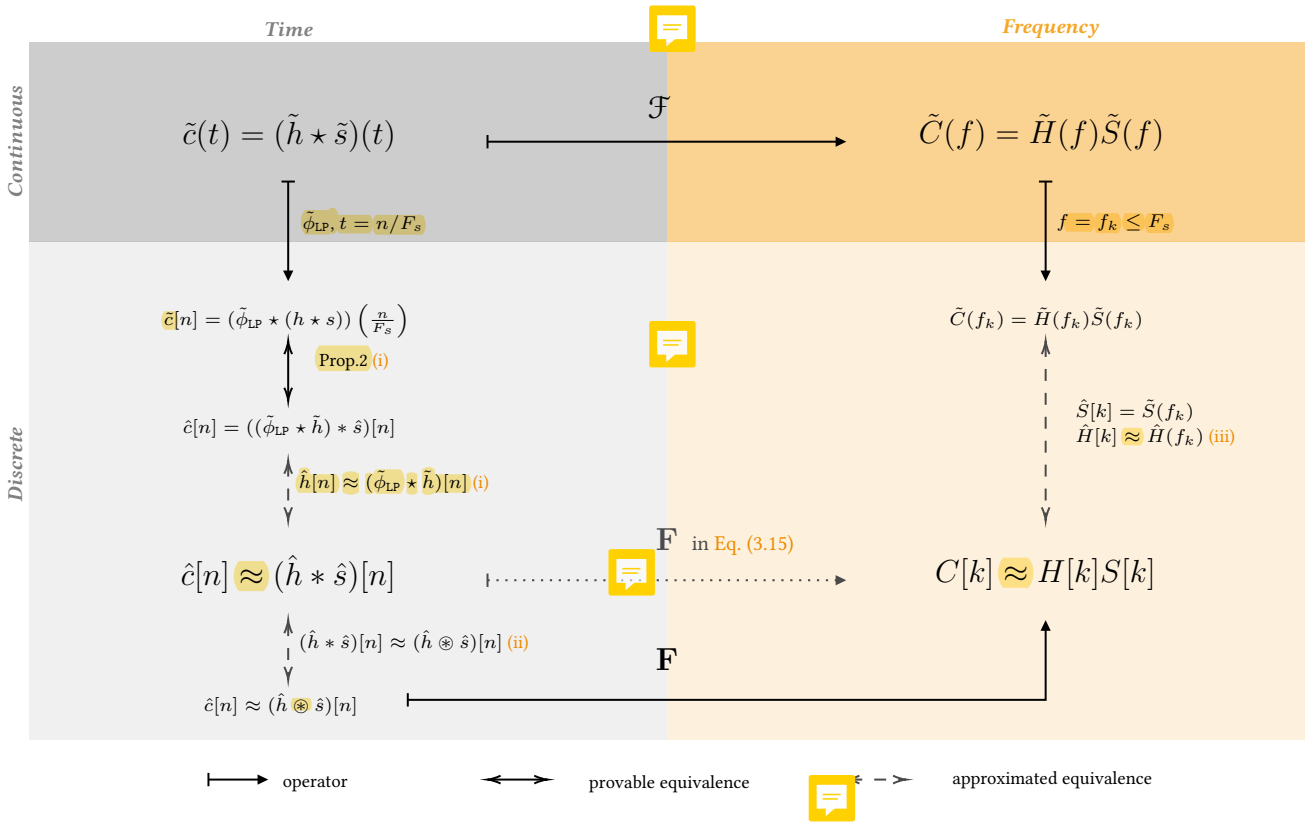
[29] This can be intrepreded as the projection onto the space spanned by a finite number of complex exponentials.

[30] 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 $n = 0, \cdots, D-L$.

where $*$ is the finite-time linear convolution operator[30] and the RIR as

$$\hat{H}_{ij}[k] \approx \sum_{r=0}^{R} \frac{\alpha_{ij}^r}{4\pi c \tau_{ij}^r} \mathrm{e}^{-\mathrm{i}2\pi f_k \tau_{ij}^r}. \tag{3.16}$$

▶ ALTHOUGH USED IN PRACTICE, this model makes use of approximations. As explained in [Tukuljac et al. 2018], issues arise from simultaneously using the closed-form RIR model derived in Eq. (3.11) and the sampled observations $\hat{x}_i[n]$. The paper mention three approximations, which are depicted in the following diagram.



