Can One Hear the Walls of a Room? Physics- and Data-Driven Inverse Methods for Acoustic Signal Processing

Antoine Deleforge

UNIVERSITÉ DE STRASBOURG

HABILITATION À DIRIGER LES RECHERCHES

Can One Hear the Walls of a Room?
Physics- and Data-Driven Inverse Methods for Acoustic Signal Processing

---

Antoine DELEFORGE

A thesis submitted in fulfillment of the requirements
for the degree of Habilitation à Diriger les Recherches

Inria branch of the University of Strasbourg
MACARON Team

Jury Members

- Roland Badeau, Professor, Dept. Image Données Signal, Télécom Paris *(Reviewer)*
- Enzo de Sena, Professor, Instit. of Sound Recording, Univ. of Surrey *(Reviewer)*
- Toon van Waterschoot, Professor, Dept. Electrical Engineering, KU Leuven *(Reviewer)*
- Simon Doclo, Professor, Dept. of Medical Physics and Acoustics, Univ. of Oldenburg *(Examiner)*
- Eric Bavu, Professor, Dept. Mécanique des Structures et Systèmes Couplés, Cnam Paris *(Examiner, President)*
- Philippe Helluy, Professor, Instit. Recherche Math. Avancée, Univ. of Strasbourg *(Examiner, Guarantor)*

The text of this thesis was
entirely human-generated.

To Denise, Valentin and Amelie
The sparks of joy dancing in my heart
I love you.

Résumé

Fermez les yeux, tapez dans vos mains. Entendez-vous la forme de la pièce? Le sol est-il fait de carrelage, ou de moquette? L'aspect *perceptif* de ces questions préservera son mystère pour le moment, et sera laissé comme un défi intrigant d'auto-expérimentation pour le lecteur. En attendant, le cœur de cette thèse (Partie II) synthétise un cheminement de recherche qui a débuté il y a huit ans, en essayant d'attaquer ces questions comme un problème *d'ingénierie*. Plus précisément, étant donnés des enregistrements microphoniques d'une source sonore dans une pièce, que peut-il être dit géométriquement et acoustiquement non pas sur la *source*, mais sur la *pièce*? Formaliser cette énigme lèvera le voile sur un riche et fascinant univers de problèmes inverses, à la croisée de l'informatique, des mathématiques et de la physique, dont la plupart restent ouverts à ce jour.

Au-delà de l'irrésistible curiosité scientifique causée par l'apparente simplicité et la nature fondamentale de ces questions, obtenir des progrès sur celles-ci pourrait bénéficier à de nombreuses et diverses applications: simplifier et améliorer le diagnostic acoustique des salles, rendre la réalité augmentée plus immersive, affiner la reproduction audio spatiale, ou encore améliorer le traitement des signaux audio pour la visioconférence, les enceintes connectées et les aides auditives. Nous présenterons ici une série de contributions algorithmiques à ce domaine, tirant profit d'outils venant du traitement du signal, de l'optimisation, et de l'apprentissage automatique, reposant à la fois sur des modèles principalement guidés par la *physique* et des modèles principalement guidés par des *données simulées*. Au passage, nous affinerons notre compréhension de la faisabilité et du caractère bien posé des problèmes associés, nous progresserons sur le développement de modèles directs et inverses trouvant un juste milieu entre complexité et réalisme, et nous identifierons certains des mécanismes sous-jacents à la généralisabilité de tels modèles aux données réelles. Nous rapporterons des résultats expérimentaux encourageant sur l'estimation des dimensions, du volume, de la surface, de l'absorption des réflecteurs et du temps de réverbération d'une salle, à partir de mesures de réponses impulsionnelles ou d'enregistrements audio, avec ou sans l'aide de connaissances géométriques. Nous étudierons également l'intérêt potentiel de connaître et d'exploiter de telles quantités pour des tâches allant au-delà de l'acoustique des salles, comme la localisation et la séparation de sources sonores. La partie principale de la thèse sera conclue en offrant des directions futures de recherche dans le domaine. Pour finir, la partie III présente de courts résumés d'autres contributions de l'auteur au domaine plus large du traitement du signal audio au cours des douze dernières années.

Abstract

Close your eyes, clap your hands. Can you hear the shape of the room? Is the floor made of tiles or carpet? The *perceptual* side of these questions will preserve its mystery for now, and will be left as an intriguing self-experimental challenge to the reader. Meanwhile, the core of this thesis (Part II) synthesizes a research journey that started eight years ago, attempting to tackle them as an *engineering* problem. Namely, given microphone recordings of a sound source in a room, what can be said geometrically and acoustically not about the *source*, but about the *room*? Formalizing this puzzle will give rise to a rich and fascinating network of inverse problems, at the crossroad of computer science, mathematics and physics, most of which remain open to date.

Beside the sheer scientific curiosity triggered by the seemingly simple and fundamental nature of these questions, making progress on them could benefit a number of applications, from simplifying and refining the acoustic diagnosis of rooms, to making audio augmented reality more immersive, to enhancing spatial audio reproduction, to improving the processing of indoor audio signals for teleconferencing, smart devices and hearing aids. We will present a series of algorithmic contributions to this field, leveraging tools from signal processing, optimization and machine learning, using both models primarily driven by *physics* and models primarily driven by *simulated data*. Along the way, some progress will be made in framing the feasibility and well-posedness of these problems, in developing forward and inverse models that strike a balance between complexity and realism, and in understanding the mechanisms that underpin the generalizability of such models to real-world data. Some promising experimental results will be reported on estimating the dimensions, volume, surface area, reflectors' absorption and reverberation time of a room from either impulse response measurements or audio recordings, with or without the aid of geometrical knowledge. We will also investigate the potential benefit of knowing and exploiting such quantities in tasks beyond room acoustics, such as localizing and separating sound sources. The main part of the thesis will be concluded by offering directions for future research in the field. Then, Part III presents short summaries of other contributions by the author to the broader field of audio signal processing over the past twelve years.

Acknowledgments

The research journey that led to the content of this thesis, with all the joy and excitement that came with it, up to the very writing of this manuscript, would not have been possible without the support and help of many people. First, I would like to thank the few people that I consider as my mentors, and whose guidance and role models will continue to inspire me as a researcher for the years to come. Radu Horaud and Florence Forbes, of course, my PhD advisors, to whom I am eternally grateful for their trust, support, and for introducing me to the indescribable joy of scientific research. Walter Kellermann, who entrusted me with increasing responsibilities and leadership during my postdoctoral years at FAU. Rémi Gribonval, the calm and brilliant leader of the PANAMA team, who gently accompanied me and opened up my horizon during my first two years as a tenured Inria researcher. Emmanuel Vincent, for his incredibly human and selfless mentorship and his inspiring genius, who taught me more than I could ever say during my 5 years with MULTISPEECH.

In the other direction of the tree, I am extremely grateful to all the master and PhD students, postdocs and engineers that I had the pleasure to officially or unofficially co-supervise over the years: Alexander Schmidt, Clément Gaultier, Saurabh Kataria, Diego Di Carlo, Helena Peić Tukuljac, Martin Strauss, Paul Mordel, Victor Miguet, Nicolas Keriven, Clément Elvira, Manuel Pariente, Corto Bastien, Joris Cosentino, Alexis Dieu, Usama Saqib, Krist Kostallari, Prerak Srivastava, Stéphane Dilungana, Khaoula Chahdi, Robin San Roman, Tom Sprunck, Marina Krémé, Louis Bahrman, Tanmay Bhonsale, Jérémy Pawlus, Barnabé Miesch, Jean-Daniel Pascal and Ilaria Fichera. Sharing with them (sometimes unreasonably long!) brainstorming sessions over physical or virtual whiteboards during which we often felt like geniuses and then complete dummies within the span of 10 minutes, has been perhaps what makes me love my job the most. To be clear, they are the ones who should be acknowledged and congratulated for the hard technical work and results behind most of this thesis' content. I thank them for their perseverance, enthusiasm, trust, and I am humbly happy that I could be even a small part of their own research journeys.

I also warmly thank the many fellow researchers and co-supervisors I had the pleasure to collaborate with since I obtained my tenured position in 2016. Yann Traonmilin, Angélique Drémeau and Paul Magron, for sharing my nerdy and engulfing passion for phase. Antoine Liutkus, for being my source separation mentor, my friend, and the central social hub of audio folks during the golden age of ICASSP. Nancy Bertin, Robin Scheibler and Ivan Dokmanic, for believing in the power of acoustic echoes and being great beer companions, two seemingly correlated virtues. Sharon Gannot for his kindness and openness, for being a wonderful conference friend, and for his encyclopedic knowledge on audio signal processing and everything else. Romain Serizel for being my awesome (pun intended) office mate for 5 years, and for generously hosting me countless nights in Nancy with his wonderful family. Olivier Warusfel for kindly welcoming me as a partner of the HAIKUS project, an opportunity that led up to many of this thesis' results, together with François Ollivier who is always here to help in case of experimental or theoretical acoustic needs. Cédric Foy for converting me to acoustics and being a dream colleague and friend over the past 6 years, constantly bringing the mood up, ready to overcome any obstacle. Sylvain Faisan for never holding back smart and hard questions, a great fuel for creative research. Yannick Privat for his award-winning mathematical wisdom and for Stone et Charden. Archontis Politis who visited me in Nancy for an incredibly fruitful 6-week research collaboration, made possible by his friendliness, talent and openness. I would also like to give a general

thank you to the countless research and non-research colleagues at FAU, Inria Rennes, Inria Nancy, Télécom Physique, Cerema and IRMA, who unfortunately cannot all fit on this list but contributed in their own way to making my day-to-day at work such a humane and lovable experience.

I am particularly grateful to Roland Badeau, Toon van Waterschoot, Enzo de Sena, Simon Doclo and Eric Bavu who kindly accepted to join the jury of my habilitation thesis, with special thanks to the three reviewers for their heartwarmingly positive and useful feedback on the manuscript. This is a true honor, as I always looked up to them and their major contributions to the field. Although we have never collaborated as of yet, I am hopeful this will change soon. I also thank Philippe Helluy for trusting me and being my guarantor.

Lastly, I would like to thank my family and friends for their continuous emotional support, that have kept my head up and leveled throughout the ups and downs of research life. My heart goes especially to my parents and sisters, who have never ceased to send me positive and loving encouragements, despite them not understanding most of what I am doing. Finally, I am forever grateful for the enormous support, compassion, strength, love and help from my wife Denise. She carried an invisible but significant part of this thesis' work on her shoulders over the years, without ever asking anything in return. Finishing this thesis would simply not have been possible without her. I dedicate this thesis to her, with all my love, admiration, and thankfulness.

Contents

Part I

Preliminaries

1 A Quick Guide to this Thesis

1.1 What?

This thesis provides an expository overview of the research I conducted after obtaining my PhD, from the year 2014 to early 2025. The full list of my associated publications is given in Chapter 4. The text is aimed at researchers broadly familiar with audio, acoustics, machine learning, and signal processing. The core of the thesis, in Part II, revolves around one intriguing question that increasingly guided and motivated my research over the years:

Can One Hear the Walls of a Room?

This part presents a unified, non-chronological synthesis of my contributions to various facets of this question, and concludes with a program for future research in Chapter 10. Complementarily, Part III offers short summaries of other research topics I explored over the period.

1.2 Who?

The contributions presented here are the result of collective efforts, and would not have existed without the many talented collaborators I had the pleasure and privilege to work with. A list of the supervisees, co-supervisors and other collaborators involved in various aspects of this work can be found in Chapter 2, while Chapter 3 provides an overview of the funded projects I contributed to. **I personally played a leading role in initiating and guiding the research covered in Part II**, having written the associated project proposals and PhD topics, and consequently acting as a primary supervisor for the involved students. Meawhile, the students were primarily responsible for the code and experiments, and countless invaluable technical insights and ideas are also due to them and to my collaborators. My involvement in the research contributions covered in Part III is more varied, and is specified at the bottom of each corresponding page.

1.3 Where?

The research was conducted in four different environments. I was first employed as a post-doctoral fellow at the chair of Multimedia Communications and Signal Processing of the Friedrich-Alexander-University with professor Walter Kellerman (*Erlangen, Jan. 2014 - Dec. 2015*). I then obtained a permanent associate researcher position at Inria, which I started in the PANAMA team (*Rennes, Jan. 2016 - Mar. 2018*), focused on mathematical models and algorithms for audio signal processing. I then joined the MULTISPEECH team (*Nancy, Apr. 2018 - Sep. 2023*), specialized in multimodal speech and language processing and synthesis. I finally joined the MACARON team (*Strasbourg, Oct. 2023 - present*), whose aim is to hybridize machine learning and numerical methods for physics modeling.

1.4 How?

How to present one's decade of research efforts in a digestible and (hopefully) valuable form? On the one hand, an exhaustive compilation of disparate summaries of articles would be tedious and offer little insight. On the other hand, attempting to fit all of one's contributions into a *Grand Unified Theory*, though appealing at first glance, would inevitably seem artificial and cumbersome, sacrificing clarity while abstracting away much of the substance. In reality, research journeys are serendipitous and rarely follow a linear, predictable path.

Nonetheless, in my case, an overarching question emerged organically from my research around 2017, gradually gaining prominence, until becoming one of its main drivers. **What can microphone recordings reveal about the *environment* in which sound propagates, as opposed to the traditionally studied *semantic* content of sound sources such as speech or music?** Or in short: *Can one hear the walls of a room*? This question still fascinates me today, and I have chosen to make it the core of this thesis (Part II), presenting contributions spanning 12 of my publications in a newly structured and unified narrative. A number of research directions for this area, some of them sparked by the introspective process of writing this very thesis, are presented in Chapter 10. To preserve fluidity, the document does not provide extensive literature reviews, which the interested reader may find in my articles [S26], [S24] or [S36]. Similarly, derivations are kept to a minimum and proofs are omitted, to focus on conceptual insights, key results, and their broader implications instead.

Part III summarizes 6 other research topics I explored since 2014, spanning 24 publications. I set to myself to present each of these topics following a constrained one-page format. Though these *"research snippets"* can be read independently, they share a number of common themes underlying my research: linear and non-linear inverse problems, gridless optimization, the hybridation of physics- and data-driven approaches, generative models, and the confrontation of such techniques to real-world data.

1.5 Notations, Conventions, Useful Facts

- References to publications I co-authored are denoted by [S1], [S2], etc. (full list in Chap.4).
- References to other publications are denoted by [1], [2], etc. (full list at the end of the thesis).
- The speed of sound is denoted by c (≈ 343 m $\cdot$ s^{-1} in air at 20°C). The density of air is denoted by ρ (≈ 1.204 kg $\cdot$ m^{-3} at 20°C).
- Scalars, vectors, matrices and sets are respectively denoted by $u, \boldsymbol{u}, \mathrm{U}$ and $\mathcal{U}$. Transposes and Hermitian transposes are respectively denoted by $(\cdot)^\top$ and $(\cdot)^\mathrm{H}$. The imaginary unit is denoted by j.
- A discrete interval $\{n, n+1, \dots, m-1, m\}$ is denoted by $[\![n, m]\!]$. Discrete (multi-)indexed sets are denoted by, *e.g.*, $\{x_{m,n}\}_{m,n=1}^{M,N}$.
- Functions in the continuous-time, discrete-time, continuous-frequency and discrete-frequency domains are respectively denoted by $u(t), u[n], \hat{u}(\omega)$ and $\hat{u}[f]$. By convention, the discrete frequency index f belongs to a set of F linearly-spaced frequencies $\mathcal{F} = \{f_1, \dots, f_F\}$ expressed in Hertz. When a discrete-time function $u[n]$ is also *finite-time*, *i.e.*, only defined for $n \in [\![0, N-1]\!]$, we may also represent it as an N-dimensional vector $\boldsymbol{u} = [u_0, \dots, u_{N-1}]^\top$.

- The continuous-time Fourier transform and its inverse are given by:

$$\hat{u}(\omega) = \int_{-\infty}^{+\infty} u(t) e^{-j\omega t} dt \quad \text{and} \quad u(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{u}(\omega) e^{j\omega t} d\omega. \tag{1.1}$$

- The discrete Fourier transform (DFT) and its inverse for a sampling frequency f_s are given by:

$$\hat{u}[f] = \sum_{n=0}^{N-1} u[n] e^{-2\pi j n f / f_s} \quad \text{and} \quad u[n] = \frac{1}{N} \sum_{f \in \mathcal{F}} \hat{u}[f] e^{2\pi j n f / f_s} \tag{1.2}$$

  where $\mathcal{F} = \{k f_s / N;\ k \in [\![-\lfloor (N-1)/2 \rfloor, \lfloor N/2 \rfloor]\!]\}$ contains N elements.

- The continuous-time and discrete-time convolutions are respectively denoted by:

$$(u * v)(t) \overset{\text{def}}{=} \int_{-\infty}^{+\infty} u(\tau) v(t-\tau) d\tau \quad \text{and} \quad (u \circledast v)[n] \overset{\text{def}}{=} \sum_{m=-\infty}^{+\infty} u[m] v[n-m]. \tag{1.3}$$

  Both operations are associative and commutative. We will often approximate the second one by *finite* discrete-time convolutions, *i.e.*, by truncating the infinite sum in (1.3) to M values. This approximation is only exact if the discrete-time functions are M-periodic (which then gives discrete *circular* convolution), or when they are finitely supported and appropriately zero-padded.

- The *convolution theorem* in continuous and finite-discrete time gives:

$$\widehat{u * v}(\omega) = \hat{u}(\omega) \odot \hat{v}(\omega) \quad \text{and} \quad \widehat{u \circledast v}[f] \approx \hat{u}[f] \odot \hat{v}[f] \tag{1.4}$$

  where $\odot$ is used for the pointwise and elementwise (Hadamard) products. Importantly, the theorem *only approximately holds* in finite time, and its accuracy increases with N. To get an equality, one either needs to employ the discrete-(infinite)-time Fourier transform (DTFT), or the discrete circular convolution, or appropriate zero-padding if the signals are finite-time.

- The impulse response of the *ideal low-pass filter* with cut-off frequency $f_s/2$ is:

$$\kappa_{f_s/2}(t) \overset{\text{def}}{=} \operatorname{sinc}(f_s t) \overset{\text{def}}{=} \frac{\sin(\pi f_s t)}{\pi f_s t}. \tag{1.5}$$

- Let $v(t)$ be a band-limited "source" signal with maximum frequency $f_s/2$ and $u(t)$ be any "kernel" function or "filter". If we *sample* these functions using

$$v[n] = v(n/f_s) \quad \text{and} \quad u[n] = (\kappa_{f_s/2} * u)(n/f_s), \tag{1.6}$$

  then a useful fact is that:

$$(u * v)(n/f_s) = (u \circledast v)[n] \qquad \text{[3, Proposition 2].} \tag{1.7}$$

  In other words, "the sampled continuous-time convolution is the discrete-time convolution of the sampled functions". Importantly, this result only holds approximately in finite time, and its accuracy increases with N.

- $\boldsymbol{r} \in \mathbb{R}$ is used to denote a position in 3D space that can be expressed either in Cartesian coordinates (x, y, z) or in spherical coordinates (θ, ϕ, r) where $\theta \in [0, 2\pi]$ denotes the azimuth, $\phi \in [-\pi/2, \pi/2]$ the elevation and $r = \|\boldsymbol{r}\|_2$ the radial distance.
- The Dirac measures in time and in 3D space are respectively denoted by $\delta(t)$ and $\delta(\boldsymbol{r})$.

Here is a list of the main acronyms used in this thesis and their definition.

- **BEADS**: Bayesian Estimation of AuDio source priors for source Separation (Algorithm, Snip. 13)
- **BLASTER**: Off-the-Grid Sparse Cross-Relation Method (Algorithm, Sec. 9.2.3)
- **CNN**: Convolutional Neural Network
- **DNN**: Deep Neural Network
- **DOA**: Direction of Arrival
- **DFT**: Discrete Fourier Transform
- **DTFT**: Discrete-Time Fourier Transform
- **MIRAGE**: MIcrophone-aRray AuGmentation with Echoes (Algorithm, Sec. 9.2.4 and 9.3.1)
- **MULAN**: MULtichannel ANnihilation (Algorithm, Sec. 9.2.2)
- **NMF**: Nonnegative Matrix Factorization
- **RIR**: Room Impulse Response
- $\mathbf{RT_{60}}$: Reverberation Time (60 dB decay)
- **RTF** or **ReTF**: Relative Transfer Function
- **SNR**: Signal-to-Noise Ratio
- **STFT**: Short-Time Fourier Transform
- **TDOA**: Time Difference of Arrival
- **TOA**: Time of Arrival
- **VAE**: Variational Auto-Encoder

2 Supervisions and Collaborations

Below is the list of people I supervised over the 2014-2025 period and that were involved in parts of this thesis, together with their contract period, their research topic, their funding source, their co-supervisors, the estimated effective co-supervision ratios, and the contributed parts. Sec. 2.4 provides a list of additional significant collaborators of this work over the same period.

2.1 Postdocs and Engineers

- **Joris Cosentino** (Nov. 2020 - Oct. 2022), research engineer on the topic "Generalized Speech Enhancement by Supervised Learning" funded by ADT PEGAUSUS (See Sec. 3). I was the sole supervisor of Joris (**100%**). Contributions to Snippet 14.
- **Marina Krémé** (Mar. 2022 - Sep. 2023), postdoc on the topic "Repairing Audio Signals using Compact Phase-Aware Models" funded by ANR JCJC DENISE (See Sec. 3). Co-supervision shared by **Paul Magron** (40%) and myself (**60%**). Contributions to Snippet 12.

2.2 PhD Students

- **Diego Di Carlo** (Oct. 2017 - Dec. 2020), on the topic "Echo-Aware Signal Processing for Audio Scene Analysis" funded by a CORDI-S Inria grant. Co-supervision shared by **Nancy Bertin** (40%) and myself (**60%**). Contributions to Sec. 8.2.1 and Chap. 9.
- **Manuel Pariente** (Oct. 2018 - Sep. 2021), on the topic "Implicit and Explicit Phase Modeling in Deep-Learning-Based Source Separation" funded by a grant from École Normale Supéreure. Co-supervision shared by **Emmanuel Vincent** (50%) and myself (**50%**). Contributions to Snippets 13 and 14.
- **Prerak Srivastava** (Oct. 2020 - Sep. 2023), on the topic "Hearing the Walls of a Room using Virtually Supervised Learning" funded by ANR PRC HAIKUS (See Sec. 3). Co-supervision shared by **Emmanuel Vincent** (40%) and myself (**60%**). Contributions to Sec. 8.1.3, 8.3 and Snippet 11.
- **Stéphane Dilungana** (Oct. 2020 - Nov. 2023), on the topic "Machine Learning and Optimization for Estimating the Acoustical Properties of a Room from Audio Signals" funded by AEx ACOUST.IA (See Sec. 3). Co-supervision shared by **Sylvain Faisan** (34%), **Cédric Foy** (33%) and myself (**33%**). Contributions to Sec. 7.1 and 8.2.
- **Tom Sprunck** (Nov. 2021 - Dec. 2024), on the topic "Can One Hear the Shape of a Room? Room Geometry Reconstruction from Acoustic Measurements using Super-Resolution and Shape Optimization" funded by ANR JCJC DENISE (See Sec. 3). Co-supervision shared by **Yannick Privat** (34%), **Cédric Foy** (33%) and myself (**33%**). Contributions to Sec. 7.2.

- **Robin San Roman** (Jun. 2021 -), on the topic "Self supervised disentangled representation learning of audio data for compression and generation" funded by a CIFRE contract with Meta. Co-supervision shared by **Romain Serizel** (30%), **Yossi Adi** (10%), **Alexandre Défossez** (30%) and myself (**30%**). Contributions to Snippet 16.

- **Jean-Daniel Pascal** (Oct. 2024 -), on the topic "Facilitating Room Acoustic Diagnosis Using Signal Processing and Machine Learning" funded by a CEREMA grant. Co-supervision shared by **Cédric Foy** (40%) and myself (**60%**). Contributions to Sec. 10.2.3.

2.3 Master Students

- **Alexander Schmidt** (Apr. 2015 - Oct. 2015), on the topic "Ego-noise reduction for the robot NAO using dictionary learning and motor data". Co-supervision shared by **Walter Kellerman** (20%) and myself (**80%**). Contributions to Snippet 15.

- **Martin Strauss** (Oct. 2017 - Dec. 2017), on the topic "Drone-Embedded Sound Source Localization for Search and Rescue". I was the sole supervisor of Martin (**100%**). Contributions to Snippet 15.

- **Joris Cosentino** (Fev. 2020 - Aug. 2020), on the topic "Filterbank designs for end-to-end audio source separation". Co-supervision shared by **Manuel Pariente** (50%) and myself (**50%**). Contributions to Snippet 14.

- **Louis Bahrman** (Apr. 2022 - Aug. 2022), on the topic "Signal Inpainting from Fourier Magnitudes". Co-supervision shared by **Marina Kreme** (33%), **Paul Magron** (33%) and myself (**34%**). Contributions to Snippet 12.

2.4 Other Collaborators

In addition to all the people listed above, here is a list of collaborators that significantly contributed to some parts of this thesis, in chronological order of first collaboration: **Heinrich Löllmann** (Snip. 15), **Hendrick Barfuss** (Snip. 15), **Stefan Meier** (Snip. 15), **Sharon Gannot** (Snip. 11, Sec. 9.1), **Florence Forbes** (Snip. 11), **Clément Gaultier** (Snip. 11), **Saurabh Kataria** (Snip. 11), **Yann Traonmilin** (Snip. 12), **Angélique Drémeau** (Snip. 12), **Helena Peić Tukuljac** (Sec. 9.2.2), **Rémi Gribonval** (Sec. 9.2.2-9.2.3), **Nicolas Keriven** (Snip. 13), **Antoine Liutkus** (Snip. 13), **Christian Rohlfing** (Snip. 13), **Robin Scheibler** (Sec. 9.3.2), **Ivan Dokmanic** (Sec. 9.3.2), **Pol Mordel** (Snip. 15), **Victor Miguet** (Snip. 15), **Clément Elvira** (Sec. 9.2.3), **Samuele Cornell** (Snip. 14), **Usama Saqib** (Sec. 15), **Jesper Rindom Jensen** (Sec. 15), **Pinchas Tandeitnik** (Sec. 9.1), **Archontis Politis** (Snip. 11) and **Pierre Fernandez** (Snip. 16).

3 Projects

Below is the list of funded projects I have been involved in over the 2014-2025 period, together with their dates, descriptions, partners, and my role in them.

FP7 STREP EARS (Jan. 2014 - Dec. 2016)

Description: The European project EARS (Embodied Audition for Robots) explored new algorithms for enhancing the auditive capabilities of humanoid robots.
Role: Postdoctoral fellow (Jan. 2014 - Dec. 2015).
Partners: University of Erlangen-Nuremberg (prof. Walter Kellermann, coordinator), Imperial College of London (prof. Patrick Naylor), Ben-Gourion University of the Negev (prof. Boaz Rafaely), Humboldt University in Berlin (Prof. Verena Hafner), and Aldebaran Robotics (Dr. Rodolphe Gelin).

ANR PRC HAIKUS (Dec. 2019 - Dec. 2024)

Description: The national research project HAIKUS (Artificial Intelligence applied to augmented acoustic Scenes) aimed at developing new methods for immersive audio-augmented reality.
Role: Local Coordinator
Partners: IRCAM (Olivier Warusfel, coordinator), Institut Jean Le Rond d'Alembert (François Ollivier).
Supervisee involved: Prerak Srivastava (PhD student).

AEx ACOUST.IA (Oct. 2020 - Sep. 2023)

Description: The Inria-funded Exploratory Action ("Action Exploratoire") ACOUST.IA aimed at simplifying and improving the acoustic diagnosis of rooms thanks to artificial intelligence and audio signal processing.
Role: Principal Investigator.
Partner: Cédric Foy (UMRAE, Cerema, Univ. Gustave Eiffel).
Supervisee involved: Stéphane Dilungana (PhD student).

ADT PEGAUSUS (Nov. 2020 - Oct. 2022)

Description: The Inria-funded technological development grant PEGASUS aimed at producing and maintaining a versatile software platform for end-to-end generalized speech enhancement by supervised deep learning.
Role: Principal Investigator.
Supervisee involved: Joris Cosentino (Engineer).

ANR JCJC DENISE (Mar. 2021 - Dec. 2024)

Description: The national research project DENISE (tackling hard problems in audio with Data-Efficient Non-linear InverSe mEthods) developed new inverse methods for room geometry estimation and phase-aware audio inpainting.
Role: Principal Investigator.
Supervisees involved: Tom Sprunck (PhD student), Marina Krémé (Postdoc).

4 List of Co-Authored Publications

Below is the list, in chronological order, of pre-prints, book chapters, journal and conference articles covered in this thesis. This represents most of my scientific production over the 2014-2025 period. The list excludes works that were done as part of, or as direct follow-ups on, my PhD thesis. It also exludes a couple of publications in which I played a relatively minor role, as well as shorter papers, communications and abstracts (<4 pages). The contributions highlighted in blue are the ones covered in Part II, while the others are more briefly presented in the research snippets of Part III.

[S1] H. W. Löllmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Speech Communication; 11. ITG Symposium*, VDE, 2014, pp. 1–4.

[S2] A. Deleforge, S. Gannot, and W. Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *23rd European Signal Processing Conference (EUSIPCO)*, IEEE, 2015, pp. 419–423.

[S3] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 355–359.

[S4] A. Deleforge and F. Forbes, "Rectified binaural ratio: A complex t-distributed feature for robust sound localization," in *24th European Signal Processing Conference (EUSIPCO)*, IEEE, 2016, pp. 1257–1261.

[S5] A. Schmidt, A. Deleforge, and W. Kellermann, "Ego-noise reduction using a motor data-guided multichannel dictionary," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2016, pp. 1281–1286.

[S6] A. Deleforge and Y. Traonmilin, "Phase unmixing: Multichannel source separation with magnitude constraints," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 161–165.

[S7] A. Drémeau and A. Deleforge, "Phase retrieval with a multivariate von mises prior: From a bayesian formulation to a lifting solution," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 4596–4600.

[S8] C. Gaultier, S. Kataria, and A. Deleforge, "VAST: The virtual acoustic space traveler dataset," in *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Springer, 2017, pp. 68–79.

[S9] S. Kataria, C. Gaultier, and A. Deleforge, "Hearing in a shoe-box: Binaural source position and wall absorption estimation using virtually supervised learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 226–230.

[S10] N. Keriven, A. Deleforge, and A. Liutkus, "Blind source separation using mixtures of alpha-stable distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 771–775.