The diagram shows a chain of operators (measurements and transforms), provable and approximated equivalences that lead to Eq. (3.15). 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* convolve the discrete signal $\hat{s}[n]$ and the discrete and low-passed version of the filter. The source signal is band-limited by nature, however, the $\tilde{h}(t)$ is not (in fact the RIR is modeled as a ~~finite~~ summation of spikes, which has infinite spectrum). Thus, the first approximation (i) considers $\tilde{h}[n] \approx (\tilde{\phi}_{\text{LP}} * \tilde{h})[n]$, in words we assume that the filter is band-limited by $\pm F_s/2$.
Tukuljac et al. made an important observation here:

> "Here, it is important to note that contrary to intuition, even in the idealized case where an infinite number of samples are available, the discrete-time filters $\hat{h}[n]$ involved in the measurement model are never streams of Diracs, but non-sparse,

infinite-length filters consisting of decimated combinations of sinc functions." [Tukuljac et al. 2018].

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. (3.16) which would require infinitely many samples and unlimited frequency support to be computed[31].

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

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

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. ~~However the information on when these frequencies occur is hidden in the transform~~. In order to jointly account for both temporal and spectral characteristic, joint time-frequency representations are used.

3.2.2 *Time-Frequency domain representation*

Time-Frequency (TF) representations aim to jointly describe the signal in time and frequency domain. 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, $w[n]$, whose is nonzero for only a period of time W shorter than the entire signal length, $W \ll N$. 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 Short Time Fourier Transform (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 processes to compute the complex

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

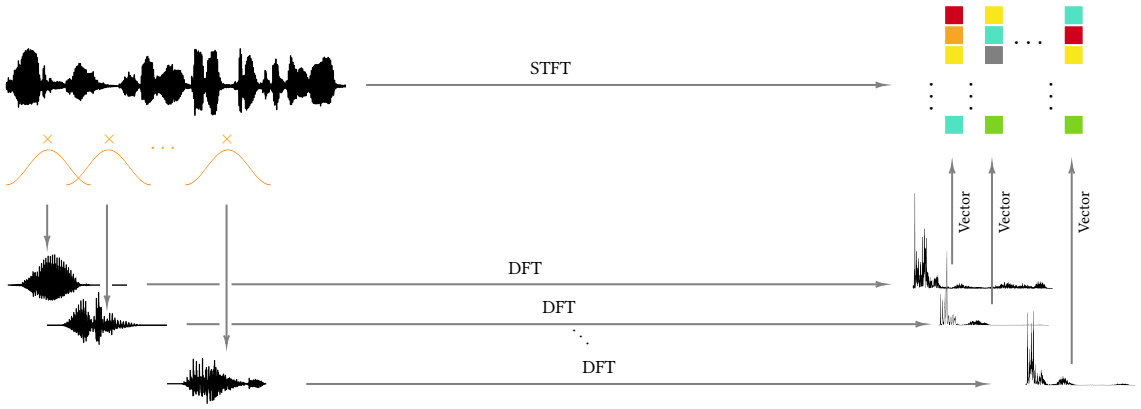


FIGURE 3.7: Schematic representation of the STFT transforms. At first the signal is windowed. Then the DFT of each frame is computed and the results stuck column-wise to form a matrix (frequency bin times frame index).

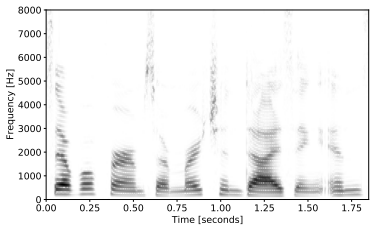


FIGURE 3.6: STFT spectrogram of an example speech signal. Higher energies are illustrated with darker colors.

For more mathematical detailed description on DFT and STFT can be found in [Oppenheim 1987]. For a audio-processing-oriented and music-processing-oriented explanation please refer also to Chapter 2 of [Vincent et al. 2018] (Chapter2) and Chapter 2 of [Müller 2015], respectively.

STFT coefficients is given by

$$x[k,l] = \sum_{n=0}^{W-1} w[n]x[n+lH]\mathrm{e}^{-\mathrm{i}2\pi kn/F} \quad \in \mathbb{C} \tag{3.17}$$

where W is the window length and H is the *hop size* which specify how much the window needs to be shifted across the signal. Equivalently, Eq. (3.17) can be expressed as DFTs of windowed frames, $x[k,l] = \mathbf{F}\, x[n,l]$ where $x[n,l] = x[n+lH]w[n]$.

Since each STFT coefficient $x[k,l]$ lives in the complex space $\mathbb{C}$, the squared magnitude of the STFT, $|x[k,l]|^2$ is commonly used for visualization and for processing The resulting two-dimensional representation is called *spectrogram*. It can be visualized by means of a two-dimensional image, whose axes represent time frames and frequency bins. In this image, the value $|x[k,l]|^2$ is represented by the intensity or color in the image at the coordinate $[k,l]$. Throughout this work both estimation and processing will be conducted in the STFT domain. 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 **??**.

The STFT has the following useful properties for audio processing:

- the frequencies scale f_k is a linear function of the frequency bin k;
- the resulting matrix allows easy treatment of the phase $\angle x[k,l]$, the magnitude $|x[k,l]|$ and the power $|x[k,l]|^2$ separately;
- the DFT can be efficienciently computed with the Fast Fourier Transform (FFT) algorithm;
- 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.~~

3.2.3 *The final model*

The model (3.15) shows how in practice the RIRs are threated 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 issues, the *convolutional STFT* for arbitrary window functions have been proposed[32] [Gilloire and Vetterli 1992]. Although mathematically exact, it is computationally and memory intensive.

[32] 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. This known in the literature as the *narrowband approximation*, namely the time-domain filtering can be approximated by complex-valued multiplication in each time-frequency bin $[l, k]$:

$$\boldsymbol{c}_j[l,k] \approx \boldsymbol{h}[k] s_j[l,k], \tag{3.18}$$

where the $\boldsymbol{h}_j(f) = [h_{1j}(f), \cdots, h_{Ij}(f)]^{\mathrm{T}}$ is the $I \times 1$ vector of the room transfer functions for ~~the~~ source j. It is sometimes practical to concatenate all this vectors into an $I \times J$ matrix $\mathbf{H}(f) = [\boldsymbol{h}_1(f), \cdots, \boldsymbol{h}_J(f)]$ called *mixing matrix*.

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

$$\boldsymbol{x}[l,k] = \mathbf{H}[l,k]\boldsymbol{s}[l,k] + \boldsymbol{u}[l,k] \tag{3.19}$$

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, ~~embed in processing frameworks~~, compute and estimate. The representation~~s~~ of ~~the~~ RIR discussed so far explicitly model early echoes e reverberation deterministically. Furthermore, alternative models are common in the audio processing literature.

3.3.1 *Steering vector model*

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

$$\boldsymbol{d}_j[k] = \left[\frac{1}{4\pi q_{1j}} \mathrm{e}^{-\mathrm{i}2\pi f_k q_{1j}/c}, \cdots, \frac{1}{4\pi q_{Ij}} \mathrm{e}^{-\mathrm{i}2\pi f_k q_{Ij}/c},\right] \tag{3.20}$$

Furthermore, assuming far-field regimes, the microphone-to-source distance q_{ij} are larger than the inter-microphone~~s~~ 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. (3.19). Dropping the dependency on j for readability and taking the first channel as reference, the Relative Transfer Function (RTF) associated the the i-th channel is defined as the element-wise ratio of the (D)FTs of the two filters [Gannot et al. 2001]

$$\tilde{h}_i[k] = \frac{h_i[k]}{h_1[k]}. \tag{3.21}$$

The 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 x_i and x_1, their signals can be re-written in term of $\tilde{h}_i$ as follows

$$\begin{cases} x_1 = h_1 * s + u_1 \\ x_i = h_i * s + u_i \end{cases} \rightarrow \begin{cases} x_1 = h_1 * s + u_1 \\ x_i = \tilde{h}_i * h_i * s + u_i \end{cases} . \tag{3.22}$$

Notice that $h_i = \tilde{h}_i * h_1$, corresponding to Eq. (3.23) in the frequency domain. Moreover although the real-world RIRs h_1 and h_i are causal, their RTF need not be so.

In ?? methods for estimation the RTF will be discussed

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" with respect to the RIRs estimation. In fact, in noiseless case, it holds that $x_i = \tilde{h}_i * x_1$.
- The RTF encodes properties of the related impulse responses and there are many advance methods to estimate them. Therefore, it may be used as a proxy for the estimations of (components of) RIRs.
- A RIR can be seen as a special case of 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 $\tilde{h}_i = h_i$[33].
- As discussed below, ~~also~~ RTFs simplify to special steering vectors in free- and far-field, which has interesting geometrical properties.