[S11] A. Liutkus, C. Rohlfing, and A. Deleforge, "Audio source separation with magnitude priors: The BEADS model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 56–60.

[S12] H. Peic Tukuljac, A. Deleforge, and R. Gribonval, "MULAN: A blind and off-grid method for multichannel echo retrieval," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[S13] R. Scheibler, D. Di Carlo, A. Deleforge, and I. Dokmanic, "Separake: Source separation with a little help from echoes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 6897–6901.

[S14] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for UAV-embedded sound source localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 1–8.

[S15] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, "Audio-based search and rescue with a drone: Highlights from the IEEE signal processing cup 2019 student competition [SP competitions]," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 138–144, 2019.

[S16] A. Deleforge, A. Schmidt, and W. Kellermann, "Audio-motor integration for robot audition," in *Multimodal Behavior Analysis in the Wild*, Elsevier, 2019, pp. 27–51.

[S17] D. Di Carlo, A. Deleforge, and N. Bertin, "MIRAGE: 2D source localization using microphone pair augmentation with echoes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 775–779.

[S18] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[S19] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[S20] A. Deleforge, "Drone audition for search and rescue: Datasets and challenges," in *QUIET DRONES International Symposium on UAV/UAS Noise*, 2020.

[S21] D. Di Carlo, C. Elvira, A. Deleforge, N. Bertin, and R. Gribonval, "BLASTER: An off-grid method for blind and regularized acoustic echoes retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 156–160.

[S22] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6364–6368.

[S23] M. Pariente *et al.*, "Asteroid: The pytorch-based audio source separation toolkit for researchers," in *21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

[S24] D. Di Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dEchorate: A calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 39, 2021.

[S25] S. Dilungana, A. Deleforge, C. Foy, and S. Faisan, "Learning-based estimation of individual absorption profiles from a single room impulse response with known positions of source, sensor and surfaces," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, vol. 263, 2021, pp. 5623–5630.

[S26] C. Foy, A. Deleforge, and D. Di Carlo, "Mean absorption estimation from room impulse responses using virtually supervised learning," *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 1286–1299, 2021.

[S27] U. Saqib, A. Deleforge, and J. R. Jensen, "Detecting acoustic reflectors using a robot's egonoise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 466–470.

[S28] P. Srivastava, A. Deleforge, and E. Vincent, "Blind room parameter estimation using multiple multichannel speech recordings," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2021, pp. 226–230.

[S29] S. Dilungana, A. Deleforge, C. Foy, and S. Faisan, "Geometry-informed estimation of surface absorption profiles from room impulse responses," in *30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 867–871.

[S30] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Gridless 3D recovery of image sources from room impulse responses," *IEEE Signal Processing Letters*, vol. 29, pp. 2427–2431, 2022.

[S31] P. Srivastava, A. Deleforge, and E. Vincent, "Realistic sources, receivers and walls improve the generalisability of virtually-supervised blind acoustic parameter estimators," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[S32] L. Bahrman, M. Krémé, P. Magron, and A. Deleforge, "Signal inpainting from fourier magnitudes," in *31st European Signal Processing Conference (EUSIPCO)*, IEEE, 2023, pp. 116–120.

[S33] R. San Roman, Y. Adi, A. Deleforge, R. Serizel, G. Synnaeve, and A. Défossez, "From discrete tokens to high-fidelity audio using multi-band diffusion," *Advances in neural information processing systems (NeurIPS)*, vol. 36, pp. 1526–1538, 2023.

[S34] P. Srivastava, A. Deleforge, A. Politis, and E. Vincent, "How to (virtually) train your speaker localizer," in *24th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[S35] R. San Roman, P. Fernandez, A. Deleforge, Y. Adi, and R. Serizel, "Latent watermarking of audio generative models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025.

[S36] T. Sprunck, A. Deleforge, Y. Privat, and C. Foy, "Fully reversing the shoebox image source method: From impulse responses to room parameters," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1023–1033, 2025.

Part II

Hearing the Walls of a Room

5 Introduction

One of the main goals of audio signal processing research is to extract information from microphone recordings. Within this vast objective, by far the main objects of interest are *sources of sound* in the environment. Extracting and enhancing their *emitted signals* involves removing the undesirable effects caused by interfering sources, by the propagation medium, or by the microphones themselves, as covered by tasks as diverse as source separation, denoising, dereverberation, declipping or bandwidth extension. *Localizing* sound sources in the receiver's coordinate frame also form a long-standing research topic. Finally, extracting the *semantic content* of sources is the subject of entire subfields including automatic speech recognition, music information retrieval and the detection and classification of more general acoustic scenes and events. Throughout this thesis, we will move the spotlight away from the main character of sound sources, and will shine it instead on the *environment* in which sound propagates. More specifically, the focus will be on indoor environments, and hence on the estimation of the geometrical and acoustical properties of the *reflectors* composing the boundary of a room. Retrieving such information from microphone recordings will require understanding the full journey of sound, from its emission by sources through its interaction with air and reflectors along its path to its reception at microphones and its subsequent digitization. This brings us to our eponymous question: *Can One Hear the Walls*[1] *of a Room?*

This voluntarily vague phrasing underlies in fact a multitude of possible instantiations into tasks, each corresponding to substantially different problems. Are the emitted source signals known? Are there one or multiple sources? One or multiple microphones? What constraints are imposed on the room's geometry and devices' positions? Is the task to localize one reflector, multiple reflectors, or global information on the boundary shape such as its surface or volume? Do we seek angles, distances or both? In two or in three dimensions? What acoustical properties of reflectors should be retrieved? These questions have sparked significant interest from the audio and acoustic signal processing community over the past 15 years, giving rise to a rich, multi-faceted, if not fully unified literature, owing to the diversity of problem formulations. The fascination of researchers for such questions has in fact deeper historical roots, and may be traced at least as far back as the famous 1969 article of the mathematician Mark Kac, "*Can One Hear the Shape of a Drum?* [4][2]". But beyond this folklore aspect, theoretical and methodological advances in estimating geometrical and acoustical room parameters from audio signals may greatly benefit a number of applications, some of which are briefly reviewed in Sec. 5.1. Sec. 5.2 then provides a brief review of existing methodologies, while Sec. 5.3 summarizes our research contributions to the field and outline the remainder of this thesis.

[1] For convenience, throughout this thesis, we will refer to any of the boundary surfaces of a room as a "wall", including the floor and the ceiling.

[2] Kac's mathematical riddle was famously answered in the negative in 1992 thanks to the discovery of isospectral shapes by Gordon, Webb and Wolpert [5]. One may view this question as a 2D version of the 3D room geometry estimation problem addressed in Sec. 7.2 of this thesis, under Dirichlet instead of Neumann boundary conditions. Two key differences are that (i) the formulation of Kac does not consider discrete room impulse responses but the availability of the infinite set of discrete Laplace-Beltrami eigen-frequencies for these boundary conditions, and (ii) we only consider cuboid boundaries while the counterexample of [5] involves non-convex polygons. To our knowledge, the general question for "3D drums" remains an open mathematical problem.

5.1 Applications

Room acoustic diagnosis [6]. Before proposing solutions to improve the acoustic characteristics of an existing room, acoustic design departments must establish a diagnosis. The currently-used approach heavily relies on external knowledge such as the geometry of the room and the absorption profiles of identifiable materials as obtained from laboratory databases. These coarse initial estimates are then fed to an acoustic simulator, and are manually and iteratively adjusted so that the re-simulated sound field matches standardized acoustic measurements made *in situ*, such as the room's reverberation time in octave band. This long, costly and approximate process could be considerably improved by the development of inverse methods directly mapping audio measurements to acoustical and geometrical parameters of interest.

Audio augmented reality [7]. This field aims at seamlessly integrating virtual sound sources in the environment of a user wearing a hear-through headset. To make the experience plausible and immersive, perceptually relevant room parameters should be used to simulate the sources [7]. An attractive and practical way to obtain these parameters is directly from audio signals recorded by microphones embedded in the user's headset. In contrast with the previous point, information on the emitted source signals or geometrical knowledge cannot be assumed in this case.

Spatial audio recording and reproduction [8]. This long-standing topic dates as far back as the early 1930s, with the development of stereophonic techniques. The aim is to not only capture and restitute a soundfield at a single point in space, but also to preserve some of its *spatial* properties. This is made challenging by the limited number of microphones at acquisition time, and the limited number of loudspeakers at rendering time. Faithfully capturing the soundfield then necessitates to fit a highly compressed model to available measurements, which involves the implicit or explicit estimation of acoustical and geometrical room parameters, or quantities derive from them. Applications include live music recording, smart-TVs or immersive binaural synthesis. A long-term goal in this area would be the ability to transfer a sound scene captured with ad-hoc receivers in an ad-hoc room to a target room, by modifying its acoustical and geometrical parameters via post-processing of the recorded signals.

Acoustic heritage preservation and digital twinning [9–13]. A lesser known application that has recently gained interest is to preserve the cultural heritage of acoustic scenes, and in particular the acoustical properties of historical monuments [13]. While a larger number of calibrated measurements with dedicated apparatus can be assumed in this setting than in the previous point, faithfully capturing a soundfield from a sparse set of measurements remains a challenge [9]. A more general goal is the creation of *digital twins* for acoustic spaces [11]. Combining the four application fields listed so far, one could envision simulating in advance the impact of an acoustic renovation solution on a space, in a spatially immersive way.

Audio-aided indoor navigation [14, S27, 15]. Another application of interest is to exploit the microphones embedded in a robot, *e.g.*, a drone, and the sounds it emits (either passively or actively) to localize sound-reflecting obstacles, which may be helpful in low-visibility conditions [S27, 15]. A generalization of this is *acoustic simultaneous localization and mapping* (A-SLAM), as pioneered in [14].

Echo-aware audio signal processing [16, S24]. This is a relatively recent line of research, aiming at exploiting knowledge on the manifestation of early acoustic reflections at microphones, referred to as *echoes*, to improve methods in traditional audio signal processing tasks such as speech enhancement or sound source localization. This is particularly relevant for indoor applications where sources and microphones can be close to reflective surfaces, *e.g.*, hearing aids, teleconferencing, smart speakers or distributed, ad-hoc microphone arrays.

5.2 Scientific Context

We present below a brief review of the main existing methodologies to estimate geometrical and acoustical parameters of interest from acoustic measurements. Following the articulation of this thesis, we segment these methodologies into two categories, *physics-driven* approaches (Sec. 5.2.1) and *data-driven* approaches (Sec. 5.2.2). For the purpose of this introductory section, we will abstractly denote by $\boldsymbol{x}$ a vector containing input acoustic measurements (or features computed from them), and by $\boldsymbol{y}$ a vector containing geometrical and acoustical parameters of interest (or quantities derived from them). Presenting in details these quantities and the models connecting them is the focus of Chap. 6.

5.2.1 Physics-Driven Approaches

The starting point of physics-driven approaches is an explicit *forward acoustic model* $\mathcal{A}$ such that $\boldsymbol{x} \approx \mathcal{A}(\boldsymbol{y})$. In some rare instances, the forward model $\mathcal{A}$ is simple enough that it can be inverted in closed form to directly recover $\boldsymbol{y}$ from $\boldsymbol{x}$. One may cite, for example, the formula linking the *time difference of arrival* τ (feature) of a plane wave impinging at a microphone pair, and the incidence angle θ (parameter) of this wave with respect to the pair: $\tau/(cd) = \cos(\theta)$ where d denotes the microphone distance. Another example closer to our topic is Sabine's empirical formula connecting the reverberation time of a room, denoted RT_{60} (feature), to the area-weighted *mean absorption coefficient* of its walls $\bar{\alpha}$ (parameter): $\mathrm{RT}_{60} = 0.16V/(S\bar{\alpha})$ where V denotes the room's volume and S the total surface area of its boundary. While inverting these formulas is trivial, the simplicity of their underlying forward model makes their range of applicability and their accuracy very limited.

For most tasks of interest, including all the ones addressed in this thesis, $\mathcal{A}$ will be non-trivial to invert, requiring the formulation of an *inverse problem*:

$$\underset{\boldsymbol{y}}{\operatorname{argmin}} \|\boldsymbol{x} - \mathcal{A}(\boldsymbol{y})\| + g(\boldsymbol{y}) \tag{5.1}$$

where $\|\cdot\|$ is an appropriate (pseudo-)norm and g is an optional regularizer on $\boldsymbol{y}$. In some cases, the problem can be formulated in such a way that $\mathcal{A}$ is *linear* but *under-determined*, requiring the use of an appropriate regularizer g, which can make (5.1) amenable to convex optimization. This is notably the approach taken in [17–20] for wall-impedance estimation and in [21] for room geometry estimation, using a sparsity-enforcing regularizer on $\boldsymbol{y}$. However, all of these approaches work by *discretizing space*, making the dimensions of $\mathcal{A}$ grow very quickly with the discrete spatial step size (cubically in 3D), critically limiting the frequency and/or spatial range over which these techniques can be applied in practice.

In most cases, however, $\mathcal{A}$ is non-linear and problem (5.1) is non-convex, requiring the development of ad-hoc algorithms specifically tailored to the problems at hand. As mentioned earlier in

this chapter, there exists almost as many instantiations of *"hearing the walls"* as there are research articles on the topic, making a unified treatment of existing methodologies challenging. We defer the reader to our survey in [S36] for a (partial) attempt at doing so. From a bird's-eye view, let us mention here that the key physical phenomenon making these problems approachable at all is that of *echoes*, *i.e.*, the materialization of early acoustic reflections of the sound wave emitted by the source into the microphones' signals. The *time of arrival* of an echo at a microphone is proportional to the length of the corresponding reflected propagation path, the *time difference of arrival* of an echo between two microphones is linked to its *direction of arrival*, and the *strength and shape* of an echo relative to the direct-path signal is linked to the acoustic attenuation caused by reflectors along its path. Based on this, the core idea of nearly all existing physics-driven methods in the field, as well as the ones presented in this thesis, is to estimate echo-related parameters from measured signals, to associate retrieved echoes to their corresponding reflection paths, and to solve for the parameters of interest based on the recovered information. Consequently, we will mostly make use of *early-time* reverberation models[3].

One distinguishing feature of the physics-driven inverse methods presented in this thesis (Chap. 7 and Sec. 9.2.2, 9.2.3) with respect to most of the prior art, is the estimation of echo-related parameters in *continuous time or space*, alleviating a number of limitations of discretization-based methods. Previous or contemporary examples of continuous approaches include [23–26].

5.2.2 Data-Driven Approaches

An attractive alternative route to explicitly inverting a forward physical model $\mathcal{A}$ connecting $\boldsymbol{y}$ to $\boldsymbol{x}$ is to directly search for an inverse model $f \approx \mathcal{A}^{-1}$ by *fitting* it to a large *training* dataset of (input, target) pairs $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_{\text{train}}}$. The resulting model is then primarily *data-driven*, and the corresponding methodological framework is called *supervised learning*, or more specifically *regression* if the involved variables are continuous. In a nutshell, regression consists in searching for a function $f_{\boldsymbol{\theta}^*}$ inside a family of parameterized functions $\mathcal{F} = \{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$ such that $f_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i) \approx \boldsymbol{y}_i$ for all $i \in [\![1, N_{\text{train}}]\!]$, using a suitable definition of "$\approx$". Ideally, the chosen family $\mathcal{F}$ and the associated search method should be such that one also has $f_{\boldsymbol{\theta}^*}(\tilde{\boldsymbol{x}}) \approx \tilde{\boldsymbol{y}}$ for a "test" pair $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ that lies outside, but reasonably close to, the training set distribution. Over the past decade, *deep artificial neural networks* optimized by variants of *stochastic gradient descent* have proven to form an endlessly versatile source of such parameterized families, with unreasonable effectiveness and generalization capabilities over a very broad range of tasks [27]. While this framework has now deeply permeated most of the field of audio signal processing, its application to acoustics, and in particular to "hearing the walls of a room", is relatively scarcer. Still, the idea has steadily gained interest over recent years, including works estimating acoustical [28, 29] or geometrical [30, 31] parameters from speech recordings or from room impulse response measurements [32].

The crux of the problem in our context is not in designing the model itself, but in gathering a sufficiently large, diverse and representative annotated training dataset $\mathcal{T}$ so that it generalizes to the use case of interest. Real-world audio datasets that include acoustical and geometrical annotations are scarce due to their prohibitive acquisition time and cost. Existing approaches hence leverage training sets that are partly [30, 28, 33, 29] or entirely [32, 31] generated by acoustic simulators, or

[3] We note however that spatio-temporal correlations in the late reverberation field may also contain geometrical and acoustical information of interest, potentially exploitable by physics-driven approaches, as recently hinted by the ongoing development of *statistical wave field theory* [22].

by statistical data augmentation schemes [34]. Our contributions to this recent endeavor are presented in Chap. 8, where the primary focus is on elucidating trade-offs between representativity, diversity, realism, size, and computational costs when designing simulated training sets for acoustic parameter estimation, and the impact of those trade-offs on the generalizability of learned models to real-world data.

5.3 Thesis Outline

Having set the stage, we are now ready to present the outline for the remainder of Part II.

Chap. 6 formalizes the original question by presenting the acoustic (Sec. 6.1, 6.2) and measurement (Sec. 6.3) models that will be used throughout the thesis, highlighting their assumptions and limitations, and detailing the geometrical and acoustical parameters of interest (Sec. 6.4).

Chap. 7 presents several *physics-driven* methods that aim to explicitly invert these models in order to retrieve those parameters, given *room impulse response* measurements. Specifically, Sec. 7.1 deals with *geometry-informed* wall-absorption estimation under forward models of increasing complexity and in the presence of geometrical errors, while Sec. 7.2 deals with the full recovery of both geometrical and acoustical parameters under a highly-simplified room model and a compact microphone array, using a *continuous-space* formulation.

Chap. 8 focuses on *data-driven* methods, and specifically on how to efficiently simulate sufficiently large, diverse and representative training datasets to supervise models that generalize well to real data. Sec. 8.1 examines the main trade-offs involved and present two of our contributions to this effort. Sec. 8.2 revisits the task of estimating the absorption of walls, this time from a single room impulse response, using two different data-driven approaches. Finally, Sec. 8.3 tackles the problem of estimating global geometrical and acoustical parameters of a room *blindly*, directly from noisy speech signals.

Chap. 9 puts the spotlight on the fundamental intermediate quantity of *acoustic echoes*, which can be viewed as a signal-space representation of the acoustic signatures of walls. As an effort to foster *echo-aware audio signal processing*, Sec. 9.1 presents a unique dataset of real acoustic measurements made in a variable-acoustics room, fully annotated by geometrical parameters and echo timings in a mutually-consistent way. Sec. 9.2 presents three methods to estimate echo parameters blindy from two-channel audio signals, operating in the *continuous-time* domain. Finally, Sec. 9.3 slightly opens up our scope by asking whether estimating echo parameters could find applications *beyond room acoustics*, into the more familiar audio signal processing tasks of sound source localization, separation and beamforming. We conclude Part II with Chap. 10, offering an outlook and a number of directions for future research.

6 Acoustic and Measurement Models

When sound propagates from a source to a receiver in a room, it interacts with boundaries and objects within the room by being absorbed, transmitted, or reflected. Hence, the receiver will not only receive sound through the *direct path*, *i.e.*, the straight line traveled from the source to the receiver at the speed of sound c, but also record delayed, attenuated and filtered copies of the emitted signal that will be referred to as *echoes* throughout this thesis. The collection of these echoes is independent of the specific signal emitted by the source (*e.g.*, a hand clap, speech or music). Instead, it contains information on the source and receiver placements, their directive responses, and the acoustic and geometric properties of the room itself. This chapter presents the models that will be used to formalize this *forward physical process*, highlighting the underlying assumptions and limitations. Developing algorithms to *invert these forward models* will be the focus of the next chapters.

6.1 The Wave and Helmholtz Equations

Suppose an omnidirectional point source located at $\boldsymbol{r}_0^{\text{src}} \in \mathbb{R}^3$ emits a signal $s(t)$ inside a room $\Omega \subseteq \mathbb{R}^3$ with boundary $\partial\Omega$. This generates a time-varying *pressure field* $p(\boldsymbol{r}, t)$ inside the domain obeying the following inhomogenous wave equation with boundary conditions:

$$\begin{cases} \dfrac{1}{c^2}\dfrac{\partial^2}{\partial t^2}p(\boldsymbol{r},t) - \Delta p(\boldsymbol{r},t) = s(t)\delta(\boldsymbol{r}-\boldsymbol{r}_0^{\text{src}}),\ \boldsymbol{r}\in\Omega & (6.1) \\ \boldsymbol{n}(\boldsymbol{r})\cdot\nabla p(\boldsymbol{r},t) + \rho\dfrac{\partial}{\partial t}\big(\beta(\boldsymbol{r},\cdot) * p(\boldsymbol{r},\cdot)\big)(t) = 0,\ \boldsymbol{r}\in\partial\Omega & (6.2) \end{cases}$$

where $\boldsymbol{n}(\boldsymbol{r})$ denotes the outward unit normal vector to the surface at $\boldsymbol{r}$, ρ denotes the air density, and $\beta(\boldsymbol{r}, t)$ denotes the time-domain *specific acoustic admittance* of the surface at $\boldsymbol{r}$. The right hand-side of (6.1), called the *source term*, can be derived by interpreting $s(t)$ as the second-order derivative of a point-supported *mass* at $\boldsymbol{r}_0^{\text{src}}$, or equivalently as the (properly normalized) acceleration of the radius of an infinitesimally-small rigid sphere centered at $\boldsymbol{r}_0^{\text{src}}$. Eq. 6.2 is called an *impedance boundary condition* and can be derived by assuming that the room boundaries are *locally reacting*[1], *i.e.*, that the following linear time-invariant relationship exists between the surface-normal component $v_{\boldsymbol{r}}(t) = \boldsymbol{n}(\boldsymbol{r})\cdot\boldsymbol{v}(\boldsymbol{r},t)$ of the *particle velocity* field $\boldsymbol{v}(\boldsymbol{r},t)$ and the pressure at any point $\boldsymbol{r}$ on the boundary:

$$v_{\boldsymbol{r}}(t) = \big(\beta(\boldsymbol{r},\cdot) * p(\boldsymbol{r},\cdot)\big)(t),\ \boldsymbol{r}\in\partial\Omega. \tag{6.3}$$

Then, (6.2) is simply the time-derivative of (6.3) combined with the following relationship between pressure and velocity, derived from the *conservation of momentum*:

$$\nabla p(\boldsymbol{r},t) = -\rho\frac{\partial}{\partial t}\boldsymbol{v}(\boldsymbol{r},t). \tag{6.4}$$

[1] These boundary conditions can be generalized to non-locally reacting surfaces by making the admittance a convolutional operator in both time and 2D space along the surface (see, *e.g.*, [2]), but this will not be explored in this thesis.

For Eq. 6.3 to be *physically realizable*, $\beta(\boldsymbol{r}, \cdot)$ should be *causal* (zero at negative times) and *stable* (absolutely integrable). Moreover, assuming the boundary does not contain any source of energy, the power dissipated by the system should be positive at all frequencies. This can be shown to imply[2] that the real part of the Fourier transform of the admittance is nonnegative, *i.e.*, $\mathrm{Re}[\hat{\beta}(\boldsymbol{r}, \omega)] \geq 0$. A system combining these three conditions is called *positive real*. Positive-real systems are *minimum phase*, which implies that they have a causal and stable inverse (in fact, their inverse is also positive real). The inverse of $\beta(\boldsymbol{r}, \cdot)$ is called the *specific acoustic impedance* at the surface and is denoted $Z(\boldsymbol{r}, \cdot)$. We have:

$$p(\boldsymbol{r}, t) = \big(Z(\boldsymbol{r}, \cdot) * v_{\boldsymbol{r}}\big)(t), \ \boldsymbol{r} \in \partial\Omega. \tag{6.5}$$

Turning now our attention to (6.1), it also implies a linear time-invariant relationship, this time between s and $p(\boldsymbol{r}, \cdot)$ for all $\boldsymbol{r}$, *i.e.*,

$$p(\boldsymbol{r}, t) = \big(h(\boldsymbol{r}, \cdot) * s\big)(t). \tag{6.6}$$

$h(\boldsymbol{r}, \cdot)$ is called the *room impulse response* (RIR) at $\boldsymbol{r}$ for a source at $\boldsymbol{r}_0^{\text{src}}$ in room Ω. Eq. 6.6 is instrumental to the topic of this thesis, because it neatly disentangles the pressure field p into s, which only depends on the source signal, and h, which only depends on the room properties and on the source and receiver positions. It is also of high practical value, since it allows one to simulate any *reverberated signal* given a *dry* source signal and a RIR. The RIR h is solution to a similar system of partial differential equations as p, with the source signal replaced by a time-domain Dirac $\delta(t)$:

$$\begin{cases} \dfrac{1}{c^2}\dfrac{\partial^2}{\partial t^2} h(\boldsymbol{r}, t) - \Delta h(\boldsymbol{r}, t) = \delta(t)\delta(\boldsymbol{r} - \boldsymbol{r}_0^{\text{src}}), \ \boldsymbol{r} \in \Omega & (6.7) \\ \boldsymbol{n}(\boldsymbol{r}) \cdot \nabla h(\boldsymbol{r}, t) + \rho \dfrac{\partial}{\partial t}\big(\beta(\boldsymbol{r}, \cdot) * h(\boldsymbol{r}, \cdot)\big)(t) = 0, \ \boldsymbol{r} \in \partial\Omega. & (6.8) \end{cases}$$

The solution $G(\boldsymbol{r}, \boldsymbol{r}_0^{\text{src}}, t)$ to (6.7)-(6.8) for all $\boldsymbol{r}_0^{\text{src}}$ is called the *Green's function* of the wave equation for these boundary conditions. Note that to guarantee the unicity of the solution, the additional boundary condition that h is causal *i.e.*, $h(\boldsymbol{r}, t \leq 0) = 0$, needs to be added. Taking the Fourier transform of (6.7)-(6.8) yields the following inhomogenous *Helmholtz* equation with so-called *Robin* boundary conditions:

$$\begin{cases} \kappa^2 \hat{h}(\boldsymbol{r}, \omega) + \Delta \hat{h}(\boldsymbol{r}, \omega) = -\delta(\boldsymbol{r} - \boldsymbol{r}_0^{\text{src}}), \ \boldsymbol{r} \in \Omega & (6.9) \\ \boldsymbol{n}(\boldsymbol{r}) \cdot \nabla \hat{h}(\boldsymbol{r}, \omega) + j\omega\rho\hat{\beta}(\boldsymbol{r}, \omega)\hat{h}(\boldsymbol{r}, \omega) = 0, \ \boldsymbol{r} \in \partial\Omega, & (6.10) \end{cases}$$

where $\hat{h}(\boldsymbol{r}, \cdot)$ is called a *room transfer function* and $\kappa = \omega/c$ is called the *wave number*. This particularizes to *Neumann*, *i.e.*, "rigid-wall" boundary conditions when $\hat{\beta} \to 0$ and to *Dirichlet*, *i.e.*, "soft-wall" boundary conditions when $\hat{Z} = 1/\hat{\beta} \to 0$. For a given boundary condition, the general solution to this problem can be expressed as an infinite discrete sum over the *eigenvalues* and *eigenfunctions* of the corresponding Laplacian operator. However, no analytical expression exists for these quantities given general $\hat{\beta}$ and Ω. In fact, studying even their basic properties constitutes an active research topic in pure mathematics, *e.g.*, [36]. For this reason, practical room acoustic simulators only compute approximate solutions, and can be broadly divided into two categories. On the one hand, *wave-based* simulators solve (6.1)-(6.2) or (6.7)-(6.8) by leveraging numerical schemes that

[2] A derivation of this fact can be found, *e.g.*, in the online version of [35] at https://sepwww.stanford.edu/sep/prof/fgdp/c2/paper_html/node5.html.

discretize time and/or space, within the domain or at the boundary. In 3D, these approaches are best suited to simulate low frequencies (say, < 1 kHz) since their computational demand scales cubically or quadratically with the discrete spatial step, which must be set proportionally to the wavelength. On the other hand, *geometrical acoustic* simulators consider the limiting case of sound propagation when the wavelength is small compared to the size of boundary features, and are hence valid at high frequencies (say, > 1 kHz). Analogously to optics, the sound waves emitted by sources are replaced by sound rays that are tracked over time, obeying certain laws of reflection and diffraction when interacting with boundaries. The *image source method* belongs to this category, and is presented in the next section.

6.2 The Image Source Method

6.2.1 The Rigid Shoebox Case

In the particular case when the room Ω is a *cuboid* ("shoebox") of size $L_\mathrm{x} \times L_\mathrm{y} \times L_\mathrm{z}$ and the walls are rigid, *i.e.*, Neumann boundary conditions ($\beta = 0$), it was shown[3] by Allen and Berkley in 1979 [1] that the (causal) solution in Ω to the inhomogeneous wave-equation (6.7)-(6.8) is the same as the (causal) solution to the following *free-field* inhomogeneous wave equation:

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}h(\boldsymbol{r},t) - \Delta h(\boldsymbol{r},t) = \sum_{k=0}^{\infty}\delta(t)\delta(\boldsymbol{r}-\boldsymbol{r}_k^{\mathrm{src}}),\ \boldsymbol{r}\in\mathbb{R}^3, \tag{6.11}$$

where the set of positions $\mathcal{R} = \{\boldsymbol{r}_k^{\mathrm{src}}\}_{k=1}^{\infty}$ in a reference frame defined by a corner of the room consists of the following union of 2^3 orthogonal lattices:

$$\mathcal{R} = \{\boldsymbol{\epsilon}\odot\boldsymbol{r}_0^{\mathrm{src}} + 2\boldsymbol{q}\odot[L_\mathrm{x},L_\mathrm{y},L_\mathrm{z}]^\top \mid \boldsymbol{\epsilon}\in\{0,1\}^3, \boldsymbol{q}\in\mathbb{Z}^3\}. \tag{6.12}$$

Intuitively, the walls of the room have been replaced by an infinite constellation of point *image sources* synchronously emitting an impulse at $t = 0$, as illustrated in Fig. 6.1. By linearity, the solution to (6.11) is the sum of translated copies of the so-called *fundamental solution* h_0 to the equation

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}h_0(\boldsymbol{r},t) - \Delta h_0(\boldsymbol{r},t) = \delta(t)\delta(\boldsymbol{r}),\ \boldsymbol{r}\in\mathbb{R}^3. \tag{6.13}$$

Assuming causality, taking the Fourier transform of (6.13), converting to spherical coordinates, solving by separation of spatial variables and going back to the time domain yields the fundamental solution, also called *Green's function*:

$$h_0(\boldsymbol{r},t) = \frac{\delta(t-\|\boldsymbol{r}\|/c)}{4\pi\|\boldsymbol{r}\|}, \tag{6.14}$$

[3]The Fourier-domain proof given in the appendix of [1] yields a solution that is, technically, non-causal. An alternative proof in the time domain that respects causality was recently presented in the PhD thesis of Tom Sprunck [37, Chap. 4].