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

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

$$\tilde{\boldsymbol{h}}_j[k] = \frac{1}{h_{1j}[k]} \boldsymbol{h}_j[k]. \tag{3.23}$$

▶ THE RELATIVE STEERING VECTORS results by combining Eqs. (3.20) and (3.23) as

$$\tilde{\boldsymbol{d}}_j[k] = \left[1, \mathrm{e}^{-\mathrm{i}2\pi f_k (q_{2j} - q_{i'j})/c}, \cdots, \mathrm{e}^{-\mathrm{i}2\pi f_k (q_{Ij} - q_{i'j})/c},\right] \tag{3.24}$$

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*[34]. In fact, the RTFs encodes the so-called Interchannel Level Difference (ILD) and the Interchannel Phase Difference (IPD)

$$\begin{aligned} \mathtt{ILD}_{ij}[k] &= 20 \log_{10} \left|\tilde{h}[k]\right| \quad [\mathrm{dB}] \\ \mathtt{IPD}_{ij}[k] &= \angle \tilde{h}[k] \qquad\quad [\mathrm{rad}] \end{aligned} \tag{3.25}$$

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

As shown in Figure 3.8, the ILD and the IPD cluster around the the steering vectors, that is direct path components. However early echoes and reverberation make them significantly diverge.

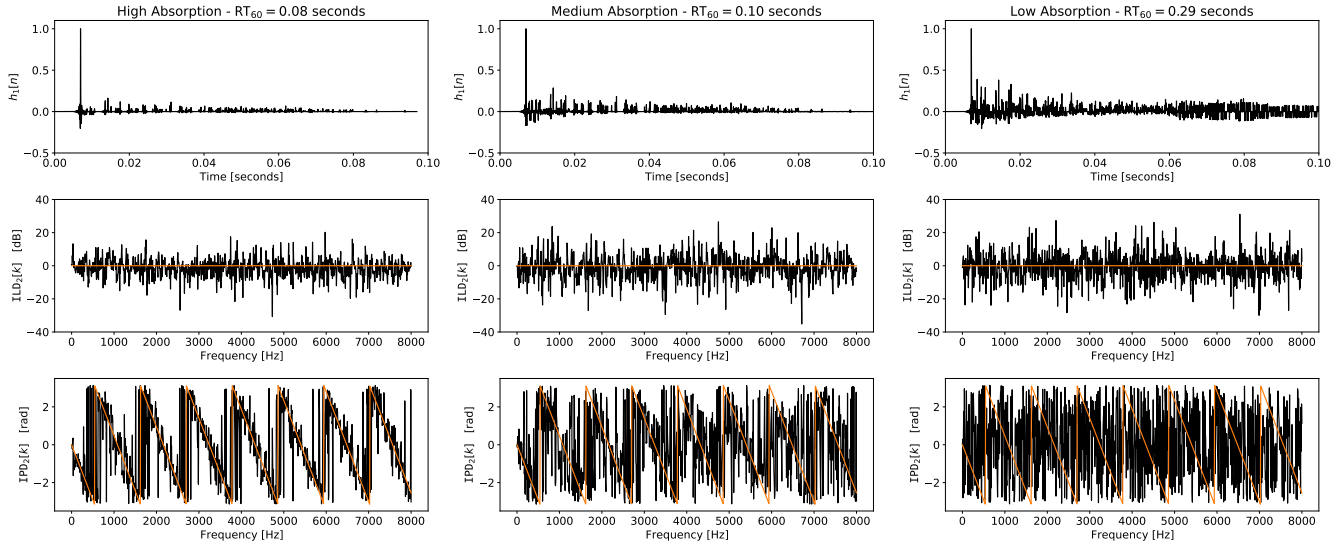


FIGURE 3.8: RIR, ILD and IPD corresponding to the pair of synthetic impulse responses of Figure 2.10 for different absorption conditions. Orange lines denote the theoretical far- and free- field ILD and IPD as defined by the relative steering vectors of Eq. (3.24)

Part II

ACOUSTIC ECHO RETRIEVAL

BIBLIOGRAPHY

4

Acoustic Echo Estimation

> "*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 this chapter formalized fundamental concepts of audio signal processing such as signal, mixtures and noise § 3.1 in the time domain. In § 3.2 we will presents the signal representation that we will use throughout the entire thesis: the STFT domain. Finally, after assuming the narrowband approximation, in § 3.3 some important models for the RIR are described.