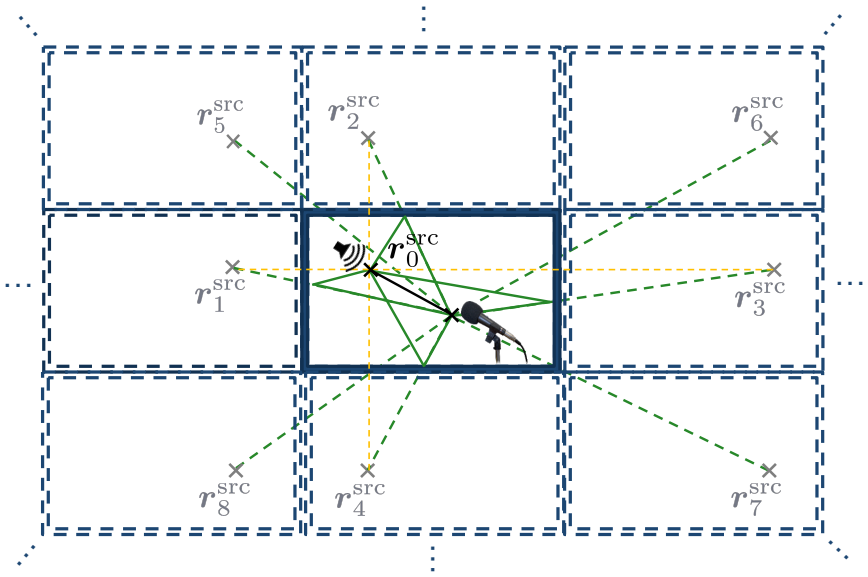


FIGURE 6.1: Illustration of the image source method in a 2D rectangular room. The direct path is depicted as a solid black line, the first-order reflection paths as solid green lines, the lines of sight to first- and second-order image sources as dashed green lines. The walls are bi-sectors of the segments from the true source to its first-order images (dotted orange lines). The method can be intepreted as replacing the room boundaries by iteratively reflected copies of the original room, eventually tiling the space.

namely, a spherical impulse propagating outward from the origin at the speed of sound. The RIR solution to (6.11) is therefore a sum of synchronous, translated copies of this spherical impulse, *i.e.*,

$$h_{\text{rigid}}(\boldsymbol{r}, t) = \sum_{k=0}^{\infty} \frac{\delta(t - \|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|/c)}{4\pi\|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|}. \tag{6.15}$$

In other words, in continuous time and at a fixed receiver location, a rigid shoebox room impulse response consists of an infinite stream of delayed Dirac impulses, with coefficients decaying as the inverse of their times of arrival. In this equation (and in its subsequent extensions hereafter), the individual terms in the sum are what we refer to as *echoes*.

Eq. 6.15 is the propagation model underlying the original *image source method* (ISM), as introduced by Allen and Berkley in 1979 [1]. While it rigorously solves (6.7)-(6.8) (in the weak, distributional sense), it is interesting to note that it does not necessary yield stable solutions to (6.1)-(6.2) for arbitrary source signals. Indeed, following (6.12), the density of Dirac impulses in a fixed-length time interval grows quadratically in t, while their coefficients only decay in $1/t$. So for example, convolving such a RIR with a rectangular source signal in (6.6) will lead to a divergent pressure field.

6.2.2 Extensions

6.2.2.1 Wall Absorption and Atmospheric Attenuation

To avoid the divergence issue mentioned in the previous section, practical implementations of the ISM only sum over a finite number K of image sources. Moreover, each Dirac impulse is multiplied by an attenuation $a_k \in [0, 1)$, yielding what we will call the *vanilla* ISM model:

$$h_{\text{vanilla}}(\boldsymbol{r}, t) = \sum_{k=0}^{K} a_k \frac{\delta(t - \|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|/c)}{4\pi\|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|}. \tag{6.16}$$

The most common way to compute the attenuations is to assign an *absorption coefficient* $\alpha_i \in (0, 1]$ to each wall i, and a corresponding reflection coefficient $R_i = \sqrt{1 - \alpha_i} \in [0, 1)$. Each a_k is then set to the following product:

$$a_k = \prod_{i \in \mathcal{W}_k} R_i = \prod_{i \in \mathcal{W}_k} \sqrt{1 - \alpha_i} \tag{6.17}$$

where $\mathcal{W}_k$ denotes the multiset of wall indices associated to image source k. This approach yields an exponential decay of attenuations in the image source order, further justifying the truncation of the sum to K and preventing any explosion over time when convolving (6.16) with a source signal.

This vanilla model can be further generalized by replacing reflection coefficients by *wall impulse response*[4] (WIR) filters $R_i(t)$ that are iteratively convolved in time to obtain *attenuation filters*:

$$a_k(t) = \left(\circledast_{i \in \mathcal{W}_k} R_i \right)(t), \tag{6.18}$$

yielding:

$$h_{\text{WIR}}(\boldsymbol{r}, t) = \sum_{k=0}^{K} \frac{1}{4\pi\|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|} \big(a_k * \delta(\cdot - \|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|/c)\big)(t). \tag{6.19}$$

The frequency-dependent absorption coefficients associated to the walls are then given by

$$\alpha_i(\omega) = 1 - |\hat{R}_i(\omega)|^2. \tag{6.20}$$

Note that while $\hat{R}_i$ captures the phase-response of walls in the frequency domain, the phase information is lost in α_i. Since the acoustic characteristics of construction materials are typically available in the form of absorption coefficients averaged over octave bands, there is a missing phase information that one needs to recreate when simulating such materials using impedance-based models. This constitutes an open research topic, see *e.g.*, [38] for some recent developments.

While both h_{vanilla} and h_{WIR} feature tractable expressions that will be amenable to inverse methods (see Chap. 7), it is important to note that they are no longer solutions to the original wave equation (6.7)-(6.8). In fact, to our knowledge, there does not exist any admittance function $\beta = Z^{-1}$ at the boundary such that they represent such solutions, even in the limiting case $K \to \infty$. One way to improve this is to approximate the spherical waves emitted by first-order image sources by plane waves impinging on each wall i with an incidence angle $\theta_i \in [0, \pi/2]$, with 0 corresponding to the normal incidence. This approximation is valid when the observation point $\boldsymbol{r}$ is sufficiently far

[4] To our knowledge, the term *wall impulse response* and a similar RIR model as (6.19) were first introduced in [21].

from the corresponding image source position[5]. There is then a well-known relationship between the incidence-dependent reflection coefficient and the impedance of each wall in the Fourier domain:

$$\hat{R}_i(\theta_i, \omega) = \frac{\left(\hat{Z}_i(\omega)/\rho c\right)\cos(\theta_i) - 1}{\left(\hat{Z}_i(\omega)/\rho c\right)\cos(\theta_i) + 1} \tag{6.21}$$

where $Z_i/\rho c$ is the *normalized impedance* at wall i. An implementation of this idea, iterated to higher image-source orders via (6.18), was recently presented in [40], and its accuracy was validated against wave-based simulators. Interestingly, it can be shown that under model (6.21), the positive-real condition on Z mentioned in Sec. 6.2.1 is equivalent to the intuitive condition that R is causal, stable, and verifies $\hat{R}_i(\theta, \omega) \leq 1$ for all θ and ω, *i.e.*, that walls *absorb* energy of plane waves from all angles at all frequencies. While $R_i(\theta, t)$ is not necessarily minimum phase for all angles, Eq. 6.21 suggests that it is for angles close to 0 and for large enough impedances (near-rigid surfaces). This is the assumption used by the two ISM-based simulators `Roomsim` [41] and `pyroomacoustics` [42] used in this thesis to handle the missing-phase problem mentioned in the last paragraph.

A last way to improve the sound propagation model is to consider atmospheric attenuation. Applied to (6.19), this can be expressed in the Fourier domain as:

$$\hat{h}_{\text{WIR-air}}(\boldsymbol{r}, \omega) = \sum_{k=0}^{K} e^{-\alpha_{\text{air}}(\omega)\|\boldsymbol{r}-\boldsymbol{r}_k^{\text{src}}\|} \hat{a}_k(\omega) \frac{e^{-j\omega\|\boldsymbol{r}-\boldsymbol{r}_k^{\text{src}}\|/c}}{4\pi\|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|} \tag{6.22}$$

where different models can be used for the air absorption coefficient $\alpha_{\text{air}}(\omega)$, depending on the air temperature, humidity and pressure of reference [43].

The inverse methods presented in Chap. 7, 8, 9 will mostly use the continuous-time propagation models h_{vanilla} or h_{WIR} as a backbone. This is justified by the fact that our interest mostly lies in early reflections, for which incidence angles are typically close to zero (making reflection coefficients roughly independent of θ in (6.21)) and travel distances are short (making atmospheric attenuation negligible). Arguably, the frequency-independence of walls assumed in h_{vanilla}, though common, is quite crude, and is only expected to hold for near-rigid surfaces, and/or over limited frequency bands.

6.2.2.2 Non-Shoebox Rooms

Another way to extend the ISM is to consider non-cuboid rooms. However, even in the rigid-wall case, the equivalence of the approach to the wave equation then disappears, except for a restricted set of room shapes. These are the right prisms whose basis is either a rectangle (shoebox case), an equilateral triangle, a right isosceles triangle, or half an equilateral triangle (see [44] for a detailed discussion on this). Beyond these shapes, at least two phenomena that are not accounted for by the classical ISM emerge. First, curvatures at the boundary, and in particular sharp edges and corners, give rise to diffraction effects. Second, even fully neglecting curved parts of the boundary and following the iterative reflection scheme of the ISM across planar parts only, *occlusions* will inevitably arise, namely, some image sources will not be visible (or rather *audible*!) in some subsets of the room domain. Such occlusions can be computed efficiently using geometrical tests for each image

[5]To go further, exact expressions of error terms for spherical waves impinging on locally reacting impedance planes can be found in [39, Sec. D.17].

source and receiver position at each reflection order, but they significantly complicate the forward propagation model, making its inversion all the more challenging.

For these reasons, all the works presented in this thesis make use of a shoebox room model. While this is a strong assumption, it remains a reasonable approximation for many typical rooms in the real world, assuming they are emptied of furniture[6].

Even restricting to shoebox rooms, an extension of high practical relevance would be to consider non-constant impedance at individual walls, accounting for the presence of, *e.g.*, windows, curtains or doors. At least for piecewise-constant variations, and neglecting possible diffraction effects at interfaces, this setting is presumably easier to handle, as one only needs to assign an appropriate attenuation filter to each image-source and receiver. Although this is not explicitly tackled in this thesis, it is believed that most of the methods presented in Chap. 7, 8 and 9 could be extended to account for this with little to no adjustment.

6.2.2.3 Directive Sources

So far we have only considered *omnidirectional point sources* of sound, also called *monopoles*, as modeled by the spatial Dirac $\delta(\boldsymbol{r} - \boldsymbol{r}_0^{\text{src}})$ on the right hand side of (6.1). This is an idealized model that does not exist in the real world[7]. Most commonly encountered sound sources such as speaking humans or loudspeakers are highly directive. Even so-called "omnidirectional loudspeakers", typically consisting of an array of loudspeakers facing outward of a sphere or placed on a carefully designed conical mount, fail to be omnidirectional for frequencies above $\sim$1 kHz. Of course, perfect point sources do not exist either, although for compact enough sources, a directive point-source model is reasonable at distances significantly larger than the device's size.

Despite the obvious invalidity of the omnidirectional assumption and the well-known significant impact of source directivity on sound fields, especially at early times, the assumption remains widely used in the acoustic signal processing literature today due to its simplicity. An important part of this thesis is dedicated to study the limitations of this model in the context of acoustic parameter estimation, and to go beyond it. Omnidirectional RIR models based on the ISM can be extended to account for source directivity relatively straightforwardly, as presented in, *e.g.*, the thesis of Dirk Schroeder [46]. Applied to our previous h_{WIR} model, this gives:

$$h_{\text{WIR-dir}}(\boldsymbol{r}, t) = \sum_{k=0}^{K} \frac{1}{4\pi \|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|} \left(g^{\text{src}}(\theta_{k,\boldsymbol{r}}^{\text{out}}, \phi_{k,\boldsymbol{r}}^{\text{out}}, \cdot) * a_k * \delta(\cdot - \|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|/c) \right)(t) \tag{6.23}$$

where $(\theta_{k,\boldsymbol{r}}^{\text{out}}, \phi_{k,\boldsymbol{r}}^{\text{out}})$ denotes the azimuth and elevation *angles of departure*[8] from image source k to the receiving location $\boldsymbol{r}$ and $g^{\text{src}}(\theta, \phi, t)$ denotes the source's *directivity pattern*, namely, an impulse response for each direction on the sphere.

Connecting this model to the original wave equation (6.1) is not immediately obvious, but can be done rigorously (at least in the free-field and rigid-shoebox cases) by replacing the monopole source

[6] This thesis, for instance, has been entirely written within (rough) shoebox rooms of different dimensions!

[7] A noteworthy attempt to emulate a point omnidirectional sound source in the context of acoustic impulse measurements was made by Javier Gómez Bolaños from Aalto University using laser-induced sparks [45]. The setup is however costly, cumbersome and hard to reproduce.

[8] Note that under this definition, these angles need to take into account geometrical reflections of the original source directivity pattern across reflective surfaces along the ISM iterations.

term $\delta(\boldsymbol{r} - \boldsymbol{r}_0^{\text{src}})$ by a *multipole* source term. A multipole source term can be formed by a linear combination of the spatial Dirac measure at $\boldsymbol{r}_0^{\text{src}}$ with its higher-order spatial derivatives projected along spatial axes. It is in fact a theorem that all point-supported measures can be expressed in this way, *e.g.*, [47, Th. 2.3.4)] . How to recover a (Fourier-domain) directivity pattern $\hat{g}^{\text{src}}(\theta, \phi, \omega)$ from such a multipole source term is discussed in [48, Sec. 6.5], using the spherical harmonic decomposition of $\hat{g}^{\text{src}}$. As explained in the book, this is not trivial for arbitrary n-poles since there is no one-to-one correspondence between Dirac derivatives and spherical harmonic basis functions.

6.3 Measurement Model

The previous sections described models of the pressure field in continuous time and space. Crucially, any acoustic measurement apparatus will only access to discrete samples of this field, through the use of one or several microphones. Modeling spatial sampling is straightforward, namely, the sound field will only be measured at a finite set of microphone positions $\boldsymbol{r} \in \{\boldsymbol{r}_m^{\text{mic}}\}_{m=1}^M$. Time sampling is then modeled using a filter applied to a continuous-time RIR h as follows:

$$x_m[n] = \big(g_m^{\text{mic}} * h(\boldsymbol{r}_m^{\text{mic}}, \cdot)\big)(n/f_\text{s}),\ n \in [\![0, N-1]\!] \tag{6.24}$$

where $g_m^{\text{mic}}(t)$ is a low-pass filter modeling the response of microphone m, f_s denotes the sampling frequency in Hz and N is the number of acquired pressure samples per microphone in time. A typical simple choice for $g_m^{\text{mic}}(t)$ is the *ideal low-pass filter* with cut-off frequency $f_\text{s}/2$ (see Eq.1.5):

$$g_m^{\text{mic}}(t) = \text{sinc}(f_\text{s} t). \tag{6.25}$$

Note that this choice should not be interpreted as modeling the actual physical response of the microphone, since this filter is not causal. A better interpretation of this model is that it approximates the signal one would obtain after perfectly *compensating* the device response via deconvolution and post-processing, assuming a cut-off frequency of $f_\text{s}/2$. As an example, applying (6.24) and (6.25) to the vanilla-ISM continuous-time RIR model h_{vanilla} in (6.16) yields:

$$x_{\text{vanilla},m}[n] = \sum_{k=0}^{K} a_k \frac{\text{sinc}(n - \|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\| f_\text{s}/c)}{4\pi \|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\|}, \tag{6.26}$$

which corresponds the default implementation of many commonly used ISM-based room acoustic simulators, such as, *e.g.*, `pyroomacoustics` [42].

Analogously to Sec. 6.2.2.3, one may further improve this model by employing a directive microphone model. Applied to the directive-source WIR ISM model $h_{\text{WIR-dir}}$ in (6.23) this gives:

$$x_{\text{WIR-dir},m}[n] = \sum_{k=0}^{K} \frac{1}{4\pi \|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\|} \Big(g^{\text{src}}(\theta_{k,m}^{\text{out}}, \phi_{k,m}^{\text{out}}, \cdot) * g_m^{\text{mic}}(\theta_{k,m}^{\text{in}}, \phi_{k,m}^{\text{in}}, \cdot) * a_k * \text{sinc}(\cdot - \|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\|/c) \Big)(n/f_\text{s}) \tag{6.27}$$

where $(\theta_{k,m}^{\text{out}}, \phi_{k,m}^{\text{out}})$ and $(\theta_{k,m}^{\text{in}}, \phi_{k,m}^{\text{in}})$ respectively denote the angles of departure and arrival from image source k to microphone m and $g_m^{\text{mic}}(\theta, \phi, t)$ denotes the directivity pattern of microphone m. From a computational perspective, the continuous-time convolutions in (6.27) can be approximated

by discrete-time convolutions of discrete-finite-time approximations of the involved filters (see Eq. 1.7 and surrounding discussion). They can be implemented by point-wise multiplications in the discrete Fourier domain using the convolution theorem (see Eq. 1.4), or via Toeplitz matrices in the discrete-time domain. Model (6.27) was notably implemented by Prerak Srivastava during his PhD thesis, leveraging measured directivity patterns from the DirPat dataset [49], as part of our work published in [S31] (see Sec. 8.1.3). The code was recently merged into the main branch of `pyroomacoustics` [42].

Finally, the above models can be convolved in time with an arbitrary dry source signal $s(t)$ as in (6.6) to produce a *reverberant signal* model. Reverberant signals can then be additively mixed and further corrupted by additive spatially-white measurement noise or diffuse background noise to emulate rich acoustic scene recordings.

6.4 Parameters of Interest

The main aim of this thesis is to estimate geometrical and acoustical parameters of interest given acoustic measurements of the forms described in Sec. 6.3. For a given reference coordinate frame, we refer to as *geometrical parameters* the room boundary $\partial\Omega \subset \mathbb{R}^3$, the receiver positions $\{\boldsymbol{r}_m^{\text{mic}}\}_{m=1}^M$, the source position $\boldsymbol{r}_0^{\text{src}}$, as well as any quantity derived from those, such as the *volume* V of the room, the *total surface area* S of its boundary, the surface areas S_i of individual walls, their distances to the source and receivers, the orientation and translation of the room boundary in a microphone's coordinate frame, the positions of images sources $\{\boldsymbol{r}_k^{\text{src}}\}_{k=1}^K$, etc.

Complementarily, we refer to as *acoustical parameters* the admittance (or equivalently impedance) along the boundary $\partial\Omega$, the source's directivity pattern g^{src}, the receivers' directivity patterns $\{g_m^{\text{mic}}\}_{m=1}^M$ as well as quantities derived from them and from geometrical parameters. Following building acoustic standards that are themselves perceptually motivated, these derived quantities are typically averaged over *octave bands*, *i.e.*, 6 frequency intervals of the form $[b/\sqrt{2}, b\sqrt{2}]$ where $b \in \{.125, .250, .500, 1, 2, 4\}$ kHz. We will adopt this convention throughout the thesis, unless stated otherwise. Derived acoustic quantities of interest include the absorption coefficients $\alpha_i(b)$ of individual walls, the area-weighted *mean absorption coefficient* $\bar{\alpha}(b) = (\sum_i \alpha_i(b) S_i)/S$, and the reverberation time $\text{RT}_{60}(b)$ calculated from RIRs using Schroeder's integration method [50].

Sabine's law famously relates some of these derived quantities[9], as empirically discovered by Wallace Sabine in the late 1890s:

$$\text{RT}_{60}(b) \approx 0.161 \frac{V}{S\bar{\alpha}(b)}. \tag{6.28}$$

Its close variant *Eyring's law* was later discovered from first principles and is known to be more precise for small rooms with large absorption:

$$\text{RT}_{60}(b) \approx 0.161 \frac{V}{S \ln(1 - \bar{\alpha}(b))}. \tag{6.29}$$

[9]Note that despite a common confusion, the absorption coefficients used in Sabine and Eyring's laws are not the same as the angle-dependent absorption coefficients one can derive from the impedance using (6.21). They are instead averaged versions of those over incident angles, under the assumption of a diffuse isotropic sound field. The correct way to measure them is to place the material of interest in a reverberation chamber.

While these laws can be used to obtain rough approximations of some quantities of interest given the others, they strongly rely on the hypothesis of a *diffuse and isotropic late sound field*. This hypothesis is reasonable in a room whose 3 dimensions are roughly equal, with similar absorption coefficients on each wall, and using omnidirectional devices placed as far as possible from the boundary. Conversely, any departure from such conditions, which occurs in most commonly encountered rooms and practical setups, decreases the validity of this hypothesis, making these laws less accurate.

In the context of this thesis, an estimation method is called *geometry-informed* if *all* of the geometrical parameters are available, possibly with some errors. A method is called *blind* when it estimates acoustical parameters from a reverberant signal without any geometrical knowledge, nor with any knowledge of the emitted source signal $s(t)$. Non-blind methods retrieve parameters from discrete RIR measurements, with or without assuming geometrical knowledge.

7 Physics-Driven Inverse Methods

Associated publications: [S29, S30, S36] and [51, Chap. 6].

This chapter presents several *physics-driven* methods (see Sec. 5.2.2) that aim to explicitly invert the RIR measurement models of Sec. 6.3 in order to retrieve the parameters of interest introduced in Sec. 6.4, given RIR observations. Specifically, Sec. 7.1 deals with geometry-informed absorption coefficient estimation from multiple RIRs under image source models (ISMs) of increasing complexity and in the presence of geometrical errors, while Sec. 7.2 deals with the full recovery of both geometrical and acoustical parameters under the vanilla ISM (6.26) using a multi-stage approach and a compact microphone array.

7.1 Geometry-Informed Absorption Profile Estimation

7.1.1 Spectrogram-Domain Approach with Idealized Devices

Associated publication: [S29]

7.1.1.1 A Robust Echo-Pruning Method

Let us consider a set of discrete RIRs $\{x_{m,j}[n]\}_{m,j=1}^{M,J}$ measured in a room by a set of M microphones and J sources, respectively located at $\{\boldsymbol{r}_m^{\text{mic}}\}_{m=1}^{M}$ and $\{\boldsymbol{r}_{j,0}^{\text{src}}\}_{j=1}^{J}$. We employ here the wall-impulse-response image-source model (6.19) and further assume that the sources and microphones are omnidirectional and calibrated such that they have a single angle-independent time-domain response jointly represented by a discrete filter $g[n]$. Following Chap. 6 and approximating continuous-time convolutions $*$ by (finite) discrete-time convolutions $\circledast$ using (1.7), this model writes:

$$x_{m,j}[n] = \sum_{k=0}^{K} \frac{f_{\text{s}}}{c\tau_{m,j,k}} \big(g \circledast \text{sinc}(\cdot - \tau_{m,j,k}) \circledast a_k\big)[n] + e_{m,j}[n] \tag{7.1}$$

where $\tau_{m,j,k} = f_{\text{s}}\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_{j,k}^{\text{src}}\|/c$ denotes the time of arrival (TOA) of the k-th image of source j to microphone m in samples, $a_k[n]$ denotes the attenuation filter of image source k in discrete time, and $e_{m,j}[n]$ captures measurements and modeling errors. Similarly to (6.18) we assume:

$$a_k[n] = \left(\circledast_{i \in \mathcal{W}_k} R_i \right)[n], \tag{7.2}$$

where $R_i[n]$ is the discrete-time reflection filter of wall i, with $i = 1, \ldots, 6$. Suppose we observe $\{x_{m,j}[n]\}_{m,j=1}^{M,J}$ and have access to the geometrical parameters of the scene in the form of the times of arrival of echoes $\{\tilde{\tau}_{m,j,k}\}_{m,j,k}$, possibly corrupted by errors. This a realistic scenario, since in practice one can never measure the positions of sources, receivers and reflectors in a room with perfect accuracy. Under (7.1), estimating the wall reflection filters $R_{1:6}[n]$ then amounts to a joint separation and multi-linear deconvolution problem under filter mismatch, for which no general inverse method

is known. To tackle it, the key idea in this work is to make the further simplifying assumption that the source-microphone response g and the image source attenuations $a_k[n]$ are *fast-decaying*, such that the *time-domain support* of echoes in RIRs have a short maximum length $L_{\max}$ in samples. For attenuation filters, this assumption is compatible with the minimum-phase model of reflection filters discussed in Sec 6.2.2.1. Indeed, minimum-phase filters are the fastest-decaying ones given prescribed Fourier magnitudes.

Under this assumption, by the convolution theorem (see Eq. 1.4), the discrete convolutions in (7.1) and (7.2) are well approximated by point-wise multiplications in the short-time discrete Fourier domain. Let us denote by $X_{m,j}$ the *squared magnitude spectrogram* of $x_{m,j}$, such that $X_{m,j}[f,n]$ is the squared magnitude of the discrete Fourier transform (DFT, Eq. 1.2) of the windowed RIR $x_{m,j}[n{:}n+2F-1]$ where $2F \gg L_{\max}$. If we assume that a *single acoustic echo*, indexed by k, occurs in this window for this source-microphone pair, then there is a single term left in the sum of (7.1) and we obtain the simple expression:

$$X_{m,j}[f,n] \approx \frac{f_{\mathrm{s}}}{c\tau_{m,j,k}} G[f] A_k[f] \tag{7.3}$$

where G and A_k respectively denote the squared magnitudes of the DFT of g and a_k. Based on (7.2), we have:

$$A_k[f] = \prod_{i \in \mathcal{W}_k} |R_i[f]|^2 = \prod_{i \in \mathcal{W}_k} (1 - \hat{\alpha}_i[f]), \tag{7.4}$$

where $\hat{\alpha}_i[f]$ is the absorption coefficient of wall i in the discrete frequency domain, namely, the quantity we want to retrieve. Let us denote by $\mathcal{J}_k^* \subset \mathbb{N}^3$ the set of microphone-source-window index triples (m,j,n) such that there is *one-and-only-one* echo occurring in $x_{m,j}[n{:}n+2F-1]$, associated to image source k. Note that this set can be computed using the available times of arrival of echoes $\{\tilde{\tau}_{m,j,k}\}_{m,j,k}$. We can now formulate the problem of retrieving the wall absorption coefficients in a given frequency bin f as a non-linear constrained least-square minimization problem:

$$\underset{\substack{\hat{\alpha}_{1:6}[f] \in [0,1] \\ G[f] \geq 0}}{\operatorname{argmin}} \sum_{k=0}^{K} \sum_{(m,j,n) \in \mathcal{J}_k^*} \left\| X_{m,j}[f,n] - \frac{f_{\mathrm{s}}}{c\tau_{m,j,k}} G[f] A_k[f] \right\|_2^2. \tag{7.5}$$

In other words, for each image source k, we attempt to fit model (7.3) to all spectrogram bins that contain an isolated echo from image source k. We found that this problem can be satisfyingly solved using a properly-initialized non-linear solver[1]. There are, however, two major hurdles that prevent this approach to work robustly:

1. The times of arrival of echoes are only available up to some geometrical errors, which means there will be mistakes in the set of isolated-echo spectrogram bins $\mathcal{J}_k^*$.
2. Even assuming perfectly known geometry, the fast-decaying assumption on echoes is limited, causing inevitable residual overlap and interference between echoes in some bins.

To counter this, we devise an algorithm inspired by random sampling consensus (RANSAC, [52]) that iteratively fits model (7.3) on *random subsets* of spectrogram bins, selected based on their probability

[1] In practice, we use the `fmincon` solver of Matlab. G and $\hat{\alpha}$ are respectively initialized by first truncating the outer sum in (7.5) to the 0-th image-source order, then to the first order. Both truncation yield non-negative *linear* least-square problems that can be solved exactly using, *e.g.*, the `nnls` solver of Matlab.

Bin set	σ_{geo}	$Q = 1$		$Q = 2$	
		MAE	CE (%)	MAE	CE (%)
$\mathcal{J}_{\boldsymbol{k}}^{*}$	0 cm	0.095	79.7	0.128	71.1
	2 cm	0.108	74.7	0.132	69.6
$\mathcal{J}_{k}^{\text{ransac}}$	2 cm	0.088	82.9	**0.082**	**84.6**

TABLE 7.1: Mean absolute error (MAE) and percentage of correct estimates (CE) on absorption coefficients obtained by minimizing (7.5) over spectrogram bin sets $\mathcal{J}_{k}^{*}$ or $\mathcal{J}_{k}^{\text{ransac}}$ and over image sources up to order $Q = 1$ or $Q = 2$.

of echo isolation, using a Gaussian error model on times of arrival. The subsets are extended and scored according to a model-fit criterion, eventually leading to a more robust isolated-echo bin set $\mathcal{J}_{k}^{\text{ransac}}$ for each k, which can be used in (7.5) instead of $\mathcal{J}_{k}^{*}$.

7.1.1.2 Results on Spectrogram-Domain Absorption Estimation

The approach is tested on a dataset of synthetic shoebox RIRs simulated with the room acoustics simulator `Roomsim` [41], implementing model (7.1). Simulated RIRs are cropped to 250 ms, sampled at a rate $f_{\text{s}} = 16$ kHz and include specular reflections up to order 20. 500 rooms are simulated with length, width and height sampled uniformly at random in $[3, 10] \times [3, 10] \times [2, 5]$ in meters. Each room contains S sources and M microphones whose positions are sampled uniformly at random in the room under the constraints of non-closeness to walls (1 meter) and non-mutual-closeness (1 meter). For each wall, absorption coefficients in 6 logarithmically-spaced octave bands centered at $.125, .250, \ldots, 4$ kHz are drawn randomly using the *reflectivity-bias* sampling strategy [S26] that will be later described in this thesis (Sec. 8.1.2). This strategy is designed to generate realistic, diverse and representative room acoustic characteristics. The corresponding ground-truth DFT-domain absorption coefficients are obtained with a linear interpolation scheme. A minimum-phase model on these coefficients is used to simulate reflections, as discussed in Sec. 6.2.2.1. A simple Dirac is used for the source-microphone filter $g[n]$. White Gaussian noise is added to the simulated RIRs to achieve a fixed peak signal-to-noise (PSNR).

To model geometrical errors, a deliberate mismatch is introduced between the geometrical parameters used to simulate the RIRs and those used to compute the TOAs $\{\tilde{\tau}_{m,j,k}\}_{m,j,k}$. A Gaussian noise with standard deviation σ_{geo} (in cm) is added to the positions of sources, microphones and walls. The true value of σ_{geo} was used to compute the set $\mathcal{J}_{\boldsymbol{k}}^{\text{ransac}}$, implying that the precision of the geometrical measurement device is known.

Spectrograms are computed using DFT windows of size $2F = 32$ samples (2 ms) and a hop size of 1 sample. This results in estimated absorption coefficients within 16 linearly-spaced frequency bands approximately centered at $f \in \{0, 470, \ldots, 8000\}$ Hz. Echoes are assumed to be of length $L_{\text{max}} = 8$ samples (0.5 ms). Preliminary experiments revealed that the proposed approach was not able to correctly estimate absorption values at the lowest (DC) frequency band $f = 0$, yielding values close to random. Hence, this frequency band is omitted in the following results. This limitation can be explained by the relatively short DFT window size employed here. On the other hand, increasing the window size showed to degrade results by decreasing the number of spectrogram bins containing isolated echoes.

Two metrics are used to evaluate the estimation of absorption coefficients over all frequencies, walls and rooms: the mean absolute error (MAE) and the total percentage of *correctly estimated* (CE)

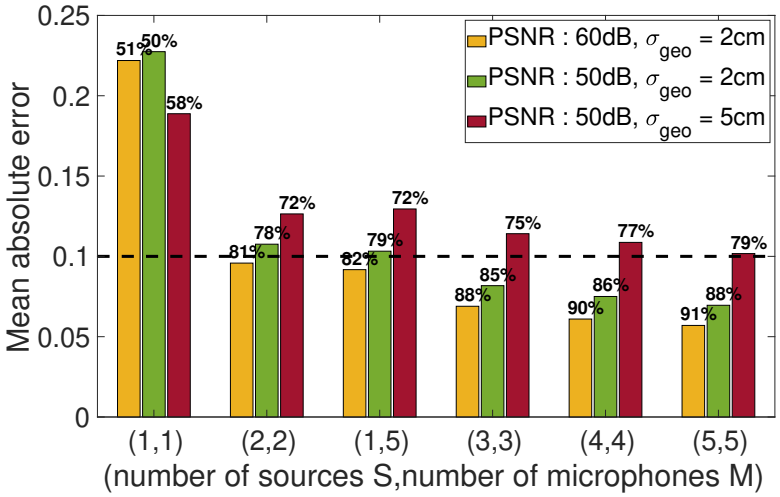


FIGURE 7.1: Mean absolute error and percentage of correct estimates (above each bar) achieved with the proposed RANSAC-inspired procedure and echoes up to order $Q = 2$ for different numbers of sources and microphones and various levels of noise and geometrical error.

coefficients, *i.e.* with an error smaller than 0.1 (recall that absorption coefficients take values between 0 and 1). This threshold is meant to reflect what would be an appropriate tolerance for acoustic diagnosis purpose. Note that in some rare cases ($< 1\%$ for $\mathcal{J}_k^{\text{ransac}}$) a spectrogram bin set computed by our approach can be empty, making the estimation of corresponding absorption coefficients impossible or very poor. In such cases, the estimate was set to the middle value of 0.5.

Table 7.1 compares the results obtained using $\mathcal{J}_k^*$ or $\mathcal{J}_k^{\text{ransac}}$ in the inner sum of problem (7.5), as well as the effect of modeling echoes up to order $Q = 1$ or up to order $Q = 2$ in the outer sum, with a geometrical error $\sigma_{\text{geo}} \in \{0 \text{ cm}, \ 2 \text{ cm}\}$, a PSNR of 50 dB, $M = 3$ microphones and $J = 3$ sources (9 RIRs). Looking at the first row, assuming perfectly known geometry (σ_{geo} =0 cm), the absorption coefficients can already be correctly estimate 80% of the time using the deterministic bin set $\mathcal{J}_k^*$ and echoes up to order 1 only. Interestingly, the results significantly degrade using echoes of order 2 as well, which highlights the issue of overlap and interference mentioned in the last section at this order. With a geometrical error σ_{geo} =2 cm, the best results are obtained using the proposed random sampling consensus approach $\mathcal{J}_k^{\text{ransac}}$ and including echoes up to order 2, with a mean error on absorption coefficients of 0.082 and 84.6% percent of correct estimation.

Fig. 7.1 then focuses on this best-performing setting. It jointly studies the influence of the number of sources S and microphones M and the robustness of the approach to measurement noise and geometrical errors. As expected, it reveals that reducing the PSNR and increasing σ_{geo} systematically degrades performance. On the other hand, increasing the number of available RIRs per room consistently improves both the MAE and CE metrics, eventually compensating the degradation. This suggests that the proposed approach succeeds in selecting the most relevant RIR spectrogram bins despite their increasing number. Note that with PSNR $= 50$ dB and $\sigma_{\text{geo}} = 2$ cm, using only 4 RIRs with $(S, M) = (2, 2)$ suffices to reach a satisfying MAE close to 0.1, with 78% of correct estimates. Interestingly, similar results are obtained for a somewhat more practical setup consisting of $S = 1$ source and $M = 5$ microphones. On the other hand, increasing to 4 sources and 4 microphones (16 RIRs) allows to preserve similar performance when geometrical errors have a standard deviation as high as 5 cm.

7.1.1.3 Conclusion on Spectrogram-Domain Absorption Estimation

This section presented a spectrogram-domain optimization method that jointly estimates the frequency-dependent absorption coefficients of the 6 walls of a shoebox room given a set of RIRs and knowledge of the geometrical parameters of the scene. On simulated RIRs, the approach was shown to yield satisfying results at 400 Hz or above, with most estimation errors below 0.1 and an encouraging robustness to noise and geometrical errors. A strong limitation of this approach is that it fundamentally hinges on the assumption that echoes are fast-decaying, such that most of their energy spans 0.5 ms. Unfortunately, this idealistic assumption does not hold on real-world measured RIR. While it may be reasonable to assume this for the response of typical microphones and of walls, the main problem is the response of sources, which may span up to 10 ms for typical loudspeakers, which are, in addition, far from omnidirectional[2]. Therefore, the most plausible avenue to make this approach applicable to real data is the development of *RIR enhancement* techniques, with a particular attention to *compensating* the (possibly known) directional responses of devices. This constitutes a non-trivial, interesting, and to our knowledge little-studied research direction, which we further discuss in Chap. 10.

7.1.2 Time-Domain Approach with Directive Devices

This section is based on yet-unpublished work from Stéphane Dilungana's PhD thesis [51, Chap. 6].

7.1.2.1 The Exact Case: A Linear Inverse Problem

To address the limitations of the spectrogram-domain approach with idealized devices presented in Sec. 7.1.1, we now turn our attention to the more realistic RIR model (6.27), that takes into account both the microphone and source directive responses. We rewrite it below for the case of M microphones, J sources and using discrete-time filters and convolutions (see Eq. 1.7):

$$x_{m,j}[n] = \sum_{k=0}^{K} \frac{f_\mathrm{s}}{c\tau_{m,j,k}} \left(g_j^{\mathrm{src}}(\theta_{k,m,j}^{\mathrm{out}}, \phi_{k,m,j}^{\mathrm{out}}, \cdot) \circledast g_m^{\mathrm{mic}}(\theta_{k,m,j}^{\mathrm{in}}, \phi_{k,m,j}^{\mathrm{in}}, \cdot) \circledast \mathrm{sinc}(\cdot - \tau_{m,j,k}) \circledast a_k\right)[n] + e_{m,j}[n] \tag{7.6}$$

where once again, $\tau_{m,j,k} = f_\mathrm{s}\|\boldsymbol{r}_m^{\mathrm{mic}} - \boldsymbol{r}_{j,k}^{\mathrm{src}}\|/c$ denotes the time of arrival (TOA) of the k-th image of source j to microphone m in samples and $e_{m,j}[n]$ captures measurements and modeling errors. Interestingly, if we assume that the geometrical parameters $\boldsymbol{\gamma} = \{\boldsymbol{\tau}, \boldsymbol{\theta}^{\mathrm{in}}, \boldsymbol{\phi}^{\mathrm{in}}, \boldsymbol{\theta}^{\mathrm{out}}, \boldsymbol{\phi}^{\mathrm{out}}\}$ and the directivity profiles $\{g_j^{\mathrm{src}}\}_{j=1}^{J}$ and $\{g_m^{\mathrm{mic}}\}_{m=1}^{M}$ are all *perfectly known*, there is a linear dependency between the measured RIRs and the attenuation filters, which include in particular the wall impulse responses of interest. Namely:

$$\boldsymbol{x} = \mathbf{H}(\boldsymbol{\gamma})\boldsymbol{a} + \boldsymbol{e} \tag{7.7}$$

where $\boldsymbol{a} = [a_1[1:L], \ldots, a_k[1:L]] \in \mathbb{R}^{KL}$ is the concatenation of all attenuation filters (assumed of length L), $\boldsymbol{x} = [x_{1,1}[1:N], \ldots, x_{M,J}[1:N]] \in \mathbb{R}^{MJN}$ is the concatenation of all observed RIRs (truncated to N samples), $\boldsymbol{e}$ is the error vector of the same size, and $\mathbf{H}(\boldsymbol{\gamma}) \in \mathbb{R}^{MJN \times KL}$ is a large matrix with a block-Toeplitz structure. For many practical use cases, say $M = J = 4$, $K = 100$, $N = 800$, $L = 32$, the linear system (7.7) is over-determined. Hence, retrieving $\boldsymbol{a}$ from $\boldsymbol{x}$ given $\mathbf{H}(\boldsymbol{\gamma})$ is amenable to any classical linear least-square solver, such as the *conjugate gradient method* (CGM, [53]). The absorption coefficients $\hat{\alpha}_i[f]$ of each wall i can then be calculated as $1 - |\hat{a}_{k(i)}|^2$ where $k(i)$

[2] See discussions on this in Sec. 6.2.2 and examples of real loudspeaker responses later in Fig. 8.4.

	MAE	CE
With exact geometrical knowledge	0.025	97.1%
With corrupted times of arrival	0.495	15.5%

TABLE 7.2: Mean absolute error (MAE) and percentage of correct estimates (CE) on absorption coefficients obtained using the conjugate gradient method to invert system (7.7) with exact geometrical knowledge $\boldsymbol{\gamma}$ or with corrupted times of arrivals.

is the associated first-order image source. Here, higher order attenuation filters, though modeled and estimated, are *not used to estimate absorption coefficients*.

We tested this approach on simulated data using the `pyroomacoustics` extension developed by Prerak Srivastava during his PhD thesis, which implements model (7.6) using measured source and microphone directivity patterns from the DirPat dataset [49] (See Sec. 8.1.2 for more on this implementation). The Genelec 8020 pattern was used for sources, and a near-omnidirectional AKG C414 pattern was used for microphones. 16 RIRs per room ($M = 4$, $S = 4$) in 100 rooms were simulated at $f_\mathrm{s} = 16$ kHz with image sources up to order 20 using random geometrical and acoustical parameters, drawn in the same way as in Sec. 7.1.1.2. RIRs were cropped to 50 ms and corrupted by additive white Gaussian at 50 dB PSNR.

To run the conjugate gradient method, the matrix $\mathbf{H}(\boldsymbol{\gamma})$ was built using the exact geometrical knowledge and the exact directivity patterns, accounting for all K *audible* image sources within the cropped RIRs, with $K \in [\![38, 290]\!]$ depending on the geometry. The attenuation filters are assumed of length $L = 32$ (2 ms). As in Sec. 7.1.1.2, we use as metrics the mean absolute error (MAE) on absorption coefficients, and the total percentage of *correctly estimated* (CE) coefficients, *i.e.* with an error smaller than 0.1. As can be seen in the first row of Table 7.2, obtained errors are very low and the recovery rate very high, which demonstrates empirically that the linear inverse problem at hand is well-posed in a wide range of configurations, at least for first-order image sources. It is important to note, however, that estimation errors were an order of magnitude larger for higher-order attenuation filters. This can be explained by the fact that higher order echoes typically arrive later and are hence more mixed in RIRs. This observation may lead one to believe that modeling all audible image sources in $\mathbf{H}(\boldsymbol{\gamma})$ is unnecessary, and that one should restrict the model to the $K = 7$ echoes of order 0 and 1 instead. Yet, our experiments showed that fixing K to 7 yielded much higher errors. This suggests that while estimation of higher order echoes might not be accurate due to fundamental ambiguities (*e.g.*, permutations and heavy interference), explicitly modeling those echoes as *short, time-domain filters* rather than mere least-square residuals is beneficial to the inverse method.

To test the sensitivity of the approach to geometrical errors, we corrupted the echo times of arrival used to construct $\mathbf{H}(\boldsymbol{\gamma})$ with random independent Gaussian delays following $\mathcal{N}(0, 2\sigma_\mathrm{geo}^2 f_\mathrm{s}/c)$ with $\sigma_\mathrm{geo} = 2$ cm. This corresponds roughly to assuming geometrical errors on the source and microphone positions with a standard deviation of 2 cm, and no errors on wall positions. Unfortunately, as can be seen in the second row of Table 7.2, the approach proves extremely sensitive to such errors, yielding results comparable to randomness. This is likely because in the time domain, the information on reflection filters is sharply localized within a few samples around the true times of arrival. This was avoided in the approach of Sec. 7.1.1, by going to the short-time Fourier domain and discarding the phase, building some natural tolerance to delay errors.

	MAE	CE
Without delay corrections	0.495	15.5%
With delay corrections	0.159	56.2%
After 6 iterations	0.122	67.2%

TABLE 7.3: Mean absolute error (MAE) and percentage of correct estimates (CE) on absorption coefficients obtained using the conjugate gradient method (CGM) to invert system (7.7) in the presence of delay errors, with or with pre-correcting delays using (7.8), and after 6 alternations with CGM.

7.1.2.2 Correcting Geometrical Errors

While the results of the previous section are encouraging regarding the well-posedness of our inverse problem in a realistic setting involving directive devices, the approach is not applicable as it is to real-world measurements, where inevitable geometrical errors will occur. To fight this, a natural approach is to attempt to *correct* the delay errors before solving the inverse problem. In the following, we will neglect the dependency of $\mathbf{H}(\boldsymbol{\gamma})$ on angular errors, and assume it mostly depends on additive delay errors on times of arrival, denoted $\boldsymbol{\Delta} = \{\Delta_{m,j,k}\}_{m,j,k}$ such that $\tilde{\tau}_{m,j,k} = \tau_{m,j,k} + \Delta_{m,j,k}$. This assumption is reasonable if we consider that the directivity profiles of sources and microphones are smooth/locally constant. Note that delay errors also affect the amplitude of echoes, due to their modified propagation time. We then propose to estimate these errors prior to estimating attenuation filters by solving the following non-linear least-square problem:

$$\underset{\boldsymbol{\Delta}}{\operatorname{argmin}} \|\boldsymbol{x} - \mathbf{H}(\boldsymbol{\gamma}, \boldsymbol{\Delta})\tilde{\boldsymbol{a}}\|_2^2 \tag{7.8}$$

where $\tilde{\boldsymbol{a}}$ is built by setting the attenuation filters of all image sources to simple Dirac impulses, *i.e.*, $\tilde{a}_{m,j,k}[n] = [1, 0, 0, \dots]$. Note that since delay errors are assumed independent, this problem can be solved independently for each individual RIR $x_{m,j}[n]$. In practice, we used the gradient descent method with momentum ADAM [54] initialized by $\boldsymbol{\Delta} = \mathbf{0}$. Due to the non-convexity and non linearity of the problem, two key additional ideas were needed to obtain satisfying results:

- To prevent delay-error values from drifting too far away from zero, we *reparameterized* them via $\Delta_{m,j,k} = \Delta_{\max} \tanh(u_{m,j,k})$ and optimized over $\boldsymbol{u}$ instead, to ensure $\Delta_{m,j,k} \in [-\Delta_{\max}, \Delta_{\max}]$. $\Delta_{\max}$ was set so that the interval contains 95% of delay errors under our Gaussian model, with standard deviation $2\sigma_{\text{geo}}^2 f_s/c$. While this constraint prevents the correct estimation of a fraction of the largest delay errors, it proved beneficial to the overall performance.

- We observed that the non-convexity of problem (7.8) mostly arose from the sharpness of echoes in time, resulting in wide variations of the loss under small delay changes. To alleviate this, both the RIRs and all the filters involved in the construction of $\mathbf{H}$ were pre-processed by *convolution with a Gaussian filter*, with standard deviation set to 1 sample. This common trick used to regularize deconvolution methods turned out to dramatically improve performance.

The results obtained after delay corrections are shown in the second row of Table 7.3 and are compared to the previously-shown results without any delay corrections in the first row. As can be seen, the approach significantly improve performance, reaching a 56.2% recovery rate.

Finally, we tried *iterating this process* by plugging estimates $\tilde{\boldsymbol{a}}$ of $\boldsymbol{a}$ by CGM into (7.8), before re-estimating $\boldsymbol{a}$. Crucially, higher-order attenuation filters were recomputed using first-order estimates

and formula (7.2) instead of using the CGM estimates directly, which are known to be poor. Encouragingly, errors are further reduced and the recovery rate reaches 67.2% after 6 iterations. Adding more iterations did not show to bring further improvement.

7.1.2.3 Conclusion on Time-Domain Absorption Estimation

We presented a time-domain inverse method that estimates the absorption coefficients of individual walls given a set of RIRs in a shoebox room, using realistic source and microphone models of known directivities and in the presence of geometrical errors. Although the mean estimation errors (≈ 0.12) are not as low as the ones obtained using the spectrogram-domain approach of Sec. é7.1.1 (≈ 0.08), the explicit use of realistic source and microphone models brings us closer to applying such physics-based inverse methods to real room-acoustic diagnosis. The main remaining difficulty at this stage is, in fact, experimental. One would need multiple RIR measurements in a near-shoebox room *combined with* full-sphere time-domain directivity patterns of all used devices *and* sufficiently precise geometrical annotations, including not only distances but also orientations of all devices in a fixed reference frame. To the best of our knowledge, this type of data is not publicly available at present. Indeed, while source and microphone manufacturers generally provide data on the directive responses of their devices, they are often limited to a restricted set of frequencies and angles, and often exclude phase responses. This is because the full-sphere, time-domain directivity measurement of a device is time-consuming, complex and costly, requiring a dedicated apparatus in a large-enough anechoic chamber and precise experimental protocols to ensure reproducibility. Obtaining such measurements is in our road map for future research (see Chap. 10). These will in turn allow a rigorous assessment of the sensitivity of the proposed approach to any remaining mismatch between model (6.27) and real-world room acoustics, such as, *e.g.*, angular errors, diffraction, or the intricate interaction between directive spherical waves and impedance surfaces.

7.2 Shoebox Room Parameter Estimation

Associated publication: [S30, S36]

In this section, we present a physics-based inverse method that jointly recovers all geometrical and acoustical parameters from a set of RIRs measured in a shoebox room with a single source and a microphone array of known internal geometry. At first this may sound like a bold objective, since it constitutes a strictly more general inverse problem than the one addressed in Sec. 7.1, which itself proved very challenging and remains far from fully solved. The catch is that we will make use of a much simpler and less realistic forward model, namely, the *vanilla* image source method (ISM) as defined in (6.26). We hence momentarily depart from the objective of real-world applicability, and focus instead on the more fundamental question of whether *"hearing the walls of a room"* is solvable at all, for a well-specified forward model, under a broad-enough range of controlled conditions.

Rather than *theoretically* investigating the well-posedness of the problem, which seems out of reach of current mathematical methods, we frame this as a question of *algorithmic invertibility*: given a set of RIRs outputted by the vanilla shoebox ISM, is there an algorithm that can recover all of its 18 input parameters, namely:

- The 3 dimensions of the room;

- The 6-degrees-of-freedom translation and orientation of the room in the microphone array coordinate frame;
- The 3-dimensional source position in the microphone array coordinate frame;
- 1 absorption coefficient for each of the 6 room surfaces ?

We exhibit a multi-stage algorithm that achieves this, and use it to provide empirical evidence that the answer to this question is *yes*, under a broad range of randomized input parameters, for sufficiently large microphone arrays and sufficiently high frequencies of sampling. The algorithm proceeds as follows:

1. Recover the 3D positions and attenuation coefficients of all the image sources that are audible in the early part of the RIRs, using a greedy, off-the-grid approach (Sec. 7.2.1);
2. From the obtained image source point cloud, estimate the directions of the 3 room axes by identifying the underlying orthogonal lattice structure (Sec. 7.2.2.1);
3. Based on the room orientation, identify the first-order image sources, which are the closest ones to the true source along each oriented axis (Sec. 7.2.2.2);
4. Based on the positions and attenuation coefficients of the zero-and-first-order image sources, recover the wall positions and their absorption coefficients (Sec. 7.2.2.3).

The code for this entire procedure is open source and available at `https://github.com/Sprunckt/acoustic-sfw`.

7.2.1 Gridless 3D Localization of Image Sources

Associated publication: [S30]

7.2.1.1 Convex Relaxation to Radon Measures

Recall from Sec. 6.2.1 that the inhomogeneous wave equation (6.7)-(6.8) with rigid shoebox boundary conditions has the same causal solution as the following free-field equation

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}h(\boldsymbol{r},t) - \Delta h(\boldsymbol{r},t) = \sum_{k=0}^{\infty} a_k \delta(t)\delta(\boldsymbol{r}-\boldsymbol{r}_k^{\text{src}}) \tag{7.9}$$

as long as $a_k = 1$ for all k. We consider here the extension to $a_k \in [0,1]$ to approximately model wall absorption, at the cost of losing direct equivalence to the wave equation (See Sec. 6.2.2.1). We now further relax this equation to an arbitrary *source mass distribution* ψ, yielding:

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}h(\boldsymbol{r},t) - \Delta h(\boldsymbol{r},t) = \psi(\boldsymbol{r})\delta(t) \tag{7.10}$$

where ψ belongs to the space $\mathcal{M}(\mathbb{R}^3)$ of (positive) *Radon measures*, *i.e.*, the topological dual of the space of continuous functions on $\mathbb{R}^3$ that vanish at infinity [55]. Generalizing the derivation of

Sec. 6.2.1, the solution of (7.10) is given by the following spatial convolution product with the Green function $h_0(t, \boldsymbol{r}) = \delta(t - \|\boldsymbol{r}\|/c)/4\pi\|\boldsymbol{r}\|$:

$$h(\boldsymbol{r}, t) = (h_0(\cdot, t) * \psi)(\boldsymbol{r}) = \int_{\boldsymbol{r}' \in \mathbb{R}^3} \frac{\delta(t - \|\boldsymbol{r} - \boldsymbol{r}'\|_2/c)}{4\pi\|\boldsymbol{r} - \boldsymbol{r}'\|_2} \psi(\boldsymbol{r}') d\boldsymbol{r}'. \tag{7.11}$$

Applying the microphone sampling model $x_m[n] = \left(g_m^{\text{mic}} * h(\boldsymbol{r}_m^{\text{mic}}, \cdot)\right)(n/f_s)$ of Sec. 6.3 gives:

$$x_m[n] = \int_{\boldsymbol{r} \in \mathbb{R}^3} \gamma_{m,n}(\boldsymbol{r}) d\psi(\boldsymbol{r}) = \langle \gamma_{m,n}, \psi \rangle \text{ with } \gamma_{m,n}(\boldsymbol{r}) \overset{\text{def}}{=} \frac{g_m^{\text{mic}}\left(n/f_s - \left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}\right\|_2/c\right)}{4\pi \left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}\right\|_2}. \tag{7.12}$$

Observe that the *non-linear*, *non-convex* function $\boldsymbol{\gamma} : \mathbb{R}^3 \to \mathbb{R}^{MN}$ can be seen as the representative of an infinite-dimensional *linear* operator[3] $\boldsymbol{\Gamma} : \mathcal{M}(\mathbb{R}^3) \to \mathbb{R}^{MN}$ that maps an arbitrary source mass distribution ψ to its corresponding observation vector $\boldsymbol{x}$. If we now particularize back (7.12) to a discrete measure as in (7.9):

$$\psi_{\boldsymbol{a}, \boldsymbol{r}^{\text{src}}}(\boldsymbol{r}) \overset{\text{def}}{=} \sum_{k=0}^{K} a_k \delta_{\boldsymbol{r}_k^{\text{src}}}(\boldsymbol{r}), \; K \in \mathbb{N} \cup \infty, \tag{7.13}$$

we recover the discrete vanilla image source model (6.26):

$$\boldsymbol{x} = \sum_{k=0}^{K} a_k \boldsymbol{\gamma}(\boldsymbol{r}_k^{\text{src}}) = \boldsymbol{\Gamma} \psi_{\boldsymbol{a}, \boldsymbol{r}^{\text{src}}} \tag{7.14}$$

$$\Rightarrow x_m[n] = \sum_{k=0}^{K} a_k \frac{g_m^{\text{mic}}\left(n/f_s - \left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\right\|_2/c\right)}{4\pi \left\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\right\|_2}. \tag{6.26}$$

Let $\mathcal{M}_*(\mathbb{R}^3) \subset \mathcal{M}(\mathbb{R}^3)$ denote the subset of sparse Radon measures of the form $\psi_{\boldsymbol{a}, \boldsymbol{r}^{\text{src}}}$, with $\boldsymbol{a} \in \mathbb{R}_{\geq 0}^{K+1}$, $\boldsymbol{r}^{\text{src}} \in (\mathbb{R}^3)^{K+1}$ and $K \in \mathbb{N}$, *i.e.*, measures that are finite positive combinations of spatial Dirac spikes. The inverse problem of recovering the attenuation coefficients (Dirac *amplitudes*) and positions of image sources given noisy observations $\boldsymbol{x}$ can now be formulated as follows:

$$\underset{\psi_{\boldsymbol{a}, \boldsymbol{r}^{\text{src}}} \in \mathcal{M}_*(\mathbb{R}^3)}{\operatorname{argmin}} \; \|\boldsymbol{x} - \boldsymbol{\Gamma} \psi_{\boldsymbol{a}, \boldsymbol{r}^{\text{src}}}\|_2^2. \tag{7.15}$$

This belongs to a general class of problems called *super-resolution* in the literature [56, 55, 57, 58], where the goal is to recover the continuous locations of a set of spikes given discrete linear observations over their measure. Rather than solving (7.15), which is a non-convex optimization problem on the attenuation coefficients and positions of the image sources, we follow the approach in [55] and consider a *convex relaxation* of this problem to the whole space of Radon measures:

$$\underset{\psi \in \mathcal{M}(\mathbb{R}^3)}{\operatorname{argmin}} \; \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\Gamma}\psi\|_2^2 + \lambda \|\psi\|_{\text{TV}} \tag{7.16}$$

[3] Strictly speaking, this operator is not well defined because $\boldsymbol{\gamma}$ is singular at each microphone position. In theory, one should change the integration domain in (7.12) to $\mathbb{R}_\varepsilon \overset{\text{def}}{=} \mathbb{R}^3 \setminus \cup_{m=1}^{M} B(\boldsymbol{r}_m^{\text{mic}}, \varepsilon)$ for a fixed $\varepsilon > 0$, and only consider measures $\psi \in \mathcal{M}(\mathbb{R}_\varepsilon)$. In practice, this adjustment is harmless as long as a minimum separation distance ε is assumed between the image sources and the microphones, and is hence ignored here for clarity.

where $\lambda \in \mathbb{R}_{>0}$ is a parameter and $\|\psi\|_{\text{TV}}$ denotes the *total variation norm* of the Radon measure ψ. For a sparse measure $\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}} \in \mathcal{M}_*(\mathbb{R}^3)$ we have in particular:

$$\|\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}}\|_{\text{TV}} = \sum_{k=0}^{K} |a_k| = \|\boldsymbol{a}\|_1. \tag{7.17}$$

Hence, the second term of (7.16) can be seen as a sparsity-inducing regularizer. By analogy with the finite-dimensional sparse setting, this problem has been coined the Beurling-LASSO (BLASSO) in [59] and offers good measure-reconstruction guarantees in low noise regimes. In particular, [60] shows that there always exists a sparse solution $\psi_{\boldsymbol{a},\boldsymbol{r}^{\text{src}}} \in \mathcal{M}_*(\mathbb{R}^3)$ to problem (7.16) and [57] studies its basins of attraction.

7.2.1.2 Resolution with the Sliding Franck-Wolfe Algorithm

In order to solve (7.16) numerically, we adapt the Sliding Frank-Wolfe algorithm proposed in [55], which is briefly reviewed below. Let us denote by $\boldsymbol{a}^{(i)} \in \mathbb{R}_{\geq 0}^{Q_i}$ and $\boldsymbol{r}^{(i)} \in (\mathbb{R}^3)^{Q_i}$ the lists of Q_i spike amplitudes and positions estimated at iteration i, with $\boldsymbol{a}^{(0)} = \boldsymbol{r}^{(0)} = \emptyset$. At iteration $i+1$, the following four steps are performed:

- **Step 1:** A new spike location $\boldsymbol{r}_{Q_i+1}$ is first added to $\boldsymbol{r}^{(i+1)}$ by maximizing the following *dual*, non-convex objective based on the current residual $\boldsymbol{y}^{(i)} \overset{\text{def}}{=} \boldsymbol{x} - \boldsymbol{\Gamma}\psi_{\boldsymbol{a}^{(i)},\boldsymbol{r}^{(i)}}$:

$$\max_{\boldsymbol{r}\in\mathbb{R}^3} \eta^{(i)}(\boldsymbol{r}) \overset{\text{def}}{=} \left[\boldsymbol{\Gamma}^*(\boldsymbol{y}^{(i)})\right](\boldsymbol{r}) = \sum_{m,n} y_m^{(i)}[n]\gamma_{m,n}(\boldsymbol{r}) \tag{7.18}$$

  where $\boldsymbol{\Gamma}^*$ denotes the *Hermitian adjoint* of $\boldsymbol{\Gamma}$. We use the parallel implementation of the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm in `scipy` for this [61].