Giving the frequency domain model of the RIR defined in the previous chapters,

$$\hat{h}[k] = \sum_{r=0}^{K} \tag{4.1}$$

the AER problem consists in estimating the echo timings $\{\tau_r\}$ and attenuations $\{\alpha_r\}$. The term AER is not typical in the audio signal processing community and it can be seen as instance of channel estimation problem or time of arrival problems.

4.1 as (sparse) RIR estimation

def estimation of the whole channel acoustic channel

methods Signal know vs. unknown. statistical methods vs. blind method

4.1.1 *Other echo-related parameters estimation*

▶ TDOA estimation

def estimation of the the difference of the direct path

methods Cross-correlation

▶ Echo Density Estimation

▶ RT_{60} and DRR estimation

4.2 Acoustic Echo Estimation is

- Acoustic Echo Retrieval definition

- Acoustic Echo Retrieval scope and placement in the signal processing pipeline
- Acoustic Echo Retrieval characteristic

4.3 Echoes in the Time, Frequency and Cepstral domains

- Time domain processing
- Frequency domain processing
- Correlation processing
- Cepstral processing

4.4 Related Works

4.4.1 *Active vs. Passive echoes estimation*

4.4.2 *Knowledge-based vs. Data-driven*

- Knowledge-driven (Physic-driven)
  - Channel (RIR) estimation and Echoes pruning - Crooco and Dokmanic
  - TDOA estimation (multipath) - Benesty
  - Spikes Retrieval - Condat
- Data-driven
  - GLLiM
  - Deep Leaning echo estimation

4.4.3 *end-2-end vs 2-steps approaches*

AER
eRTF + AER
Pruning methods

4.5 Related Works

summarize Crocco's presentation

4.5.1 *AER as a RIR Estimation problem*

TX signal: known vs. not known
TX signal not known: statistical methods and blind methods

4.5.2 *AER as a Spike Estimation problem*

4.5.3 *Virtually-supervised and Data Augmentation*

4.6 Data and Metrics

4.6.1 *Spike-based metrics*

Bibliography

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, 24).

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, 23).

Bal, Guillaume (2012). "Introduction to inverse problems". In: *Lecture Notes-Department of Applied Physics and Applied Mathematics, Columbia University, New York* (cit. on p. 3).

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. 25).

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. 17).

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 p. 29).

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

Fourier, Jean Baptiste Joseph (1822). *Théorie analytique de la chaleur*. F. Didot (cit. on p. 32).

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 p. 37).

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. 37).

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 pp. 25, 26).

Habets, Emanuel AP (2006). "Room impulse response generator". In: *Technische Universiteit Eindhoven, Tech. Rep* 2.2.4, p. 1 (cit. on pp. 21, 24).

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. 22).

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. 26).

Kitic, Srdan (2015). "Cosparse regularization of physics-driven inverse problems". PhD thesis. Rennes 1 (cit. on pp. 2, 4).

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. 17).

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

Kuttruff, Heinrich (2016). *Room acoustics*. CRC Press (cit. on pp. 8, 16–18, 20).

Müller, Meinard (2015). *Fundamentals of Music Processing*. Springer Verlag. ISBN: 978-3-319-21944-8 (cit. on p. 36).

Oppenheim, Alan V (1987). *Signals and Systems: An Introduction to Analog and Digital Signal Processing*. MIT Center for Advanced Engineering Study (cit. on p. 36).

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

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. 3).

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, 21, 22, 26).

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. 23).

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).

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. 30).

Thomas, Matthew Reuben (2017). "Wayverb: A Graphical Tool for Hybrid Room Acoustics Simulation". PhD thesis. University of Huddersfield (cit. on pp. 21, 23).

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. 34, 35).

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

Välimäki, Vesa, Julian Parker, Lauri Savioja, Julius O Smith, and Jonathan Abel (2016). "More than 50 years of artificial reverberation". In: *Audio engineering society conference: 60th international conference: dreams (dereverberation and reverberation of audio, music, and speech)*. Audio Engineering Society (cit. on pp. 22, 26).

Vincent, Emmanuel, Tuomas Virtanen, and Sharon Gannot (2018). *Audio source separation and speech enhancement*. John Wiley & Sons (cit. on pp. 29, 36).

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. 25).

Watson, LT, JA Ford, and M Bartholomew-Biggs (2001). *Nonlinear Equations and Optimisation*. Vol. 4. Elsevier (cit. on p. 2).

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. 27).

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. 34).