- **Step 2:** The whole list of amplitudes $\boldsymbol{a}^{(i+1)}$ is updated by minimizing (7.16) over $\boldsymbol{a}$ only. This amounts to a classical non-negative LASSO convex optimization problem for which efficient solvers are available (we use the `scikit-learn` implementation [62]).

- **Step 3** (*Sliding*): The value of the cost function in (7.16) is further decreased by jointly refining all the values in $\boldsymbol{a}^{(i+1)}$ and $\boldsymbol{r}^{(i+1)}$ through non-convex local search. We use the bounded version of BFGS in [61] to preserve positive amplitudes.

- **Step 4:** The spikes whose amplitudes are lower than a threshold a_{min} are removed from $\boldsymbol{a}^{(i+1)}$ and $\boldsymbol{r}^{(i+1)}$.

To improve the optimization and reduce the computational time, we introduce a number of modifications, briefly outlined below.

- The non-convexity of **Step 1** makes it very sensitive to initialization. In practice, we initialized BFGS by multiple points on discrete spherical grids centered around the microphones, whose radius was determined based on time-domain peaks in the RIRs.

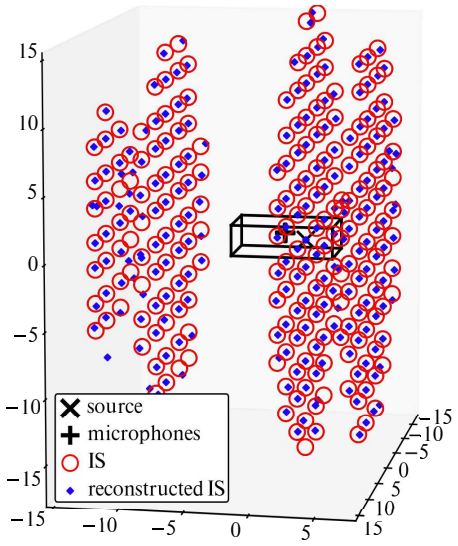


FIGURE 7.2: 3D plot of a room and the corresponding target and reconstructed image sources using the algorithm of Sec 7.2.1.2 for a 32-channel spherical microphone array with diameter 16.8 cm, $f_s = 16$ kHz, $T_{\text{max}} = 50$ ms and no noise. The corresponding RIR at one microphone is shown in Fig. 7.3.

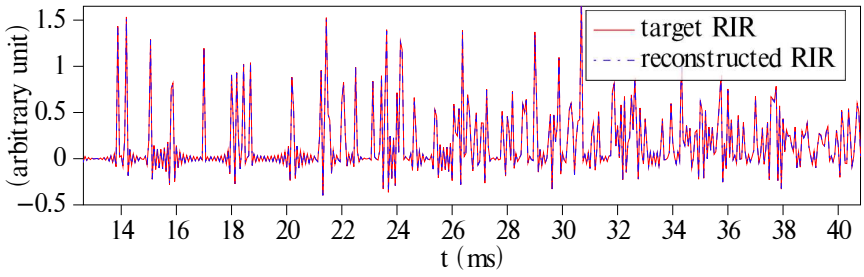


FIGURE 7.3: Excerpt from a room impulse response (RIR) at one microphone and its reconstruction using (6.26), under the setup described in Fig. 7.2.

- The algorithm is first ran on an early-cropped version of the RIRs, where echoes are better separated. The resulting spikes are then used as an initialization to run the algorithm on progressively longer signals, and this is repeated until the desired signal length N is reached.
- The sliding step 3 is only performed *on the very last iteration*, as suggested in [58].
- Spikes with an amplitude less than 0.1 are deleted before and after sliding to decrease the number of false positives.

λ will be fixed to $3 \cdot 10^{-5}$ throughout all the following experiments, based on a preliminary manual tuning.

7.2.1.3 Results on Image Source Localization

We present here some of the numerical results obtained by applying the algorithm described in the previous section to a set of 200 simulated shoebox rooms containing an omnidirectional source and a spherical array of 32 microphones. The geometry of the array is the same as the em32 Eigenmike®

TABLE 7.4: Mean room volume ($\overline{\text{V}}$), Recall (R), Precision (P) and mean radial ($\overline{\text{RE}}$), angular ($\overline{\text{AE}}$), Euclidian ($\overline{\text{EE}}$) and attenuation coefficient ($\overline{\text{ACE}}$) errors among the recovered sources for varying numbers of image sources using the algorithm of Sec. 7.2.1.2, with f_s=16 kHz, d=16.8 cm and no noise.

# of IS	$\overline{\text{V}}$(m^3)	R(%)	P(%)	$\overline{\text{RE}}$(mm)	$\overline{\text{AE}}$(°)	$\overline{\text{EE}}$(mm)	$\overline{\text{ACE}}$
0-150	214	94.3	81.8	0.069	0.38	94	0.042
150-300	102	92.1	83.1	0.099	0.36	91	0.029
300-500	56	86.1	78.1	0.151	0.38	97	0.025
500-1323	30	57.3	51.6	0.300	0.46	108	0.027

(diameter of 8.4 cm), but scaled by various factors. We use a unique ideal low-pass filter with cutoff frequency $f_s/2$ (Eq. 1.5) to model the response of all the microphones, which are assumed omnidirectional *i.e.*, $g_m^{\text{mic}}(t) = \text{sinc}(f_s t)$ for all m. As in the experiments of Sec. 7.1, the rooms' lengths and widths in meters are sampled uniformly at random in $[2, 10]$, while the heights are taken in $[2, 5]$. The absorption coefficient α_i of each individual wall $i \in [\![1, 6]\!]$ is sampled uniformly at random in $[0.01, 0.3]$. The source and the array are then placed randomly in each room, with a separation constraint of 1 m to the walls and between each other. The array is also rotated on itself at random. While full-lenght RIRs are simulated with echoes up to order 20, the proposed algorithm is only fed with the first $T_{\text{max}} = (N-1)/f_s = 50$ ms of each channel. This allows us to consider as targets all the image sources that are *audible* by all the microphones, *i.e.*, whose distances are inferior to $cT_{\text{max}} = 17.15$ m, independently of the room dimension. For each test room, the ground truth image source positions and attenuation coefficients are obtained using the pyroomacoustics simulator [42]. An observation vector is then built using (6.26) and adding white Gaussian noise with a desired peak signal-to-noise ratio (PSNR). An example of room, image source point cloud, RIR, recovered image sources and reconstructed RIR is shown in Fig. 7.3 and Fig. 7.2.

To evaluate the efficiency of the method, a source is considered *recovered* if at least one estimated source is at an angular distance of less than $2°$ and a radial distance of less than 1 cm from it with respect to the array center. We then calculate the *recall* (ratio of true image sources recovered), the *precision* (ratio of estimated sources assigned to a recovered source, discarding doubles), as well as the mean radial, angular and Euclidean errors and the mean error on attenuation coefficients, where the means are calculated *over recovered sources only*. Because the number of image sources that are audible within 50 ms of RIRs varies widely depending on the room's volume, the test set is sliced into four subsets, as detailed in the first two columns of Table 7.4. The remaining columns report the metrics for a sampling frequency $f_s = 16$ kHz, an array diameter $d = 16.8$ cm (x2) and no noise.

A recall rate of over 90% for large and medium sized rooms is obtained. The precision is over 80%, indicating few false positives and a reasonable prediction of the number of audible sources. As expected, the recall and precision significantly drop in smaller rooms, where the *echo density* [63] is higher, making the image sources harder to separate. The strength of the proposed gridless approach is revealed by the mean radial and angular errors, which are below tenths of millimeters and fractions of degrees. As a first comparison, the best previously reported results we are aware of in a similar simulated setting are in [64], where an average of 25 nearest image sources are localized with a mean angular error of $4.3°$. Note however that [64] is a *blind* method. As a second comparison, ignoring any basis-mismatch issue and assuming *perfect localization*, a sparse method in discrete space such as [21] would require a spatial grid of at least 111 million points to achieve errors below $1°$ and 1 cm over the same range. This is four orders of magnitude larger than the initialization grids used in the

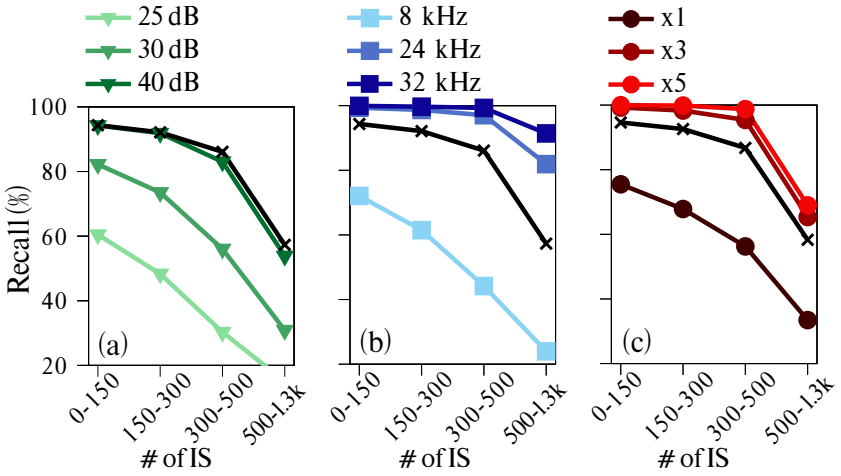


FIGURE 7.4: Recall for varying PSNR (a), sampling frequency (b) and microphone array scaling (c). The default values (×) are noiseless, 16 kHz and x2.

proposed approach.

Notice that the obtained mean Euclidean errors are of a few centimeters. This is because they grow with the source distances, as expected due to the compact spherical geometry of the array. The attenuation coefficients of the recovered sources, which take values in $[0, 1]$[4], are also accurately estimated, with mean errors around 0.03. Logically, these errors are slightly larger in large rooms because attenuation coefficients are larger in that case, due to fewer reflections on the walls. Note that only 3 out of 1200 first-order image sources were missed on this test, while all 200 true sources were recovered.

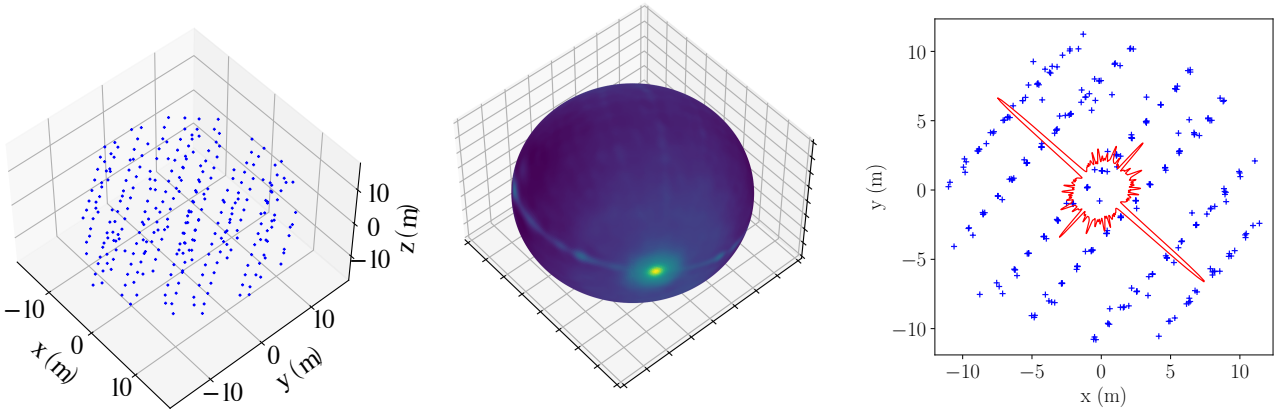
Finally, the impact of PSNR, f_s, and array diameter on the recall is reported in Fig. 7.4. Remarkably, it can be observed that either increasing f_s to 32 kHz or the array diameter to 42 cm (x5) brings the recovery rate near 100% for rooms with up to 500 image sources. Conversely, decreasing by half these parameters significantly degrades performance. This is expected as they are known to control the source localization accuracy for compact microphone arrays. Adding noise to the observations does not significantly affect the recovery rate at 40 dB PSNR, but quickly degrades it for PSNRs below 30 dB. Nevertheless, it was observed that the recall values for a PSNR of 30 dB could be restored near the noiseless level by simply considering an angular recovery threshold of $6°$ instead of $2°$. This shows an encouraging stability of the method, given that for such PSNRs the peaks of many echoes in the RIRs fell below the noise standard deviation.

7.2.2 Fully Reversing the Shoebox Image-Source Model

Associated publication: [S36]

Equipped with the algorithm presented in the previous section, we are now faced with the following problem: how to recover the geometrical parameters of the scene, as well as the absorption coefficients of individual walls, given an (unlabeled) point cloud of image source positions and their amplitudes, possibly corrupted by errors and spurious or missing estimates? We present an approach in three steps, that are detailed in each of the following subsections.

[4]Following (6.17), recall that attenuation coefficients are products of reflection coefficients of the form $\sqrt{1-\alpha_i}$ where here α_i is drawn in $[0.01, 0.3]$.



(A) (B) (C)

FIGURE 7.5: (A) Reconstructed image-source point cloud using the algorithm of Sec. 7.2.1.2. (B) Associated J_3^σ score plotted on the sphere (brighter is higher). (C) Projection of the estimated sources on $\hat{\boldsymbol{e}}_1^\perp$ (blue) and the associated 2D $J_{2,\hat{\boldsymbol{e}}_1}^\sigma$ score (red). Sharp peaks are observed in the directions of wall normals.

7.2.2.1 Room Orientation Estimation

Let us first consider the task of recovering the room orientation from the unlabeled image source point cloud. The key idea is to estimate its underlying *orthogonal grid structure*, which is apparent in the example of Fig. 7.2. The projected coordinates of image sources onto a normal vector to a wall will form clusters, each cluster corresponding to a plane of image sources parallel to this wall. In contrast, projecting image sources onto a randomly chosen vector will not form clusters, but instead spread out over the entire range of possible values. In other words, the room axes are orthogonal to image-source planes that are generated by pairs of facing walls. They are hence expected to *maximize the number of orthogonalities* to vectors formed by *pairs of image sources*. Our method seeks to exploit this structure by scoring room-axis candidates according to their orthogonality to the directions generated by image-source pairs. To this aim, let us define the *orthogonality-indicator* function f_D as follows:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D, \quad f_D(\boldsymbol{u}, \boldsymbol{v}) = \begin{cases} 1 & \text{if } \boldsymbol{u} \perp \boldsymbol{v} \\ 0 & \text{otherwise.} \end{cases} \tag{7.19}$$

Let $\mathcal{G} \subset \mathbb{R}^3$ be a finite set of image source locations estimated by the algorithm of Sec. 7.2.1.2. Let us consider the following optimization problem:

$$\max_{\|\boldsymbol{u}\|_2=1} J_3(\boldsymbol{u}, \mathcal{G}), \quad \text{where} \quad J_3(\boldsymbol{u}, \mathcal{G}) = \sum_{\boldsymbol{s},\boldsymbol{p} \in \mathcal{G}} f_3(\boldsymbol{u}, \boldsymbol{s} - \boldsymbol{p}). \tag{7.20}$$

Recall from Sec. 6.2.1 that the infinite image source point cloud for a shoebox room is given by:

$$\mathcal{G}_\infty = \{\boldsymbol{\epsilon} \odot \boldsymbol{r}_0^{\text{src}} + 2\boldsymbol{q} \odot [L_\text{x}, L_\text{y}, L_\text{z}]^\top \mid \boldsymbol{\epsilon} \in \{0,1\}^3, \boldsymbol{q} \in \mathbb{Z}^3\}. \tag{6.12}$$

It can be shown ([S36, Proposition 1]) that if $\mathcal{G}$ is a subset of $\mathcal{G}_\infty$ generated by $\boldsymbol{q} \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}$ where $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are integer intervals, then the solution $\boldsymbol{u}^*$ of problem (7.20) is indeed a wall normal, *i.e.*, a room axis. While it is technically possible to carefully craft adversarial subsets of $\mathcal{G}_\infty$ that would

Algorithm 1 Orientation estimation

Require: Image sources $(\boldsymbol{r}_k)_{k=1}^K$
Ensure: Estimated room orthonormal basis $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3$
$\hat{\boldsymbol{e}}_1 \leftarrow \operatorname{argmin}_{\boldsymbol{u}\in\mathcal{S}^2_{\text{discr}}} J_3^{0.01}$
for $\sigma \in [0.01, 0.005, 0.0005]$ **do**
  $\hat{\boldsymbol{e}}_1 \leftarrow \text{local\_descent}(\hat{\boldsymbol{e}}_1, J_3^\sigma)$
end for
$\hat{\boldsymbol{e}}_2 \leftarrow \operatorname{argmin}_{\boldsymbol{u}\in\mathcal{S}^1_{\text{discr}}} J_{2,\hat{\boldsymbol{e}}_1}^{0.01}$
for $\sigma \in [0.01, 0.005, 0.0005]$ **do**
  $\hat{\boldsymbol{e}}_2 \leftarrow \text{local\_descent}(\hat{\boldsymbol{e}}_2, J_{2,\hat{\boldsymbol{e}}_1}^\sigma)$
end for
$\hat{\boldsymbol{e}}_3 = \hat{\boldsymbol{e}}_1 \times \hat{\boldsymbol{e}}_2$

make this fail, if we assume that the estimated set $\mathcal{G}$ misses image sources at random, the probability of encountering such situations is vanishingly small and hence not an issue in practice. However, the estimated locations will typically be *noisy*, while f_D only captures *exact* orthogonalities, making it unusable in practice. Hence, we approximate it using a Gaussian kernel

$$f_D^\sigma(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-\frac{1}{2\sigma^2}\left(\frac{\boldsymbol{u}.\boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}\right)^2\right), \tag{7.21}$$

such that $\lim_{\sigma\to 0} f_D^\sigma = f_D$ in the pointwise sense. The scale parameter σ controls the tightness of the approximation and plays a regularizing role with respect to the error committed in the localization of image sources. A small σ will yield a noisy loss function if the source localization error is high. Conversely, a large σ will results in poorer precision on room orientation recovery. As we are searching for an optimal *unit* vector, the regularized score function J_3^σ can be re-parameterized in spherical coordinates by two angles $(\theta, \phi) \in [0, 2\pi[\times[0, \pi[$ in the microphone array's reference frame:

$$J_3^\sigma(\theta, \phi, \mathcal{G}) = \sum_{\boldsymbol{s},\boldsymbol{p}\in\mathcal{G}} f_3^\sigma(\boldsymbol{u}(\theta, \phi), \boldsymbol{s} - \boldsymbol{p}) \tag{7.22}$$

where $\boldsymbol{u}(\theta, \phi)$ is the unit vector defined by (θ, ϕ). We use the parallel `scipy` implementation of the *BFGS* algorithm [61] to maximize J_3^σ. Due to the non-convexity of the problem, we initialize the optimization algorithm on a finely meshed half-sphere $\mathcal{S}^2_{\text{discr}}$. To further reduce the chance of stopping at a local minimum, we begin with a high value of the scale parameter and perform the optimization with gradually decreasing values[5], namely, $\sigma = 0.01, 0.005, 0.0005$. This process yields an accurate, off-the-grid reconstruction of a first room axis $\hat{\boldsymbol{e}}_1$, given a sufficiently accurate image-source point cloud. We can then proceed in a greedy manner by projecting $\mathcal{G}$ onto $\hat{\boldsymbol{e}}_1^\perp$ and maximizing:

$$J_{2,\hat{\boldsymbol{e}}_1}^\sigma(\theta, \mathcal{G}) = \sum_{\boldsymbol{s},\boldsymbol{p}\in\mathcal{G}} f_2^\sigma(\boldsymbol{v}(\theta), \mathcal{P}_{\hat{\boldsymbol{e}}_1^\perp}(\boldsymbol{s} - \boldsymbol{p})) \quad \forall\theta \in [0, 2\pi[. \tag{7.23}$$

to obtain a second room axis $\hat{\boldsymbol{e}}_2$, where $\boldsymbol{v}(\theta)$ is the unit vector of angle θ inside the plane $\hat{\boldsymbol{e}}_1^\perp$. The third vector is finally obtained by cross product: $\hat{\boldsymbol{e}}_3 = \hat{\boldsymbol{e}}_1 \times \hat{\boldsymbol{e}}_2$. The full process is summarized in

[5]These same values of σ are used throughout all experiments without any specific tuning.

Algorithm 1. Fig. 7.5 shows examples of score functions J_3^σ and $J_{2,\boldsymbol{u}}^\sigma$. As can be seen, they feature maxima along the room axes.

7.2.2.2 First-Order Image Source Identification

Once the room orientation has been estimated, we seek to identify which of the estimated image sources are of first order. We leverage the fact that the zero-th order image source, *i.e.*, the true source, can be straightforwardly identified. Indeed, it is necessarily the closest one to the microphone array's center. It is also accurately localized, since the direct path is generally well separated from reflections in RIRs. We then cast a cone from the true source in each reconstructed direction $\hat{\boldsymbol{e}}_d$ and their opposite $-\hat{\boldsymbol{e}}_d$. The image source closest to the true source within each cone is picked as a first-order candidate. If the cone is empty (implying that source localization errors are too great) we progressively extend the cone's width until it contains at least one source. As the reconstruction algorithm sometimes produces clusters of sources around the true image-source locations, we assume that any source close to an estimated first order source is a reconstruction artifact. We thus proceed to merge the closest estimated sources. Let $\boldsymbol{r}^*$ be a candidate first-order source, $\mu \in \mathbb{R}_+^*$ a threshold (set to 50 cm in our experiments), and $\{\boldsymbol{r}_1^*, \dots \boldsymbol{r}_P^*\}$ the set of reconstructed sources such that $\|\boldsymbol{r}_p^* - \boldsymbol{r}^*\| < \mu,\ \forall p \in [\![1, P]\!]$. We use a heuristic inspired by [65] to merge the corresponding Diracs and their amplitudes:

$$\hat{a} = \sum_{p=1}^{P} a_p^*, \quad \hat{\boldsymbol{r}} = \sum_{p=1}^{P} \frac{a_p^*}{\hat{a}} \boldsymbol{r}_p^*. \tag{7.24}$$

This procedure gives us estimates for the locations of the six first-order image sources and their associated reflection coefficients.

7.2.2.3 Room Parameter Recovery

Once the positions and amplitudes of zero-th and first order image sources are estimated, recovering the remaining room parameters is relatively straightforward. Indeed, each wall is a bisector between the true source and a first order image source. Let $\hat{\boldsymbol{r}}_{i-}$ and $\hat{\boldsymbol{r}}_{i+}$ be the first order image sources corresponding to room axis i, in the directions $-\hat{\boldsymbol{e}}_i$ and $\hat{\boldsymbol{e}}_i$ from $\hat{\boldsymbol{r}}_0$ respectively. The associated room length is given by:

$$\hat{L}_i = \hat{\boldsymbol{e}}_i.(\hat{\boldsymbol{r}}_{i+} - \hat{\boldsymbol{r}}_{i-})/2. \tag{7.25}$$

Setting the intersection of the walls corresponding to $\hat{\boldsymbol{r}}_{1-}, \hat{\boldsymbol{r}}_{2-}, \hat{\boldsymbol{r}}_{3-}$ as a reference vertex of the room, the translation vector of the room with respect to the source is:

$$\hat{\boldsymbol{\tau}}_{\text{room}} = \frac{1}{2} \begin{pmatrix} \hat{\boldsymbol{e}}_1.(\hat{\boldsymbol{r}}_0 - \hat{\boldsymbol{r}}_{1-}) \\ \hat{\boldsymbol{e}}_2.(\hat{\boldsymbol{r}}_0 - \hat{\boldsymbol{r}}_{2-}) \\ \hat{\boldsymbol{e}}_3.(\hat{\boldsymbol{r}}_0 - \hat{\boldsymbol{r}}_{3-}) \end{pmatrix}. \tag{7.26}$$

This completes the recovery of all 18 input parameters that were used to generate the multichannel RIR: the room orientation $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \hat{\boldsymbol{e}}_3$; the 3D source position $\hat{\boldsymbol{r}}_0$; the room translation with respect to the source $\hat{\boldsymbol{\tau}}_{\text{room}}$; the room dimensions $\hat{L}_1, \hat{L}_2, \hat{L}_3$; and the 6 wall absorption coefficients $\hat{\alpha}_i = 1 - \hat{a}_i^2$ for $i = 1, \dots, 6$.

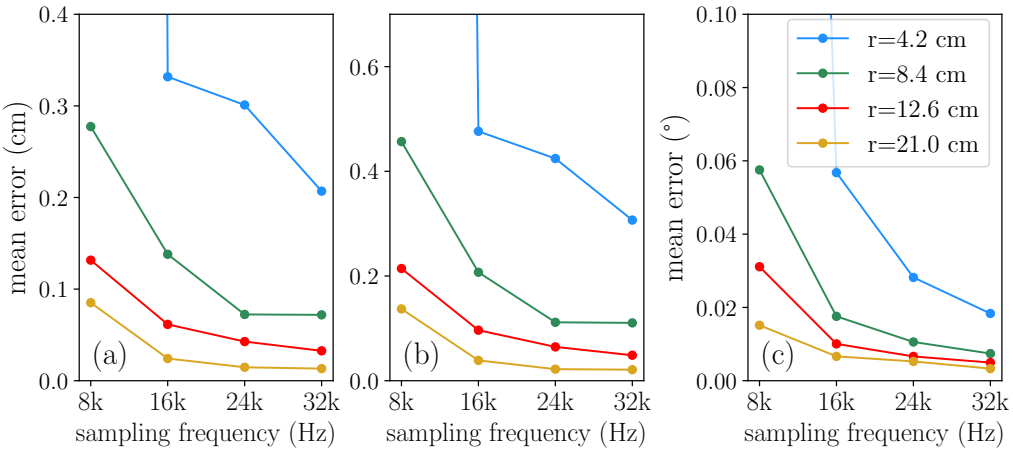


FIGURE 7.6: Mean absolute errors on room dimensions (a), mean Euclidean errors on room center (b), and mean angular error on room orientation (c) as a function of the sampling frequency for varying array radii.

7.2.2.4 Results on Room Parameter Estimation

The proposed room-parameter estimation procedure is now tested on the same set of 200 random room geometries as the one described in Sec. 7.2.1.3, with the same 32-channel microphone array of varying radius, on RIRs simulated using the same shoebox image-source model at the same frequencies of sampling f_s. We first use the algorithm presented in Sec. 7.2.1.2 to estimate the positions and amplitudes of all image sources audible within the first 50 ms of the input multichannel RIR. These are then fed to the 3-step procedure described in the last 3 sections to recover the 18 geometrical and acoustical parameters of interest. The following evaluation metrics are employed:

1. **Orientation and dimensions.** We compute the mean angular errors between the recovered directions $\hat{e}_d$ and the appropriately-matched ground-truth room-axis vectors by taking the arccosine of the dot products. We also compute the mean absolute errors on estimated room dimensions.

2. **Wall absorptions.** Having matched recovered first-order sources to walls, we also compute the mean absolute errors on estimated absorption coefficients $\hat{\alpha}_{1:6}$.

3. **Room translation.** In order to evaluate the room translation estimation, we calculate the room's center in the array's reference frame based on the estimated parameters. We then calculate the mean Euclidean distance to the ground truth center.

4. **RIR extrapolation.** Lastly, we evaluate the global accuracy of the method by re-simulating a RIR $\hat{\boldsymbol{x}}$ corresponding to a new random source-array placement in the room using the vanilla image-source method (6.26) with estimated parameters as input. Using the same sampling rate, we compute the signal-to-error ratio to the true RIR $\boldsymbol{x}$ at the new location:

$$\mathrm{SER}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = 10 \log_{10} \left(\frac{\sum_{i=1}^{NM} x_i^2}{\sum_{i=1}^{NM} (\hat{x}_i - x_i)^2} \right). \tag{7.27}$$

Fig. 7.6 presents the algorithm's performance on the geometry estimation task for varying sampling frequencies and microphone array radii. In accordance with the image source localization results

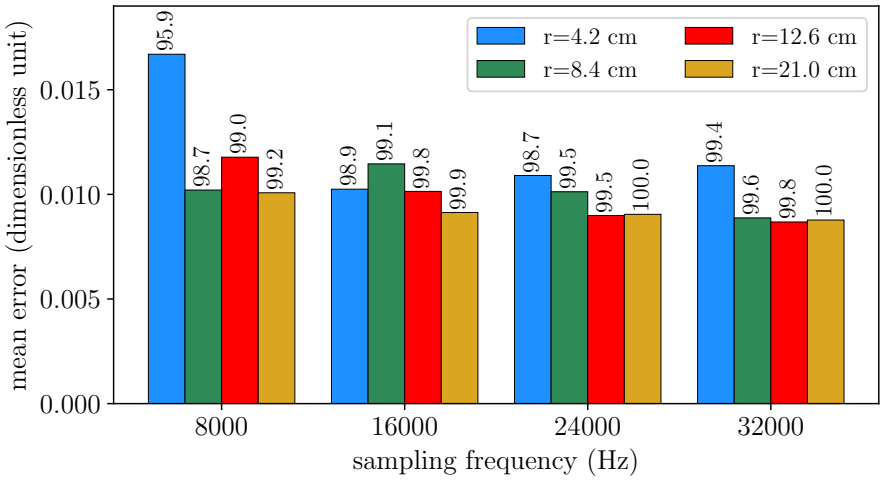


FIGURE 7.7: Mean absolute error on absorption coefficients recovered below a 0.3 threshold for varying array radii and frequency of sampling. The recall for this threshold is indicated above each bar, in percent.

reported in Sec. 7.2.1.3, the accuracy of the estimation improves as the radius or the sampling frequency grow. The lowest resolution (radiud 4.2 cm and $f_s = 8$ kHz) yields a few reconstruction failures that heavily impact the mean errors. For instance, 3.5 % of the errors on wall distance estimation are over 50 cm in that case. These large-error cases vanish for all the larger array sizes and sampling rates considered, the mean error steadily converging towards zero for all three metrics. This empirically supports our claim that the shoebox image-source method is indeed *fully algorithmically reversible* for large enough arrays and frequencies of sampling.

For a frequency of sampling of 24 kHz and the lowest radius, the mean room dimension estimation error is already very low, around 3 mm. This number goes down to 0.15 mm when dilating the array by a factor of 5. Meanwhile, as shown in Fig. 7.6(c), the mean error on room orientation remains under $0.06°$ in all experiments, except for the very lowest resolution. The errors on room center localization are slightly higher, at 4.2 mm with the smallest array at 24 kHz and 0.22 mm after $5\times$ dilation. This increase is expected because estimating the room center couples errors on orientation estimation and source-wall distance estimation.

We then evaluate the estimation of wall absorption coefficients. Here, we observed some rare ($< 1\%$) failures of absorption recovery even for relatively high array resolutions and frequency of sampling. In order to get a more meaningful picture of the error committed, we thus only compute the mean errors over coefficients estimated with an error below 0.3, and consider the rest as outliers (recall that in our simulations, the coefficients take values in $[0.01, 0.3]$). We also compute the recall rates for this threshold. Both metrics are displayed in Fig. 7.7. The obtained mean errors are around 0.01 with a 100% recall rates for the largest array and frequencies of sampling above 24 kHz. While these are very low errors, we do not observe the same convergence towards zero as on geometrical errors. One possible explanation is that we kept the spike estimation algorithm described in Sec. 7.2.1.2 untouched, including the spike pruning steps that discard low amplitude Diracs before and after the final gradient descent. While the first pruning step does seem to help the optimization algorithm, the second step, which aimed at reducing false positives, might cause an issue on amplitude estimation. Rather than deleting the spikes and losing the corresponding amplitudes, a lead for improvement would be to merge the spikes by, *e.g.*, adapting the heuristic presented in [65].

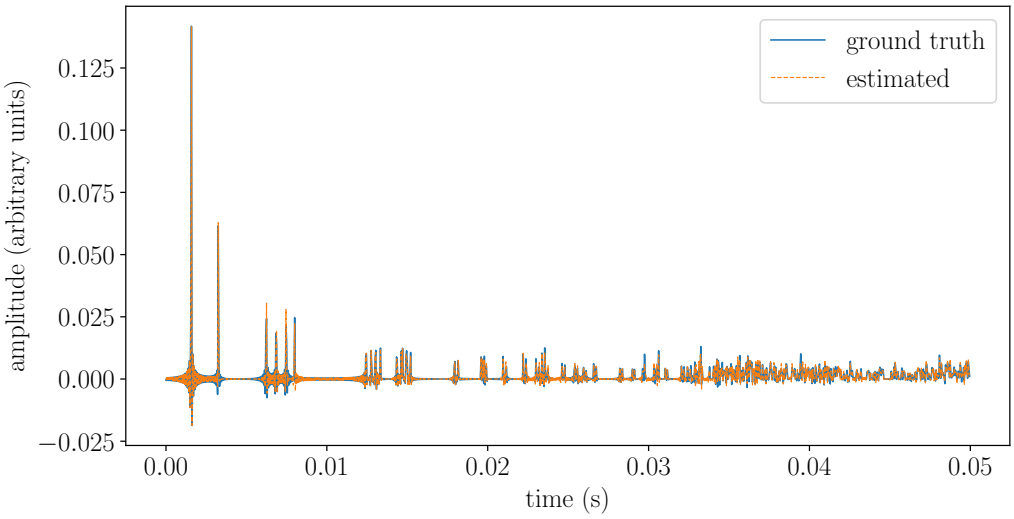


FIGURE 7.8: Example of RIR extrapolation inside the room of Fig. 7.5(A) (4.2 cm array radius, 24 kHz frequency of sampling).

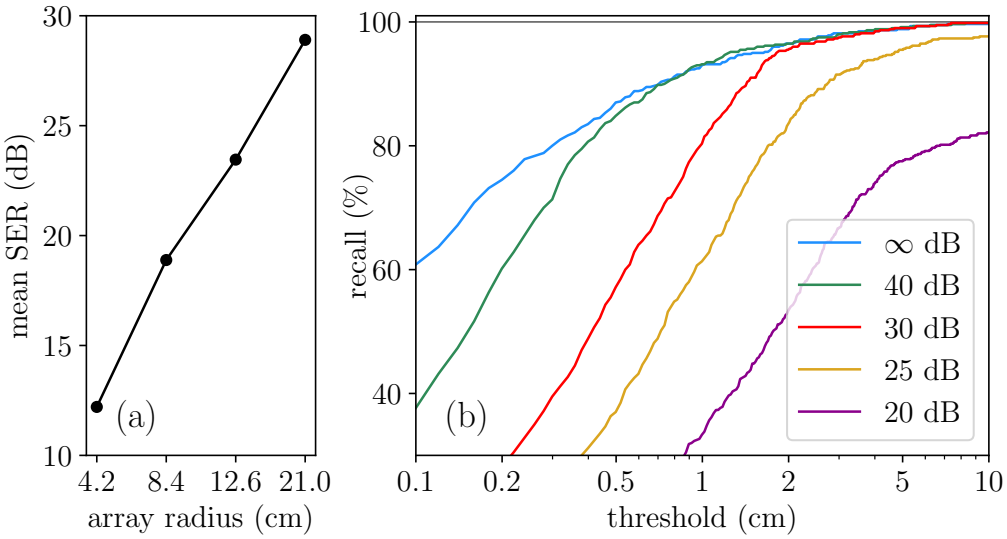


FIGURE 7.9: (a) Mean signal-to-error-ratio of RIR extrapolation for varying array radii at $f_s = 24$ kHz (b) Recall on room dimension recovery as a function of threshold for varying PSNRs for an array radius $r = 4.2$ cm and a frequency of sampling $f_s = 24$ kHz.

We now proceed with evaluating the ability of the method to extrapolate RIRs to arbitrary source-array placements in the same room. The results are shown in Fig. 7.9(a). We again observe a strong convergence of RIR extrapolation errors towards zero as the array size increases, bringing further support to the claim that the shoebox image-source method has been successfully reversed. Note that we did not observe such convergence as a function of the frequency of sampling. This is expected, since the RIR extrapolation task itself, as assessed by the proposed metric, becomes harder as the frequency of sampling increases. An example of RIR extrapolation result is presented in Fig. 7.8. As can be seen, the extrapolated RIR very closely matches the ground truth.

We finally study the impact of noise on room size estimation. The sampling frequency and array radius are respectively set to 24 kHz and 4.2 cm, and we proceed to varying the peak signal-to-noise

ratio (PSNR) of input signals using additive white Gaussian noise uncorrelated across channels. This is meant to coarsely emulate generic signal degradation but is not representative of typical degradations present in measured RIRs. Note also that the use of PSNR occludes the fact that in a RIR, the peaks of first-order echoes, all of which are necessary for room-size estimation, are typically much lower than the global peak. To give an idea of this, in each RIR from the 25-dB-PSNR test set, the PSNR of the weakest first order echo taken in isolation ranges from 1.5 dB to 19 dB (11.7 dB on average), which constitutes a significant degradation from a signal processing perspective. Fig. 7.9(b) presents the recall curves for the recovery of each individual room dimension L_i, for different recovery thresholds. As expected, the algorithm's performance deteriorates when the noise increases and a severe drop appears below 20 dB PSNR. Nevertheless, the algorithm still manages to recover 95.5% of all room dimensions with an error below 5 cm under a PSNR of 25 dB, suggesting a reasonable robustness of the approach.

7.2.2.5 Comparison to a Baseline

We now compare the accuracy of the proposed algorithm with the landmark Euclidean distance matrix (EDM)-based method introduced by Dokmanic et al. in 2013 [66], using the code provided by the authors[6]. This method takes as input a set of unlabelled times of arrival (TOAs) on multiple RIRs, and returns the 3D locations of first order image sources. Direct comparison on the synthetic dataset from the previous section turned out to be unfeasible. Indeed, the computational cost (in time and memory) of the EDM-based method explodes when the number of reflections is too large. Moreover, the method makes the strong assumption that only TOAs from image sources of orders lower than or equal to two are provided. Even when only considering these low-order sources, the number of considered combinations can become very high if the reflections are tightly clustered together due to the room's configuration, which frequently happens in our dataset. Finally, we observed that the computational cost of the method also drastically increases with the number of channels. In particular, applying the algorithm to the 32-channel spherical array used in other experiments, or even to arrays of more than 8 microphones, was not feasible without running out of time and memory (in [66], arrays of 5 microphones were used).

To reach a fair compromise between the computational feasibility and the performance of this baseline, we consider a non-spherical microphone array of 8 microphones consisting of two squares stacked on top of each other, the top square being rotated by an angle of $\pi/4$. The corresponding array diameter is 37.5 cm. This configuration is also meant to illustrate the applicability of our approach to a different array geometry and number of microphones. In order to avoid choosing a peak-picking technique to process the input of the EDM-based method, we place it in an oracle setting. Namely, *we provide it with the true times of arrival of all image sources up to order 2* that are in recording range (partial oracle labeling), rounded to the nearest discrete-time sample at 32 kHz. Note that working in discrete time is a fundamental limit of such approaches. We run the two algorithms on the same room configurations as before, only altering the array's geometry but retaining the same location for its center.

For each method, we compute the precision and recall for a 20 cm error threshold on the source and first-order image sources localization and labelling. While the proposed algorithm always returns exactly 6 first-order sources, the EDM-based method can wrongfully label second-order reflections as

[6] https://infoscience.epfl.ch/record/186657/

TABLE 7.5: Recall, precision and mean Euclidean errors (MEE) for correctly recovered first-order image sources (O_1) and MEE for the true source (O_0) using [66] or the proposed method.

	O_1 Rec.	O_1 Prec.	O_1 MEE	O_0 MEE
[66]	84.4%	59.7%	65.7± 41.3 mm	35.1± 26.0 mm
Ours	97.2%	97.2%	2.41± 5.71 mm	0.289± 0.584 mm

first-order reflections, causing a loss in precision. The results for these experiments are listed in Table 7.5. The localization errors obtained with the EDM-based method are over an order of magnitude larger than with the proposed approach. This highlights that, even using oracle TOA information, the considered task is far from trivial when considering fully randomized room parameters. The proposed algorithm obtains a mean Euclidean error below 3 mm, which is below $\frac{343}{2\times 32000} \approx 5.1$ mm, the theoretically lowest achievable radial error by any discrete-time method at this frequency of sampling, indicating that super-resolution is achieved.

The number of rooms for which all 6 first-order sources were retrieved without spurious second-order ones was 25.5% for the EDM-based method. Hence, the method could not be used to recover the full geometry of most of the rooms. In contrast, this ratio reached 95.5% of the rooms using the proposed method. For those rooms, the mean geometrical reconstruction errors obtained by it, keeping the same metrics as previous section, were respectively 0.34 ± 0.6 mm for the room dimensions, 0.61 ± 0.6 mm for the room translation and $0.016 \pm 0.05°$ mm for the room orientation. These results are in line with those obtained with the 32-element spherical microphone array of comparable radius and sampling frequency. This seems to indicate that when the array resolution is sufficient, adding microphones does not significantly improve the accuracy of correctly recovered sources. However, adding microphones does seem to reduce some of the geometrical ambiguities and hence to increase the number of correctly identified sources.

7.2.3 Conclusion on Shoebox Room Parameter Estimation

All-in-all, this section presented a four-stage algorithm that, given a discrete multichannel RIR simulated by the vanilla shoebox image-source method for a known microphone-array geometry, recovers all 18 geometrical and acoustical input parameters used to simulate it. Extensive numerical experiments with randomized input parameters revealed that near-exact recovery is consistently achieved by the method, for large enough array sizes and sampling rates. This constitutes, to our knowledge, the first evidence that the historical image-source method of Allen and Berkley [1] is *algorithmically reversible*, for a wide range of configurations.

Crucially, the proposed approach is currently not directly applicable to real measured RIRs. This is mainly because the inverse image-source localization method of Sec. 7.2.1 uses the vanilla image source method (6.26) as a forward model, which makes a number of simplifying assumptions that do not hold in reality. A path towards real-data applicability can nevertheless be envisioned. For this, the inverse method would need to be extended to take into account both angular and frequency dependencies of receiver, source, and wall responses, using the extended image-source models presented in Sec. 6.2.2.1 and 6.2.2.3. Even assuming the responses of the source and microphones are known, and using a plane-wave approximation for the angular dependencies of wall responses, the number of unknowns in the problem would then be significantly increased. Namely, one would need to additionally

estimate the source (and image sources) orientations, as well as a frequency-dependent impedance for each wall. Leveraging additional geometrical and physical constraints on these unknown or incorporating stochastic data-driven models are promising leads to make the corresponding inverse problem tractable. Another avenue for future research is to go beyond shoebox geometry. This would notably require tackling the combinatorial problem of image-source occlusions (see Sec. 6.2.2.2). It would also make the task of room orientation recovery more difficult, and would call for the development of more general point-cloud-to-geometry techniques.

7.3 General Conclusion on Physics-Driven Inverse Methods

This chapter presented three optimization-based inverse methods that aim to recover geometrical and acoustical parameters of interest from room impulse responses, given forward physical models of varying degrees of realism. An obvious limitation of the presented work is that the methods were only tested on simulated data that, although corrupted by noise, discrete-sampling effects and geometrical errors, closely followed the assumed forward models. Their direct application to real, measured data is still out of reach as of yet, and will require further research. Nonetheless, each approach does shed a bit of light on this thesis' central question: *Can One Hear the Walls of a Rooom?*

The first method in Sec. 7.1.1 indicates that extracting echoes from room impulse responses based on their expected times of arrival and looking at their short-term discrete Fourier magnitudes may provide sufficient, relevant, and robust acoustical data on acoustic reflectors, as long as the frequency responses of devices can be *a priori* compensated. The second method in Sec. 7.1.2 gives hope that as long as the directive responses of devices are precisely known, a full-time-domain optimization-based approach may jointly correct geometrical errors and estimate the reflectors' responses. Finally, the third method in Sec. 7.2 strongly suggests that the fundamental inverse problem of *"hearing the walls"* from the early part of a multichannel RIR is not only well-posed but also effectively solvable algorithmically, in the limit of idealized devices (omnidirectional), idealized reflective surfaces (near-rigid) and idealized room geometry (empty shoebox). This latter fact was not *a priori* obvious, and a negative result would have required reframing our central question to make it solvable.

Taking a step back, we seem to have reached a fundamental limit that is common to any purely physics-based inverse method, namely, that the complexity of the real world results in an inevitable mismatch between the theoretical forward model we want to invert and observed data. Such mismatch, by definition, cannot be derived and modeled from first principles. This leaves as the only possible route the construction of stochastic models from the *data themselves*. In the next chapter, we will continue exploring our central question but coming from the opposite direction offered by this route.

8 Virtually Supervised Learning

Associated publications: [S25, S26, S28, S31, S34]

The previous chapter highlighted some limitations of physics-based approaches, namely, they involve the resolution of difficult, non-convex inverse problems and they are not readily applicable to real data due to model mismatch. An attractive alternative route is offered by the paradigm of *supervised learning*, and in particular *regression*. In a nutshell, given a *training set* of input-output pairs $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_{\text{train}}}$, regression consists in searching for a function $f_{\boldsymbol{\theta}^*}$ inside a family of parameterized functions $\mathcal{F} = \{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$ such that $f_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i) \approx \boldsymbol{y}_i$ in a suitable sense. Ideally, the chosen family $\mathcal{F}$ and the associated search method should be such that one also has $f_{\boldsymbol{\theta}^*}(\tilde{\boldsymbol{x}}) \approx \tilde{\boldsymbol{y}}$ for a *test* pair $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ that is outside, but reasonably close to, the training set distribution. In the context of this thesis, this paradigm may constitute a promising route to map an acoustic measurement $\boldsymbol{x}$ (see Sec. 6.3) to acoustical and geometrical parameters of interest $\boldsymbol{y}$ (see Sec. 6.4).

Artificial and in particular *deep* neural networks (DNN) form an endlessly versatile source of parameterized families of functions $\mathcal{F}$. Along the past decade, performing supervised regression by training a DNN to minimize a loss function on a sufficiently large, diverse and representative training set via variants of stochastic gradient descent has proven extraordinarily effective, with impressive generalization capabilities in practice. Of course, *sufficiently large, diverse and representative* constitutes the crux of the problem. In our context, gathering enough real training data would involve performing a large number of acoustic measurements fully annotated by geometrical and acoustical parameters in many different environments. Such data, already very scarce for a *single room* due to their prohibitive acquisition time and cost, are non-existent for hundreds of rooms. This leaves little hope that pure *real-data-driven* approaches can successfully be applied to our problem.

Given this premise, we will explore in this chapter an approach which we coined *virtually-supervised learning* in some of our early works[1] [S8, S9]. The idea is to leverage an acoustic simulator[2] to build a large annotated training dataset, use this set to train a regression model, and use the model for inference on real data. Since, by definition, acoustic simulators perform a *forward mapping* from acoustic parameters $\boldsymbol{y}$ to acoustic measurements $\boldsymbol{x}$, they form a potentially unlimited source of training data to learn the *inverse mapping* from. The interesting research question is then: *How to simulate sufficiently large, diverse and representative training datasets to supervise models that generalize well to real acoustic data?*

Sec. 8.1 will examine the main trade-offs that one must consider to build training sets for virtually-supervised acoustic learning, and present two of our contributions to this effort. Sec. 8.2 revisits the task of estimating absorption coefficients from RIR that was explored in Sec. 7.1, but from the angle of virtually-supervised learning. Finally, Sec. 8.3 tackles the problem of *blindly* estimating geometrical and acoustical parameters from noisy speech signals.

[1] Our works in [S8, S9], published in 2017, did not make use of deep learning but of a probabilistic locally-linear regression model. The focus was on sound source localization, although [S9] jointly performed mean absorption estimation. We will not cover these articles here, although the subsequent works presented in this chapter are greatly indebted to and inspired by them. They are briefly summarized in our research Snippet 11 on sound-source localization.

[2] We note that an interesting intermediate route, as explored in *e.g.*, [30, 28, 34, 33, 29], is to leverage datasets that *combine* real and simulated data. While these studies show that leveraging a few hundred annotated real RIRs for training can improve the generalizability of room parameter estimators to real data, such annotations remain costly to acquire even in small amount. This chapter focuses on purely simulated training sets instead.

8.1 Training Data Generation

8.1.1 Simulation Trade-Offs

Simulating a sufficiently large, diverse and representative training dataset for the purpose of virtually-supervised learning and real-data generalization requires navigating two important trade-offs. The first one is the targeted level of *realism*, which typically grows with the *computational cost* of the chosen simulator. The second one is the targeted *representativity* of the generated set, which may require an explosion of its *size*, and in turn, of the memory and computational demand. We examine these two trade-offs in the next subsections.

8.1.1.1 Realism/Computation Trade-Off

When simulating RIRs, more realism typically implies higher, sometimes prohibitive computational costs. Existing room acoustic simulators can be divided into three categories[3]. The first category, referred to as *wave-based* solvers, directly solve the wave equation (6.1)-(6.2) by discretizing space, time and/or frequency. These notably include finite-element methods, boundary-element methods or finite-difference time-domain methods. While these can in principle simulate any boundary conditions to arbitrary precision, their computational demand is inversely proportional to the space-discretization step used, which determines the smallest attainable wavelengths. Because of this, these approaches are impractical above ≈1 kHz, in particular for large volumes. A second category, belonging to so-called *geometric acoustic* approaches, includes variants of the image source method (ISM, see Sec. 6.2). While the ISM enables fast simulation of RIRs at early times, the required number of image sources grows cubically in the reverberation time, and the model does not handle advanced boundary effects such as *scattering*. The last category includes *energy-based methods*, also known as *ray-tracing* or *particle filtering*. They also belong to geometric acoustic but leverage Monte-Carlo sampling. The sound waves emitted by sources are replaced by a large number of sound rays that are tracked and stochastically altered as they interact with boundaries. These approaches are asymptotically accurate at high frequencies (say above ≈1 kHz), and can in principle model arbitrary acoustic conditions, including surface scattering. However, their precision, in particular on early-time and near-field effects, is tied to the number of rays employed, directly impacting computational demand.

The complementary strengths and weaknesses of existing approaches led to the development of hybrid solvers. Let us review two examples of research works that use such hybrid solvers in the context of virtually-supervised acoustic learning, and nicely illustrate the necessity of a realism/computation trade-off. The first one is the GWA dataset [68], which features ≈2 million highly-realistic single-channel RIRs, simulated by combining a wave-based solver at low frequencies and a geometric-acoustic solver at high-frequencies. Although highly-optimized, this effort required a combined ≈1,300 hours of CPU and GPU usage. The authors showed that models trained on this dataset offered better generalization to real data than identical models trained with a number of less realistic but faster solvers, on a variety of single-channel speech processing tasks. One limitation of GWA is that it assumes omnidirectional devices. Extending it to more realistic devices or to a microphone array would require regenerating the entire dataset anew. The second example is the VAST dataset, presented in our early work [S8]. It includes ≈100k RIRs generated by the hybrid simulator

[3] We refer the reader to [67] for an extensive review of existing room acoustic simulation concepts, which is out of the scope of this thesis.

`Roomsim` of Schimmel et al. [41], which combines the ISM with a directive binaural receiver and an efficient ray-tracing technique called *diffuse-rain* to model scattering effects. This effort required ≈700 CPU hours. We showed that a sound-source localization model trained on this dataset offered better generalization to real data than the same model trained on an anechoic binaural dataset. Here again, a limitation of VAST is that it is restricted to a specific binaural receiver and an omnidirectional source.

The models presented in this chapter will either be trained using the aforementioned hybrid simulator `Roomsim` [41], or some of the ISM extensions discussed in Chap. 6, as implemented in `pyroomacoustics` [42], the latter being faster due to the absence of ray tracing. These options proved to form an adequate compromise for the purpose of this research, fast enough to enable reactive experimentation with training sets containing tens of thousands of RIRs, while offering promising real-data generalizability when combined with appropriate representativity trade-offs, as discussed in the next section. Nonetheless, as we shall see, striking an ideal level of realism is ultimately tied to the task and application at hand, and should be studied on a case-by-case basis.

8.1.1.2 Representativity/Size Trade-Off

While deep-learning-based supervised regression models are known for their remarkable *interpolation* capability within the training-set distribution, they are also known for their limited *extrapolation* capability. A large diversity in the training set is hence desirable for the learned model to generalize well to many different situations. However, more diversity also implies more data, to obtain a set that is *representative* of the targeted use case. Indeed, for a fixed sampling density of the parameterized observation space, the number of required training samples grows exponentially in the number of parameters, a classical manifestation of the *curse of dimensionality*. We list below three identified axes along which a diversity/representativity cursor must be set, and discuss the choices made in the remaining work of this chapter.

Geometrical sampling. The volume of real-world rooms vary widely, from small water closets ($\sim$6 m^3) to typical office rooms (20-60 m^3) to large entrance halls ($>$600 m^3). The volume has a direct impact on RIRs, affecting their duration (see Sabine and Eyring's formula in Sec. 6.4) and their *echo density* (see, *e.g.*, [63]). Indoor environments also come in a variety of shapes, from coupled or L-shaped rooms to auditoriums and cathedrals, which impact reverberation via scattering effects and the distribution and visibility of image sources (see Sec. 6.2.2.2). In this chapter as in the rest of this thesis, our focus will be on rooms that are typical of office, school, restaurant or accommodation buildings. Accordingly, training sets will be designed by sampling shoebox room dimensions uniformly at random with width, length and height inside the range $[2, 10] \times [2, 10] \times [2, 5]$ in meters. Once a room size is picked, we typically place receivers and sources uniformly at random in the room, while preserving a pairwise distance of 30-50 cm between sources, receivers and walls to avoid undesired edge effects. When directional or multichannel devices are considered, they are also randomly rotated on themselves.

Acoustical sampling. The absorption profiles of materials encountered in rooms also vary widely, from the hard concrete walls of a hangar, to linoleum floors, to carpets and curtains. To account for this diversity, we will propose in Sec. 8.1.2 a sampling strategy designed to be representative of the acoustics encountered in the typical buildings listed above. This will purposefully exclude

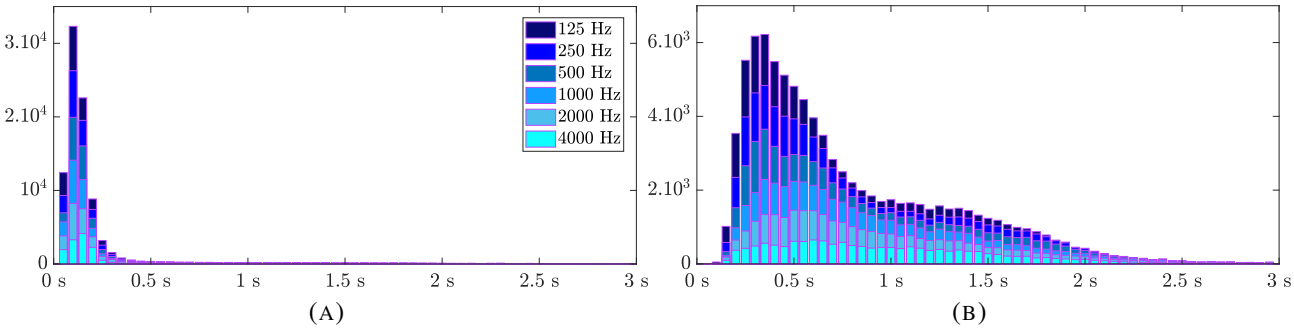


FIGURE 8.1: Histograms of $\mathrm{RT}_{60}(b)$ reverberation times obtained from RIRs simulated by `Roomsim` [41] based on 15,000 rooms drawn from the geometrical sampling described in Sec. 8.1.1.2, using either a *uniform* sampling of absorption coefficients (A) or the proposed *reflectivity-biased* sampling scheme (B), described in Sec. 8.1.2.

materials that are only encountered in highly specialized rooms such as (semi-)anechoic chambers or studios. With the exception of Sec. 8.3.3 which does not use the ray-tracing simulator, we will also implicitly account for the acoustic effects of smaller motives or objects, including furniture, by assigning *scattering coefficients* to each surface, drawn uniformly at random in $[0, 0.3]$ for octave bands at and below 500 Hz, and in $[0.2, 1]$ for octave bands at and above 1 kHz. This choice is guided by scattering profiles measured in real furnished rooms, as reported in [69].

Source and receiver sampling. As explained in Sec. 6.2.2.3 and 6.3, the directive responses of sources and receivers has an important and often neglected impact on acoustic measurements, which may strongly affect room parameter estimation methods, as already observed in Chap. 7. But to what extent can the representativity of devices in the training set impact virtually-supervised learning ? To help answering this question, we will present in Sec. 8.1.3 an extension of the `pyroomacoustics` ISM-based simulator that will enable the simulation of various real sources and microphones. We will then precisely study the impact of incorporating these directivities at train time for virtually-supervised blind room parameter estimation in Sec. 8.3.3.

8.1.2 Surface Absorption Profiles

Associated publication: [S26]

When designing a training set with the aim of faithfully representing the diversity of commonly encountered room acoustics, how should one sample the absorption coefficients $\alpha_i[b]$ of the 6 walls ($i = 1 \dots 6$) in 6 octave bands ($b = 0.125, \dots, 4$ kHz) ? The most obvious and straightforward choice would be to draw each of the 36 coefficients independently and uniformly at random in $[0, 1]$. To our knowledge, this *uniform* sampling approach remains the most widely-used one by current virtually-supervised methods, *e.g.*, [32]. While one may intuitively expect this scheme to maximize acoustic diversity, we found that, on the contrary, it leads to a significant bias in resulting reverberation times. As can be seen in the histogram of Fig. 8.1(A), the resulting $\mathrm{RT}_{60}(b)$ distribution over 15,000 simulated RIRs is tightly clustered around 150 ms, which is a highly unusual value, typical of semi-anechoic chambers. This is because using this technique, drawing four or more reflective absorption profiles within a same room (*e.g.*, $\alpha_i(b) < 0.12$ for all b) is very unlikely. Yet, highly reflective profiles are frequently encountered in real buildings. These are characteristic of hard surfaces made of, *e.g.*,

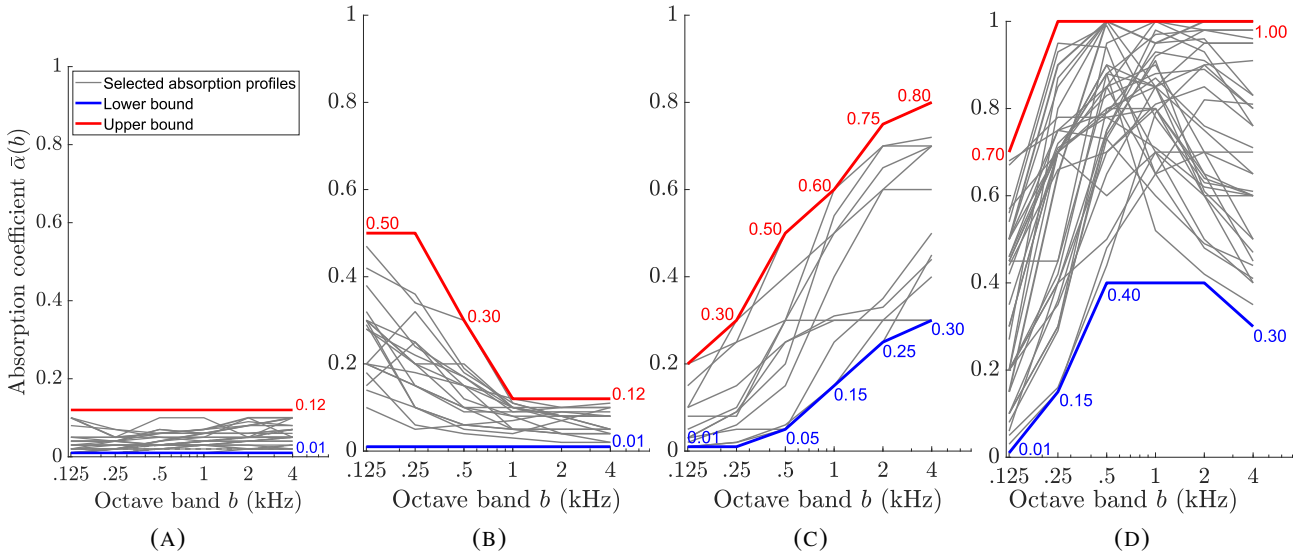


FIGURE 8.2: Absorption profiles of 92 commonly encountered reflective, wall, floor and ceiling materials together lower (blue) and upper (red) bounds. (A) 26 reflective profiles, (B) 19 wall profiles, (C) 12 floor profiles, (D) 35 ceiling profiles. The full lists of materials and profiles used for these figures are available at `https://members.loria.fr/ADeleforge/files/jasa2021_supplementary_material.zip`.

concrete, bricks or tiles. The absorption profiles of 26 such materials are plotted in Fig. 8.2(A). As can be seen, they are all roughly frequency-independent with absorption coefficients below 0.12.

Based on this observation, we designed the following *Reflectivity-Biased* (RB) sampling strategy, that is extensively used in Chap. 7 and 8 of this thesis:

1. for each surface type (wall, floor, ceiling), toss a coin;
2. on heads, draw *reflective* frequency-independent absorption profiles uniformly at random in $[0.01, 0.12]$ for all surfaces of this type;
3. on tails, draw *non-reflective* frequency-dependent absorption profiles uniformly at random within ranges depending on the surface type, as defined by the blue and red curves in Fig. 8.2(B,C,D).

As can be seen in Fig. 8.1(B), this scheme results in a more diverse and representative distribution of reverberation times, spread over the 150 ms-2.5 s range. Note that with this scheme, walls are either all reflective or all non-reflective, with distinct profiles. To produce even more diversity while preserving a similar $\mathrm{RT}_{60}(b)$ distribution, a variant of these schemes will be used in Sec. 8.2.2 and 8.3, where each surface is independently and randomly assigned a type among reflective, wall, floor and ceiling, with respective probabilities $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{12}$ and $\frac{1}{12}$.

8.1.3 Device Responses

Associated publication: [S31, S34]

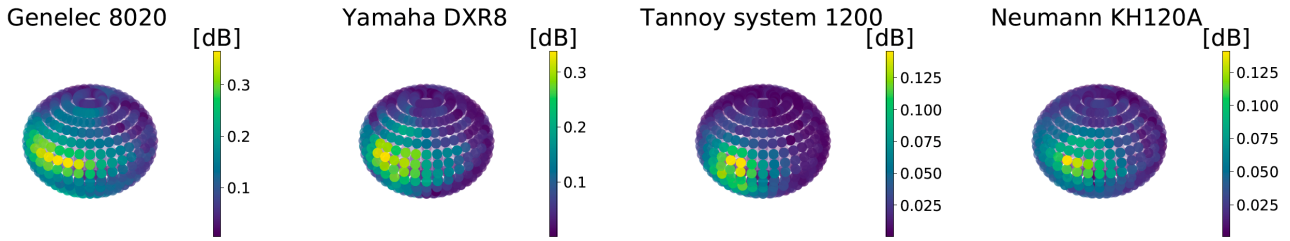


FIGURE 8.3: Magnitude responses of 4 loudspeakers from the DirPat dataset [49] at 2 kHz, over the rectangular grid $\mathcal{G}$ of spherical coordinates.

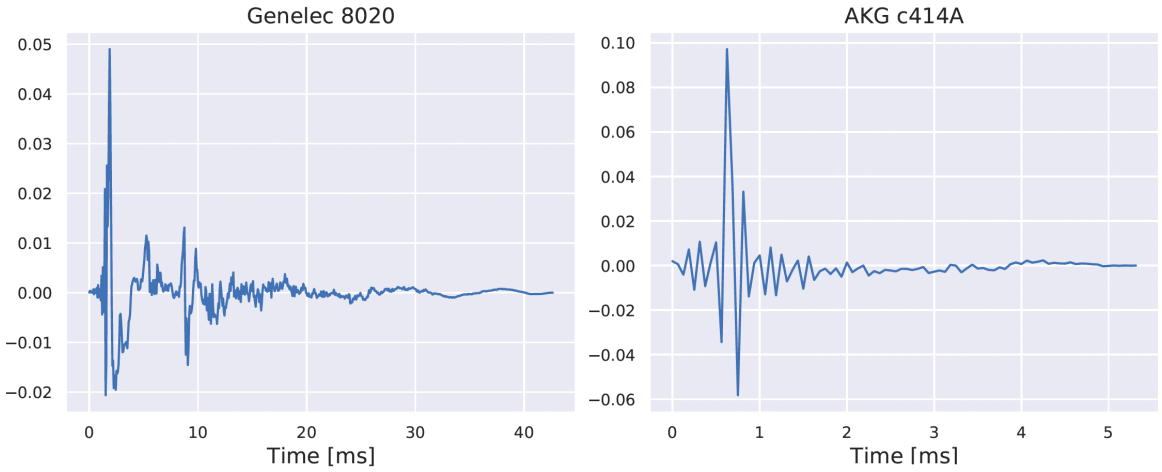


FIGURE 8.4: Finite impulse response filters taken at a random point on the sphere for the loudspeaker Genelec8020 and the microphone AKG C414A in the DirPAT dataset [49].

Our second contribution to the realism and diversity of simulated acoustic training sets concerns the directive response of devices. During his PhD thesis, Prerak Srivastava contributed to the open-source `pyroomacoustics` simulator [42] by implementing the directive, wall-impulse-response extension of the image-source method presented in Sec. 6.3, which we recall here:

$$x_{\text{WIR-dir},m}[n] = \sum_{k=0}^{K} \frac{1}{4\pi \|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\|} \Big(g^{\text{src}}(\theta_{k,m}^{\text{out}}, \phi_{k,m}^{\text{out}}, \cdot) * g_m^{\text{mic}}(\theta_{k,m}^{\text{in}}, \phi_{k,m}^{\text{in}}, \cdot) * a_k * \text{sinc}(\cdot - \|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\|/c) \Big)(n/f_s). \tag{6.27}$$

While the toolbox already enabled analytic frequency-independent patterns for g^{mic} and g^{src} such as cardiod or figure-of-eight, our extension enables the use of *measured* directivity patterns from the DirPat dataset [49]. DirPat includes measurements for a variety of loudspeakers and guitar amplifiers, a Brüel & Kjær head and torso mouth simulator, an AKG C414 microphone in four settings (Omni, Cardiod, Supercardiod, F-8), the em32 Eigenmike®, and two other microphones. For each source and receiver, the measurements are available as time-domain finite impulse responses on a discrete spherical grid $\mathcal{G}$ with 30 regularly-spaced azimuth and 16 or 18 regularly-spaced elevations, $\{g[\theta, \phi, n]\}_{(\theta,\phi)\in\mathcal{G}}$. To obtain a better coverage of incidence and exit angles in simulations, we used spherical harmonic interpolation to map the discrete Fourier transform of each impulse response at each frequency f to a denser and more uniform Fibonacci grid $\mathcal{G}'$ with $1,000$ points, yielding $\{\hat{g}[\theta, \phi, n]\}_{(\theta,\phi)\in\mathcal{G}'}$. Interpolation is implemented using Voronoi-cell-based weighted least squares as

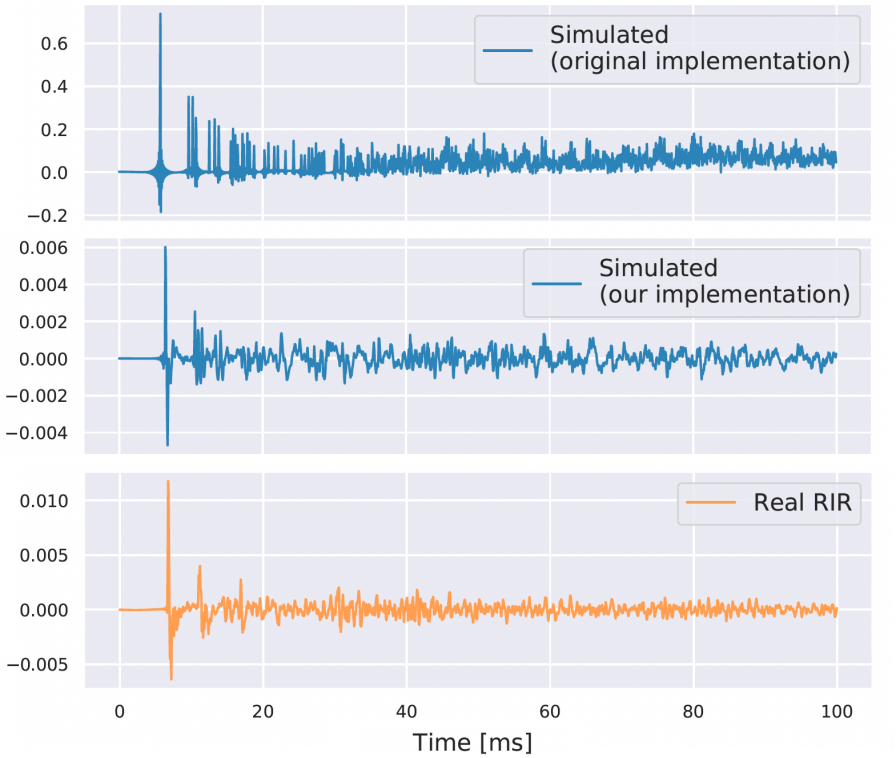


FIGURE 8.5: Qualitative comparison between: (top) a RIR generated using the vanilla-ISM model (6.26), (middle) our extension with wall impulse responses and DirPat directivities from [49] following (6.27) and (bottom) a real measured RIR from the dEchorate dataset (Sec. 9.1). The top two RIRs were simulated with the `pyroomacoustics` package [42], manually setting geometrical and acoustical parameters to roughly match those of the real one.

described in [70, Chap. 4]. At simulation time, the responses for image source k with continuous exit and incidence angles are then obtained by picking the nearest neighbor on $\mathcal{G}'$.

The magnitude responses over $\mathcal{G}$ of 4 loudspeakers from DirPat at 2 kHz are shown in Fig. 8.3. Their responses are clearly far from omnidirectional. Examples of time-domain responses from a loudspeaker and a microphone from DirPat are shown in Fig. 8.4. The loudspeaker response is significantly spread over ≈20 ms, which will inevitably cause overlap between the echoes of a RIR measured with this loudspeaker. A qualitative comparison between a RIR simulated with the vanilla image-source method (6.26), our DirPat-based extension following (6.27), and a real RIR is shown in Fig. 8.5. The spreading and distribution of peaks is noticeably closer to reality using the directive-device model.

8.2 Absorption Estimation from a Room Impulse Response

8.2.1 Mean Absorption

Associated publication: [S26]

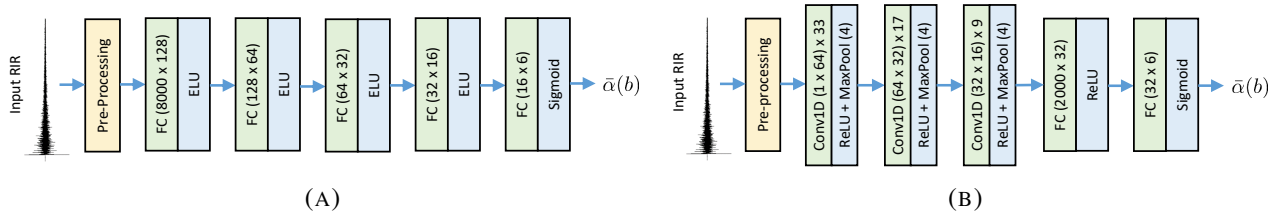


FIGURE 8.6: Layer design, non-linearities and dimensions of the neural network architectures used to learn the RIR$\rightarrow \bar{\alpha}$ mapping. (A): multilayer perceptron (MLP), (B): convolutional neural network (CNN). FC stands for *fully connected*, Conv1D for *1D convolutional*, ELU for *exponential-linear unit*, ReLU for *rectified linear unit*.

In this section, we train neural network models via virtually-supervised learning to map a single-channel room impulse response to its corresponding *area-weighted mean absorption coefficients* $\bar{\alpha}(b) = (\sum_i \alpha_i(b)S_i)/S$ in 6 octave bands, as defined in Sec. 6.4. This quantity is treated here as an analytical parameter that globally summarizes the acoustic properties of all surfaces in the room. On the one hand, it is a less informative quantity than the individual absorption profiles $\alpha_i(b)$ that were targeted in Sec. 7.1. On the other hand, we consider here a more commonly encountered scenario where only a single RIR measurement with no extra geometrical information is available. In such scenario, recovering individual profiles seems out of reach, due to obvious geometrical and permutation ambiguities. Meanwhile, $\bar{\alpha}$, calculated using the classical Sabine or Eyring's formula (see Sec. 6.4), is a commonly used acoustic parameter by field practitioners. For example, if the absorption profiles of the floor and walls can be retrieved from an existing database of materials (see Fig. 8.2), and if the boundary surface areas S_i are approximately known, a coarse estimation of the ceiling's profile and the impact of a proposed acoustic solution can be inferred based on $\bar{\alpha}$'s definition.

There are however known limitations to these classical formulas. First, they require to know the room's geometry by means of S and V. Second, they require an accurate estimation of the reverberation time, which may not be available when the measured RIR features an insufficient or non-linear decay of its Schroeder curve [50]. Last, as discussed in Sec. 6.4, their validity strongly hinges on the assumption of a *diffuse and isotropic sound field*, which does not hold in many practical situations, including "shoebox" rooms. For these reasons, a direct and generally usable RIR$\rightarrow \bar{\alpha}$ mapping could be of practical interest.

8.2.1.1 Neural Network Models and Training

A first question is how to represent the RIR at the neural network's input. Ideally, one seeks a representation that preserves or enhance features that are relevant while removing unnecessary or redundant ones. In learning-based audio signal processing applications, phase-less time-frequency representations such as magnitude spectrograms or Mel-frequency cepstral coefficients are commonly used. Since frequency-dependent values are sought, such representations seem attractive at first glance. However, discarding the phase runs into the risk of removing fine-grain time and amplitude information contained in early echoes due to overlap and interference, as experienced and discussed in Sec. 7.1.1. Alternatively, one could consider invertible complex time-frequency representations such as the short-term Fourier transform (STFT). Our preliminary experiments in that direction were however not conclusive, possibly due to the difficulty of handling complex values in neural networks,

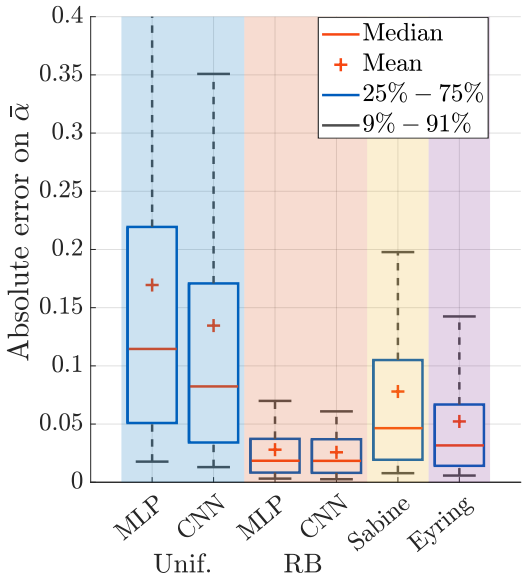


FIGURE 8.7: Box plots of the absolute errors on mean absorption coefficients $\bar{\alpha}(b)$ obtained using Sabine and Eyring formulas and the proposed MLP and CNN models virtually trained using either the *uniform* (Unif.) or the *reflectivity-biased* (RB) absorption sampling schemes (Sec. 8.1.2), on a simulated test set of 500 RIRs.

or because any choice of STFT parameters implies a non-obvious compromise between time and frequency resolution. Consequently, we choose to let the networks learn their own internal representation of time-domain RIRs, in an end-to-end fashion.

As the only pre-processing, input RIRs are resampled to 16 kHz, cropped to 500 ms, and normalized to have a maximal value of one, yielding 8000-dimensional real vectors. The normalization is done to facilitate learning and to prevent models from relying on the RIR's absolute amplitude, which is often inaccessible in practical applications due to unknown source and microphone gains. Random white Gaussian noise with a signal-to-noise ratio (SNR) of 30 dB is also added to every simulated RIR throughout this study, as a regularizer. This is to make learned models more robust and prevent them from relying on vanishingly small values, which would be inaccessible in practical applications.

We consider two common neural network architectures with simple designs and comparable number of parameters and depth, namely, a multi-layer perceptron (MLP) and a convolutional neural network (CNN). Their respective architectures are depicted in Fig. 8.6(A) and (B). Each network outputs an estimated vector of mean absorption coefficients $\widetilde{\boldsymbol{\alpha}} \in [0,1]^6$, which is fitted to the target $\bar{\boldsymbol{\alpha}}$ using a mean-squared error loss-function. Networks are trained on a set of 15,000 RIRs generated using the geometrical and acoustical sampling described in Sec. 8.1.1.2 and the `Roomsim` simulator [41], with batches of size 1000 and the ADAM optimizer [54] with a learning rate of 0.001. Training is stopped before 400 epochs using early stopping on a separate evaluation set of 5,000 RIRs generated the same way. Note that only omnidirectional devices are considered here, studying the impact of directivity being left to Sec. 8.3.3.

8.2.1.2 Results on Mean Absorption Coefficient Estimation

We now evaluate the MLP and CNN models virtually supervised by either the *uniform* (Unif.) or the *reflectivity-biased* (RB) absorption sampling schemes defined in Sec. 8.1.2. They are compared to the

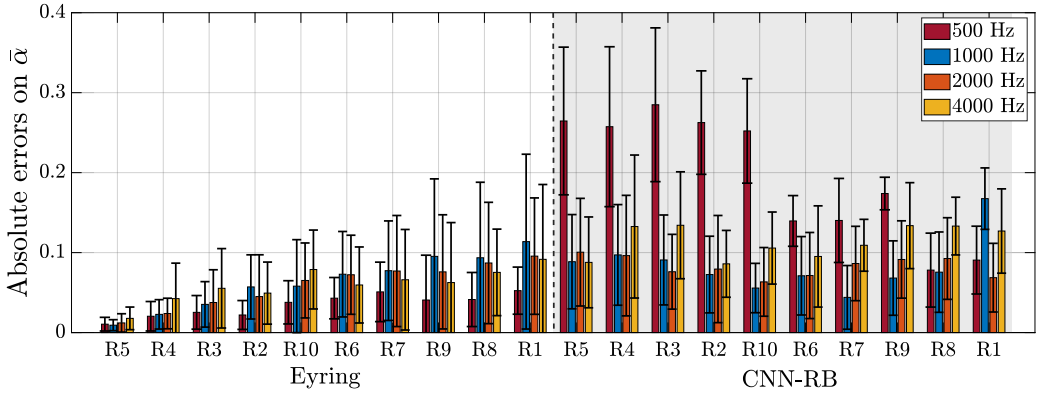


FIGURE 8.8: Mean and standard deviation of $\bar{\alpha}(b)$ estimation errors over measured RIRs from the dEchorate dataset [S24] (Sec. 9.1), in 4 octave bands, using Eyring's formula or the CNN with reflectivity-biased (RB) training. Only the RIRs whose Schroeder curves are in $\mathcal{A}$, for which the reverberation times could be computed, are included.

Sabine and Eyring formulas (Sec. 6.4) to which the true room volume V, true total surface area S, and Schroeder $\mathrm{RT}_{60}(b)$ [50] are provided.

We first use a simulated test set of 500 RIRs meant to emulate acoustics and geometries plausibly encountered in real buildings. Five representative geometries are selected with the following (L_x, L_y, L_z) dimensions in meters: $(4, 5, 3)$, $(10, 2, 3)$, $(10, 5, 3)$, $(5, 8, 2.5)$, $(10, 10, 5)$. The absorption profiles of the walls, floor and ceiling are drawn uniformly at random from the database displayed in Fig. 8.2. The obtained absolute errors on mean absorption coefficients by the different methods are shown in the form of box plots in Fig. 8.7. As can be seen, networks trained on the naive Unif. training set do not succeed in outperforming classical approaches. However, mean estimation errors twice smaller than Eyring's method and with much less variance are obtained using the networks trained on the RB set. As expected, Sabine's estimates show to be slightly less accurate than Eyring's overall. The MLP and CNN models appear to perform very comparably on this set, with a slight edge for the CNN model.

To evaluate the generalizability of our virtually-supervised models to real measured RIRs, we use the 10 non-furnished variable-acoustic room configurations from the dEchorate dataset [S24], that will be described in details in Sec. 9.1 of this thesis, representing 900 test RIRs. In this experiment, the octave bands centered at 125 and 250 Hz will not be considered because the measured RIRs did not exhibit sufficient power in those bands for reliable $\mathrm{RT}_{60}(b)$ estimation. A major difficulty in evaluating the considered models on real *in situ* measures is the unavailability of ground truth for the mean absorption coefficients, which would require to know the true absorption profiles of every material in the room. While some of them could be inferred from manufacturer's data, only coarse values of $\bar{\alpha}(b)$ would be obtained in this way. To overcome this difficulty while ensuring that a single, stable and reliable mean absorption profile is used as a reference for each room, we propose a technique based on the aggregation of multiple RIR measurements.

For each room configuration, the Schroeder curves of the 90 measured RIRs in 4 octave bands are traced. Then, the Schroeder curves are visually inspected and separated into two sets. Set $\mathcal{A}$ contains Schroeder curves featuring a sufficient linear log-energy decay from -5 dB to -15 dB at least. Set $\mathcal{B}$ contains all the other curves. In practice, 49% of the 3600 Schroeder curves were discarded to the set $\mathcal{B}$ in this way. These mostly corresponded to challenging measurement situations contained in the dEchorate dataset, such as a receiver near a surface, or a loudspeaker facing towards a surface

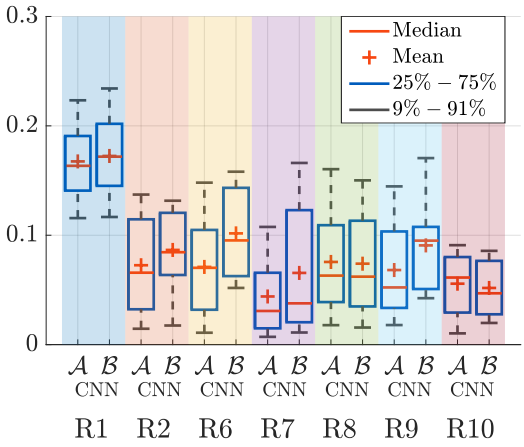


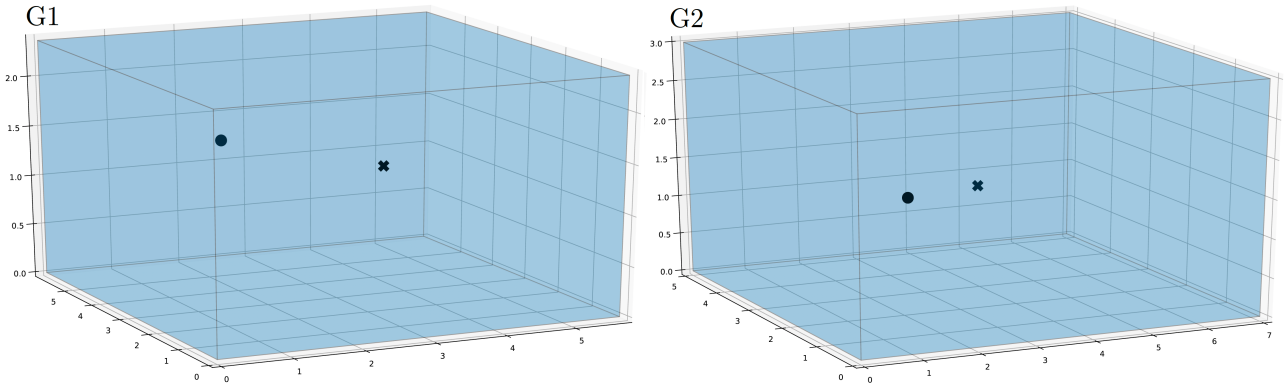
FIGURE 8.9: Box plots of the absolute errors on the mean absorption coefficients $\bar{\alpha}(1\text{ kHz})$ obtained with the CNN with reflectivity-biased training. Errors obtained on RIRs whose Schroeder curve at 1 kHz is in dataset $\mathcal{A}$ vs. $\mathcal{B}$ are compared.

and away from receivers. These situations yield to the known phenomenon of *double decay*, which prevents reliable RT_{60} estimation. Then, for each room configuration and each octave band b, the reference mean absorption coefficient $\bar{\alpha}_{\text{ref}}(b)$ is taken to be the median value of Eyring's model based on the $\text{RT}_{60}(b)$ computed from Schroeder curves in $\mathcal{A}$ only, as extrapolated from the -5 dB to -15 dB range, using the known room's volume and total surface area. This median value $\bar{\alpha}_{\text{ref}}(b)$ is taken over at least 5 and on average 47 estimates, yielding reliable and robust ground truth values. We observed that a diversity of mean absorption coefficients between 0.12 and 0.52 was represented over the 10 rooms and 4 octave bands, which matches quite well the range of values displayed in Fig. 8.2.

On real RIRs, the MLP performed significantly worse than the CNN model, yielding errors up to twice as large. We hence focus on the CNN, trained with the RB sampling scheme. Crucially, the Eyring model could not be applied to the Schroeder curves in dataset $\mathcal{B}$, since their reverberation time cannot be meaningfully computed. Results obtained using Eyring's formula over the RIRs whose Schroeder curves are all inside dataset $\mathcal{A}$ are given in Fig. 8.8. Rooms are sorted left-to-right from the most reverberant one to the least reverberant one. Errors increase as the reverberation time decreases, as expected. Nevertheless, they remain reasonably low (below 0.1) under all configurations, despite measurements being taken from many different source-receiver placements and orientations in the room. This suggests that the diffuse-isotropic sound field assumption underlying Eyring's formula is reasonable for the RIRs corresponding to $\mathcal{A}$. The right part of the figure shows the corresponding results for the RB-trained CNN. Except for the two least-reverberant rooms, obtained errors are two-to-three times larger at 500 Hz than in other octave bands. This could be due to a wave phenomenon that could not be learned by the neural network trained on Roomsim. At and above 1 kHz, the CNN performs at a steady level across all room, with reasonable errors around 0.1. While encouraging in absolute, this contrasts with Eyring's formula which performed better in highly reverberant rooms, although one should keep in mind that this formula was used as a reference.

Finally, Fig. 8.9 compares errors obtained by the RB-trained CNN at 1 kHz, depending on whether the corresponding Schroeder curve was placed in dataset $\mathcal{A}$ or $\mathcal{B}$[4]. Interestingly, we do not observe a significant difference between the two datasets, which means that the CNN is unaffected by the non-linear or insufficient log-energy decays of Schroeder curves. This suggests that the network learned

[4]Rooms R3, R4, and R5 are omitted here because an insufficient number of curves were placed in $\mathcal{B}$ for these rooms.



(A) (B)

FIGURE 8.10: Room geometries together with their source (•) and microphone (×) positions used for virtually-supervised absorption profile estimation. The room dimensions are $5.70 \times 5.97 \times 2.36$ meters for G1 and $7 \times 5 \times 3$ meters for G2.

to rely on more elaborate and more robust features than those used by classical reverberation-theory formulas, which could not be applied to this dataset.

8.2.2 Individual Absorption Profiles in Fixed Geometry

Associated publication: [S25]

In this section, we revisit the task explored in Sec. 7.1, namely, estimating the individual absorption profiles of the 6 surfaces in a shoebox room under (approximate) knowledge of the geometrical parameters, but this time from the angle of virtually-supervised learning. As in Sec. 8.2.1, we consider here the challenging but practical setting where a single RIR is available for the room. This task is likely impossible without the knowledge of geometrical parameters, at least because of the obvious permutation ambiguity between the different reflective surfaces in the room. Indeed, a single channel RIR contains distance information in its echoes' times of arrival, but no directional information. An attempt to alleviate this ambiguity was made in the learning-based approach of [32], where the absorption coefficients of the 6 walls in a fixed octave band were predicted in *increasing order*. Besides the sheer difficulty of this task[5], the authors acknowledged the limited usefulness of arbitrarily ordered coefficients, and outlined an approach to link them to their respective surfaces, using geometrical knowledge.

In this section, the geometry will be implicitly informed to our regression model by training it on a simulated RIR dataset with a *fixed* room-device geometry, up to slight random variations to enforce robustness. This presents the advantage of intrinsically resolving the profile-to-surface assignment, the coefficients being always predicted in the same order, but the disadvantage of being geometry-specific, requiring a new training for a different geometry.

[5]The authors of [32] report errors only 30% to 60% smaller than chance. This is under a *uniform* sampling of absorption coefficients at both train and test times, which we showed to be unrepresentative of typical room acoustics in Sec. 8.1.1.2. Errors would likely get closer to chance under a more representative reflectivity-biased sampling, due to the skewness towards zero of the corresponding distribution.

	G1→G1	G2→G2	Rand→Rand	Mean
Mean absolute errors	0.058	0.068	0.13	0.14

TABLE 8.1: Mean absolute errors on absorption coefficients obtained by training on G1 and testing on G1, training on G2 and testing on G2, training on random geometries and testing on random geometries, and by a dummy mean estimator.

8.2.2.1 Neural Network Models and Training

We use a CNN architecture very close to the one presented in Fig. 8.6(B) but with one more layer, with similar training parameters as in Sec. 8.2.1.1 (see [S25] for details). While the network's input is identical, the output is now a vector in $\mathbb{R}^{36}$ containing the absorption coefficients of the 6 walls in 6 octave bands in a fixed order. We generate a training set of 10,500 RIR vectors in $\mathbb{R}^{8000}$ using the simulator `Roomsim` [41] and the acoustical sampling scheme described in Sec. 8.1.1.2 and 8.1.2, but this time with the fixed geometry G1 shown in Fig. 8.10(A). For each simulated RIR, this base geometry is perturbed by additive white Gaussian noise with standard deviation σ_{geo}=2 cm on the positions of the source, microphone and walls. As in Sec. 8.2.1, the RIR signals are also perturbed by additive white Gaussian noise before being normalized, this time with a peak signal-to-noise ratio drawn uniformly at random in $[40, 50]$ dB. A disjoint validation set (for early stopping) and a disjoint test set (for evaluation) of 2,250 RIRs each are generated using the same procedure. This entire data generation and training process is also done using the base geometry G2 of Fig. 8.10(B)

8.2.2.2 Results on Absorption-Profile Estimation

Table 8.1 shows the mean absolute errors on absorption coefficients obtained by training on G1 and testing on G1, training on G2 and testing on G2, and training on *random geometrical sampling* and testing on *random geometrical sampling* (see Sec. 8.1.1.2). The errors are compared to the ones obtained by a dummy *mean estimator* that constantly returns the mean absorption coefficient across the entire training set for each octave band. As can be seen, the model successfully performs the task when training and testing on matched geometries, regardless of the geometry, with mean errors well below 0.1. We did not observe significant differences across octave bands. This suggests that sufficient information is present in a single RIR to recover the full absorption profiles when the geometry is fixed. To refine this observation, we ran the same experiment but cropping the input RIR using various cutoff times. Performance were unaffected as long as the cutoff time was larger than T_1 ≈64 ms, and quickly degraded to reach that of the mean estimator for cutoff times below $T_0 \approx 16$ ms. This suggests that the model primarily leverages information contained in early echoes and in particular first-order ones, which all occurred before T_1 in our datasets. This coincides with the wisdom used to design the physics-driven methods of Chap. 7.

Meanwhile, the results of Table 8.1 also reveal that the model performs close to the mean estimator under randomized geometries. This provides evidence towards the hypothesis made at the beginning of this section, namely, that this task is impossible without geometrical knowledge.

8.2.3 Conclusion on Virtually-Supervised Absorption Estimation from a RIR

This section explored the use of virtually supervised learning for the estimation of absorption coefficients from a single RIR. The focus was on the generation of adequate simulated datasets rather than

on the neural network architecture themselves, which were kept simple. We obtained encouraging generalization capabilities to real measured RIRs for mean absorption coefficients at 1000 Hz and above, with potential to overcome situations in which classical reverberation-theory formulas struggle, *e.g.*, when devices are close to reflectors and/or facing away from each other, causing nonlinear reverberation decays. However, we could not obtain satisfying results at lower frequencies, possibly due to unaccounted physical phenomenons at train time. We also obtained evidence that a single early RIR may contain enough information to estimate all the individual absorption profiles jointly, as long as the geometry of the room and the device placement is fixed and known with sufficient precision. However, the approach requires training a model on a simulated dataset of thousands of RIRs for any given geometry, a process representing days of CPU time. This calls for the development of more advanced and more expressive neural network architectures, that would enable *conditioning* a generic model by the geometry.

A limitation of both works is that only omnidirectional devices were considered at train time, calling for further research to study the impact of directivity on these task. Doing so will however require the development of an efficient simulator that can handle both the acoustic sampling scheme of Sec. 8.1.1.2, including scattering and devices with arbitrary directivity[6].

Another key challenge going forward, for both tasks, will be to obtain ground truth data of high-enough quality to further validate the generalizability of the learned models. The only currently available way to do this would be to build a dedicated room entirely made of materials whose coefficients are known, as measured by their manufacturers in a lab. Even then, there might be variations in the effective absorption properties of materials once mounted on site. Moreover, there exists ambiguities in the very definition of absorption coefficients, as explained in Sec. 6.2.2.1 and 6.3. A better-defined problem would be to recover the materials' *impedance*, for which even less data is available. An alternative promising route for evaluation would be to use *contrastive* metrics. Does the method output similar absorption profiles across different measurement positions in a fixed room? What about different rooms made of similar materials? And if nothing changes except the absorption of one of the surfaces, does the method correctly predict this change? Finally, another approach for evaluation would be to feed the estimated parameters back to a well-controlled, highly-realistic forward model. Its output could then be compared to measured data using various metrics, which could include subjective ones, depending on the considered application.

8.3 Blind Estimation of Global Room Acoustic Parameters

Associated publication: [S28, S31]

So far in this thesis, we have only considered the estimation of acoustical or geometrical parameters of interest from one or several room impulse responses, possibly aided by geometrical knowledge. While one can reasonably expect the availability of such measurements in the context of room acoustic analysis, diagnosis or during the calibration of devices, there are many other applications in which only noisy recordings of unknown sound sources will be available, *i.e.*, the *blind* setting. In this section, and in fact for the remainder of Part II, we will turn our attention towards the blind,

[6] At the time of writing, `pyroomacoustics` [42] can account for each effect separately but not jointly. Such an extension is planned but not trivial to implement efficiently, as time-frequency-angle histograms are required to gather rays at the sources and receivers. This has been done in C/C++ in `Roomsim` [41] but only for one specific binaural head and the software is no longer maintained.

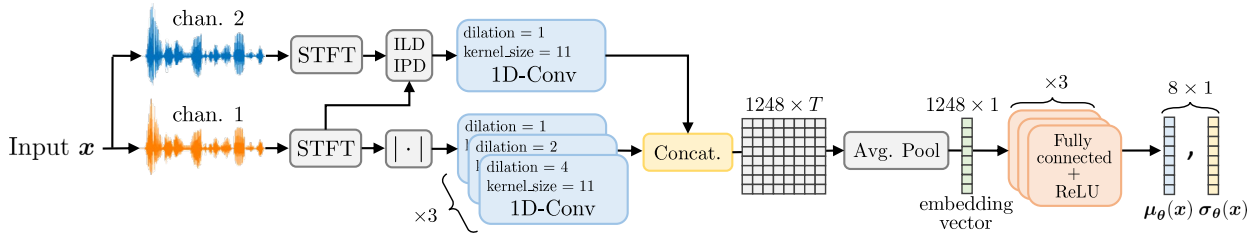


FIGURE 8.11: Diagram of the proposed neural network architecture for global room parameter estimation from multiple, multichannel recordings. The architecture jointly outputs 8 parameters and their estimated variance: the room's volume V, total surface area S and reverberation time $\mathrm{RT}_{60}(b)$ in 6 octave bands.

geometrically-uninformed setting. Of course, the problem then becomes strictly harder, and one cannot reasonably expect to retrieve the same amount of information with the same accuracy as from RIRs.

With this in mind, this section explores the use of virtually-supervised learning for the blind estimation of *global* room parameters, *i.e.*, parameters attached to the entire room rather than to individual reflectors. We focus here on the room's volume V, its total surface area S, and its reverberation time in 6 octave bands $\mathrm{RT}_{60}(b)$ (see Sec. 6.3). This specific set of parameters was identified as a *reverberation fingerprint* in [7], *i.e.*, a compact way to characterize rooms for the realistic binaural rendering of virtual sources on audio-augmented reality headphones (see Sec. 5.1). With this use case in mind, we will consider a practical setup where two-channel, noisy recordings of a single unknown speech source are available, possibly from multiple unknown *viewpoints*, *i.e.*, different source-receiver locations inside the room. Sec. 8.3.1 describes the proposed neural network model and its training, Sec. 8.3.2 studies the impact of the number of channels and measurements on its performance, and Sec. 8.3.3 studies the impact of simulation realism on its real-data generalizability.

8.3.1 Neural Network Model and Training

The proposed neural network architecture[7] is depicted in Fig. 8.11. Single-channel and inter-channel features are extracted from the time-domain two-channel input signal $\boldsymbol{x}$ in the form of spectrograms. We use short-time discrete Fourier transforms (STFT) with 96 ms sliding Hann windows and 50% overlap to obtain a complex spectrogram $\{X_m[f_i, n]\}_{i=1,n=1}^{F,N}$ for each channel m, with $F = 769$ positive frequency bins and $N = 63$ time frames for a 3 s input signal (our architecture works on arbitrary input length). Then, single-channel features are computed as $|X_1[f, n]|$. Inter-channel features are obtained by concatenating inter-channel level differences (ILD) and phase differences (IPD):

$$\mathrm{ILD}[f, n] = \log |X_1[f, n]| - \log |X_2[f, n]| \tag{8.1}$$

$$\mathrm{IPD}[f, n] = \left[\mathrm{Re}, \mathrm{Im}\left(\frac{X_1[f, n]X_2^*[f, n]}{|X_1[f, n]X_2^*[f, n]|}\right)\right] \tag{8.2}$$

[7] Technically, this is the architecture used in [S31] while the original architecture proposed in [S28] also returned area-weighted mean absorption coefficients $\bar{\alpha}(b)$. We choose to omit results obtained on $\bar{\alpha}(b)$ in Sec. 8.3.2 to keep the presentation streamlined, and because of the inherent difficulty of evaluating absorption estimation on real data (see detailed discussion in Sec. 8.2.1.2).

These features are then processed through 1D convolutional blocks (1D-Conv), which were proposed in the Conv-TasNet architecture in the context of speech separation [71]. These blocks consist of separable convolutions (depth-wise and point-wise) intertwined with rectified linear unit (ReLU) activations and followed by layer normalization [72]. The latter proved to be crucial in our experiments, as it creates scale-invariant representations. For single-channel features, three 1D-Conv blocks with increasing dilation factors along the frequency-axis and a kernel size of 11 are used, while only one block is used on inter-channel features, as this showed to give best results. The obtained representations are concatenated along the frequency axis and average-pooled along the time-axis to yield a time-independent, 1248-dimensional *embedding vector*. The embedding vector is finally passed through 3 fully-connected layers to obtain $2 \times D = 16$ outputs consisting of the estimated room parameters $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}^D$ and the estimated variances $\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}^D$ (or uncertainties) on these parameters.

The network parameters $\boldsymbol{\theta}$ are optimized by minimizing the following Gaussian negative log-likelihood loss function:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) &= -\log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = -\log \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\boldsymbol{x})) \\ &\stackrel{c}{=} \frac{1}{2} \sum_{d=1}^{D} \log \sigma^2_{d,\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{(y_d - \mu_{d,\boldsymbol{\theta}}(\boldsymbol{x}))^2}{\sigma^2_{d,\boldsymbol{\theta}}(\boldsymbol{x})}\end{aligned} \tag{8.3}$$

where $\boldsymbol{y} \in \mathbb{R}^D$ denotes the true room parameters. A benefit of this approach is that it adaptively weights errors on individual parameters. The estimated variances can also be used to fuse estimates obtained from J independent observations $\{\boldsymbol{x}_j\}_{j=1}^J$ of the same room using the following formula derived from Bayes' rule:

$$p_{\boldsymbol{\theta}}(y_d|\bar{\boldsymbol{x}} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J]) = \mathcal{N}(y_d; \bar{\mu}_{d,\boldsymbol{\theta}}(\bar{\boldsymbol{x}}), 1/\bar{\gamma}^2_{d,\boldsymbol{\theta}}(\bar{\boldsymbol{x}})) \tag{8.4}$$

$$\text{with } \bar{\mu}_{d,\boldsymbol{\theta}}(\bar{\boldsymbol{x}}) = \sum_{j=1}^{J} \frac{\gamma^2_{d,\boldsymbol{\theta}}(\boldsymbol{x}_j)}{\bar{\gamma}^2_{d,\boldsymbol{\theta}}(\bar{\boldsymbol{x}})} \mu_{d,\boldsymbol{\theta}}(\boldsymbol{x}_j), \;\; \bar{\gamma}^2_{d,\boldsymbol{\theta}}(\bar{\boldsymbol{x}}) = \sum_{j=1}^{J} \gamma^2_{d,\boldsymbol{\theta}}(\boldsymbol{x}_j)$$

where $\gamma^2_{d,\boldsymbol{\theta}}(\boldsymbol{x}_j) = 1/\sigma^2_{d,\boldsymbol{\theta}}(\boldsymbol{x}_j)$ is the estimated *precision* for observation $\boldsymbol{x}_j$ and $\bar{\boldsymbol{\mu}}_{d,\boldsymbol{\theta}}(\bar{\boldsymbol{x}})$ is the fused estimate.

To train the network, we first generate a set of 20k rooms annotated with their respective ground truth parameters, simulated using `Roomsim` [41], the geometrical and acoustical sampling schemes described in Sec. 8.1.1.2 and 8.1.2, an omnidirectional source, and an omnidirectional-microphone pair with an aperture of 22.5 cm to match the considered headset use case. For each room, 5 RIRs corresponding to different random source-receiver positions are generated, yielding 100k RIRs in total. For each room and each octave band, a unique reverberation time is estimated by taking the median value over the 5 source-receiver positions available. The values are obtained by linear regression over the -5 dB to -25 dB decay of Schroeder curves [50]. This yields training parameters in the respective ranges $V \in [18, 400]$ m^3, $S \in [48, 360]$ m^2 and $\text{RT}_{60}(b) \in [0.2, 3.2]$ s.

The obtained RIRs are then downsampled to 16 kHz and convolved with random speech excerpts from the LibriSpeech corpus [73]. The resulting 3 s two-channel reverberated signals are then corrupted with both static microphone noise, i.e., independent additive white Gaussian noise on each channel, and spatially-diffuse babble noise, i.e., speech-shaped noise convolved with the late part (>50 ms) of an additional random RIR in the room. For the noise levels to be realistic, signals from sources that are placed further away from the receiver should exhibit lower signal-to-noise ratios

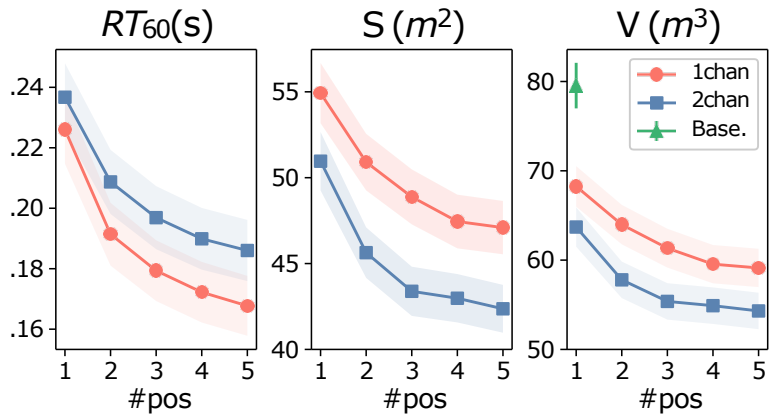


FIGURE 8.12: Mean absolute errors achieved on simulated data by the proposed virtually-supervised model with one- or two-channel inputs, as a function of the number of source-receiver positions fused in each room. Shaded areas indicate 95% confidence intervals. Volume estimation errors are compared to the baseline [30] (Base.).

(SNRs). To achieve this, for each room, we first generate a *reference* signal using a random speech source placed 1 meter in front of a receiver (not used in the final dataset). Static and diffuse noise levels for this signal are set to obtain SNRs drawn uniformly at random in $[70, 90]$ dB and $[30, 60]$ dB, respectively (different levels for each source-receiver position in the room are set). These noise levels are then kept fixed for the final mixtures, irrespective of the distance from the speech source to the receiver. This resulted in an overall SNR range of $[-10, 65]$ dB across the dataset. The 100k annotated two-channel noisy speech signals are divided into training, validation and test sets of respective size 80k, 10k and 10k with no room or speech-signal overlap between them.

To avoid issues due to scale differences, the ground truth parameters are normalized at training time by dividing them by their standard deviations over the training set, which are saved and multiplied with the network output at test time. The network is trained using the ADAM optimizer [54] with a learning rate of 10^{-4} and batches containing 120 speech signals selected at random from the training set. A dropout rate of 0.2 and 0.4 was used in conv-blocks and in fully connected layers to avoid over-fitting. We used a patience of 15 epochs on the validation set for early stopping. Training generally converged in 100–150 epochs. The code to reproduce this model and the subsequent experiments is on Prerak Srivastava's GitHub `https://github.com/prerak23/`.

8.3.2 Impact of the Number of Channels and Measurements

We first evaluate the proposed approach on the 2,000 unseen rooms of the simulated test set, using between 1 and 5 two-channel 3-s noisy speech recordings at different source-receiver positions per room, based on the fusion method given in Eq. 8.4. Two variants of our model are compared: the full two-channel architecture depicted in Fig. 8.11, and the same architecture without the upper inter-channel processing part of the network, i.e., using only one channel. These variants are compared to our implementation of the single-channel, single-position blind room volume estimation method of [30] trained on the same data. The metric used is the mean absolute error on each of the room parameter, where errors are averaged over all octave bands for the RT_{60}.

As can be seen in Fig. 8.12, increasing the number of fused observations per room significantly reduces errors on all parameters. Using five positions and two channels, mean absolute errors of 0.18 s

Method	Features	# pos	RT_{60}	S	V
[30]	Single channel	1	-	-	137.8
Ours	Single channel	1	0.134	129.6	154.5
Ours	Single channel	5	0.097	125.8	149.1
Ours	Both	1	0.101	89.4	107.6
Ours	Both	5	**0.062**	**50.2**	**68.8**

TABLE 8.2: Mean absolute error achieved over 3 rooms from the real dEchorate dataset, using 1 or 5 viewpoints per room, with or without inter-channel features. Bold numbers indicate the best statistically significant result per column, based on 95% confidence intervals.

Method	Features	# pos	RT_{60}	S	V
[30]	Single channel	1	-	-	10.0
Ours	Single channel	1	0.161	27.2	31.8
Ours	Single channel	5	0.090	19.6	23.0
Ours	Both	1	0.100	34.7	39.7
Ours	Both	5	0.054	16.5	18.9

TABLE 8.3: Standard deviation of parameter estimates for room "011100" of the real dEchorate dataset (see Sec. 9.1).

for RT_{60}, 42 m^2 for S and 54 m^3 for V are obtained. The proposed model significantly outperforms the one in [30] for volume estimation, reducing the error by 13% using one channel and one observation, and by 31% using two channels and five observations. Interestingly, we observe that using two channels instead of one significantly reduces surface-area and volume estimation error but does not significantly impact reverberation time estimation, according to 95% confidence intervals. This may be interpreted by the fact that the latter mostly govern the late, spatially-diffuse regime of RIRs, and hence should have limited correlation with inter-channel cues that mostly capture spatial characteristics. Conversely, S and V are inherently spatial quantities as they relate to the room's geometry and hence early echoes, which do correlate with inter-channel cues.

To check how well our virtually-supervised model generalizes to real data, we use again the dEchorate dataset [S24], which will be described in details in Sec. 9.1. This time, we use wet speech recordings from the dataset rather than RIRs, recorded in a variable-acoustic room of size $5.7 \times 6 \times 2.4$ m ($S = 125$ m^2, $V = 82$ m^2). 5 arrays of 6 omnidirectional microphones and 6 directional loudspeakers are placed inside the room, yielding $5 \times 6 = 30$ multi-channel speech recordings per room configuration. The ground truth RT_{60} of each room configuration in the four octave bands from 500 Hz to 4 kHz are computed as in Sec. 8.2.1.2, and for the same reasons, results in lower octave bands are omitted. In our experiments, we use $3 \times 30 = 90$ three-second speech recordings, corresponding to the 2-channel sub-arrays with aperture 22.5 cm and to the 3 room configurations involving 3 or more reflective surfaces, as these most closely match the considered scenario. For these rooms, the RT_{60} ranges from 0.25 to 0.66 s.

Table 8.2 reports mean absolute errors using the proposed approach with or without inter-channel features. We report results using either 1 or 5 source positions and a fixed receiver. For the latter, we exclude one out of the 6 available source positions for each test, so that there are 90 tests in each case. The mean absolute volume estimation error using the single-channel, single-position baseline [30] is reported as well. Encouragingly, errors obtained with our approach are of comparable orders to those obtained on simulated data. In the single-channel, single-position case, volume estimation errors obtained with our model are comparable to [30]. We are able to systematically reproduce the

Train. set	walls	src	mic	RT_{60} (.5 kHz)	RT_{60} (1 kHz)	RT_{60} (2 kHz)	RT_{60} (4 kHz)	S	V
D1	$\mathcal{N}$	$\mathcal{O}$	$\mathcal{O}$	0.193	0.160	0.108	0.185	71.00	75.68
D2	$\mathcal{RB}$	$\mathcal{O}$	$\mathcal{O}$	0.182	0.140	0.128	0.198	45.11	55.16
D3	$\mathcal{RB}$	$\mathcal{O}$	$\mathcal{M}$	0.115	**0.098**	0.078	0.156	52.76	61.82
D4	$\mathcal{RB}$	$\mathcal{M}$	$\mathcal{O}$	0.133	0.112	**0.066**	0.155	**21.46**	**18.57**
D5	$\mathcal{N}$	$\mathcal{M}$	$\mathcal{M}$	0.151	0.133	0.084	0.159	35.88	31.11
D6	$\mathcal{RB}$	$\mathcal{M}$	$\mathcal{M}$	**0.080**	**0.103**	**0.064**	**0.140**	32.69	30.57

TABLE 8.4: Mean absolute errors obtained by the proposed virtually-supervised model trained on 6 different datasets when estimating RT_{60} (sec), S (m^2) and V (m^3). The models are evaluated on 560 sets of 3 real two-channel speech recordings from the dEchorate dataset [S24]. Bold numbers indicate the best statistically significant result per column, based on 98% confidence intervals.

observation that increasing the number of source-receiver positions significantly decreases errors. We also observe again that using inter-channel features significantly improves the estimation of S and V, while having less impact on RT_{60}.

Finally, Table 8.3 shows the standard deviations of estimated values by the same methods over the 30 recordings from the room with 3 reflective surfaces. Encouragingly, the relatively low standard deviations reveal the ability of the models to provide parameter estimates that are stable within a room, and do not depend much on the source-receiver position. Moreover, it can be seen that, as expected, using five observations in a room instead of one systematically decreases the standard deviation of estimates.

8.3.3 Impact of Simulation Realism

We will now study the impact of simulation realism, and in particular the use of real directivity patterns on devices (Sec. 8.1.3), and of the reflectivity-biased sampling strategy on walls (Sec. 8.1.2) on the generalizability of the proposed virtually-supervised model to real data. To this end, we retrain the model of Sec. 8.3.1 on 6 different datasets, numbered D1 to D6, simulated using `pyroomacoustics`[8] [42] and the same geometrical sampling scheme, but incorporating different levels of wall, source and receiver realism, as summarized in the first four columns of Table 8.4. This time, each training set consists of 30k random rooms with 3 source-receiver placements per room.

The wall absorption profiles are sampled using the *naive* scheme ($\mathcal{N}$) in D1 and D4, and the *reflectivity-biased* scheme ($\mathcal{RB}$) in D2, D3, D5 and D6, as described in Sec. 8.1.2. Datasets D1, D2 and D3 use omnidirectional sources while dataset D1, D2 and D5 use omnidirectional receivers. For each source in D4, D5 and D6, a random directivity pattern among the Genelec 8020, Neumann KH120A and Yamaha DXR8 loudspeakers of the DirPat dataset [49] is randomly oriented with pointing direction parallel to the floor. For each individual receiver in D3, D5 and D6, the directivity pattern of the omnidirectional AKG C414 microphone of DIRPAT is rotated uniformly at random over the sphere.

The model trained on D1, ..., D6 is evaluated on a subset of real two-channel speech recordings from the dEchorate dataset [S24], similarly to the previous section. We use this time 4 room configurations with 2 to 5 reflective surfaces, yielding reverberation times ranging from 250 to 810 ms. The

[8] As mentioned in Sec. 8.2.3, the current version of `pyroomacoustics` cannot jointly model scattering coefficients and directivity. The former effect is hence dropped in the experiments of this section.

microphones in this dataset are omnidirectional AKG CK32, and each room configuration includes 6 Avanton MixCubes loudspeaker and one lightweight Brüel & Kjær omnidirectional loudspeaker. Note that none of these receivers and sources are present in the directivity dataset used to build the training sets. For evaluation purposes, we consider all 140 possible combinations of three two-element arrays recording a fixed source in each of the 4 rooms, resulting in 560 test cases in total, with three viewpoints per test.

The results are shown in Table 8.4. We first see that the most realistic training set D6 yields best or second best results for all quantities. Interestingly, the dataset D4, which is identical but with a simpler omnidirectional microphone model, results in better source and volume estimation than D6 on the real test set, suggesting that mismatched microphone responses at train and test times can degrade the accuracy. A solution could be to use more diverse microphone responses at train time. More generally, by comparing D2 to D3 on the one hand and D4 to D5 on the other hand, we see that using measured microphone directivities significantly improves RT_{60} estimation over both test sets, but tends to degrade S and V estimation. By comparing the results obtained using D1 and D2 on the one hand and D5 and D6 on the other hand, we see that the more realistic *reflectivity-biased* ($\mathcal{RB}$) sampling strategy for wall absorption coefficients significantly and consistently outperforms the naive scheme ($\mathcal{N}$) across all target quantities. Finally, by comparing results obtained using D2 and D4, we observe that increasing the realism of source directivity consistently improves results across all target quantities. This trend is dramatically confirmed by comparing the results obtained with D3 and D6, especially in geometry estimation. Overall, the results reveal that every added layer of simulation realism at train time on the source, receiver and walls improve the generalization capability of the model to real data.

Finally, comparing[9] the results obtained using D2 with those in the last two rows of Table 8.2 suggests that, encouragingly, training our model using `Roomsim` [41] or `pyroomacoustics` [42] yields comparable generalization performance to real data. This could indicate that modeling scattering does not strongly impact the estimation of these room parameters in the context of virtually-supervised learning, although more evidence is needed to validate or refute this hypothesis. We note that a similar observation regarding the impact of scattering was made in [74] in the context of sound source localization. However, a plausible explanation in that case is that the relevant information is concentrated in the direct path, while reverberation, including scattering, may be viewed as an adversarial effect against which the trained model could build robustness through other data-augmentation schemes. Things are less clear when it comes to global room parameters for which, in principle at least, the reverberation field in its entirety could be exploited. Recent theoretical developments in understanding spatio-temporal correlations in late reverberation fields could help shedding light on this issue [22].

8.4 Conclusion on Virtually Supervised Learning

In this chapter, we explored the use of virtually-supervised learning for three acoustic parameter estimation tasks: (i) mapping a single-channel RIR to the room's area-weighted mean absorption coefficients, (ii) mapping a single-channel RIR to the 6 absorption profiles of the walls under a fixed

[9] Note however that the number of rooms in each training set, the number of observations per room used at test time, and the subsets of dEchorate used for evaluation, though comparable, are not strictly identical, making the comparison limited.

and approximately known geometry, and (iii) mapping a set of noisy speech recordings to the volume, surface-area and reverberation time of the room. Along the way, we developed two methods to improve the representativity of wall absorption profiles and the realism of device directivities in the training data. Both techniques proved to consistently and significantly improve the generalizability of trained models to real data on task (iii). While simple multi-layer perceptron or convolutional neural network models with no specific data pre-processing were used for tasks (i) and (ii), a more advanced architecture allowing for variable input size and multi-task training was developed for task (iii).

While encouraging results were obtained on real data for tasks (i) and (iii), further work is needed to handle frequencies around 500 Hz (Schroeder's frequency) and below, and the acquisition of a dedicated dataset will be required to evaluate task (ii) on real data. The approach used for task (ii) is also very costly computationally, requiring the generation of a training set for each geometry. This calls for the development of more flexible architecture that would allow the estimation of geometry-dependent parameters. For all tasks, further studies are needed to better understand the impact of measurement noise and of scattering effects on the results.

9 Echo-Aware Audio Signal Processing

Associated publications: [S12, S13, S17, S21, S24]

In this last contribution chapter, we focus on a concept that resonated[1] throughout this thesis, namely, *acoustic echoes*. One way to explain why echoes are such a useful and pervasive concept in the physics-driven room-parameter estimation literature [66, 75, 76] is that they form a perfect *intermediate quantity* between observations and targets. While they live in signal space, they disentangle individual reflection paths from which acoustical and geometrical parameters can be retrieved. Moreover, estimating echoes from measured signals and estimating acoustic properties from echoes represents two subtasks of very different nature and roughly comparable complexity, yielding a natural and effective decomposition of the problem.

Developing benchmarks to evaluate progress on each subtask independently seems important to foster research in this area. Sec. 9.1 presents our contribution to this effort, in the form of a real RIR dataset called *dEchorate*, recorded in a variable-acoustics room. The RIRs in the dataset are annotated by both the geometrical parameters and the times of arrival of first-order echoes in a *mutually consistent way*, which is a unique feature to the best of our knowledge.

Then, Sec. 9.2 presents three of our contributions to the first subtask. More specifically, it succinctly introduces three methods (as well as two baselines) to blindly estimate the times of arrival and amplitudes of early echoes from multichannel audio signals, two of them physics-driven, the last one data-driven.

We then ask whether such estimated echoes could have applications beyond their role as intermediate quantities in room parameter estimation. Sec. 9.3 explores this question through three more contributions: echo-aware sound-source localization, echo-aware speaker separation, and echo-aware beamforming. The section highlights key results without detailing the methods, using them as proof-of-concept illustrations of the potential of echo-aware audio signal processing. We finally conclude this chapter in Sec. 9.4.

9.1 dEchorate: An Annotated Echo Dataset

Associated publication: [S24]

9.1.1 Data Acquisition

The recording setup is placed in an empty cuboid room of dimensions $6\times6\times2.4$ meters located at the acoustic lab at Bar-Ilan University. Inside the room, 30 omnidirectional AKG CK32 microphones are mounted on 6 static non-uniform linear arrays of 5 elements each, placed parallel to the ground. 4 Avanton MixCubes directional loudspeakers are facing the center of the room and 2 more are pointing towards the walls, to study the case of early reflections being stronger than the direct path. Each loudspeaker and each array is positioned closer to one of the walls in such a way that the origin of

[1] Pun intended!

(A) (B) (C)

FIGURE 9.1: Photos of the setup used to record the dEchorate dataset. From left to right: the overall setup, one microphone array, the setup with revolved panels.

	Surfaces:	Floor	Ceil	West	South	East	North
one-hot	010000	✗	✓	✗	✗	✗	✗
	001000	✗	✗	✓	✗	✗	✗
	000100	✗	✗	✗	✓	✗	✗
	000010	✗	✗	✗	✗	✓	✗
	000001	✗	✗	✗	✗	✗	✓
incremental	000000	✗	✗	✗	✗	✗	✗
	010000	✗	✓	✗	✗	✗	✗
	011000	✗	✓	✓	✗	✗	✗
	011100	✗	✓	✓	✓	✗	✗
	011110	✗	✓	✓	✓	✓	✗
	011111	✗	✓	✓	✓	✓	✓
f.	010001*	✗	✓	✗	✗	✗	✓

TABLE 9.1: Room encoding in the dataset: each binary digit indicates if the surface is absorbent (0, ✗) or reflective (1, ✓). In configuration 010001*, furniture (f.) were used.

the strongest echo can be easily identified. Moreover, their positioning is chosen to cover a wide distribution of source-receiver-wall distances. Photos of the setup are shown in Fig. 9.1.

The main feature of the room is the possibility to change the acoustic profile of each of its facets (walls, floor, ceiling) by flipping double-sided panels with one *reflective* face (made of Formica Laminate sheets) and one *absorbing* face (made of perforated panels filled with rock-wool). This allows achieving diverse values of RT_{60} ranging from 0.1 to almost 1 second. In this dataset, the panels of the floor were always kept absorbent. 10 different acoustic configurations are considered, as summarized in Table 9.1. The dataset also features an eleventh recording session in which office furniture are positioned to simulate a typical meeting room with chairs, tables, a coat hanger and a head-and-torso manikin.

For each source-microphone-room combination, one RIR is measured using the *exponential sine sweep* technique with three repetitions [77], yielding a dataset of $6 \times 30 \times 11 = 1,980$ unique RIRs at 48 kHz and 32 bits per sample. An example of such RIR is shown in Fig. 9.2. Additional recordings of white noise, babble noise and random speech excerpts from the LibriSpeech corpus [73] are also performed, including some emitted with an additional lightweight Brüel & Kjær omnidirectional loudspeaker and 4 additional 6301bx Fostex loudspeakers. The dEchorate dataset and associated

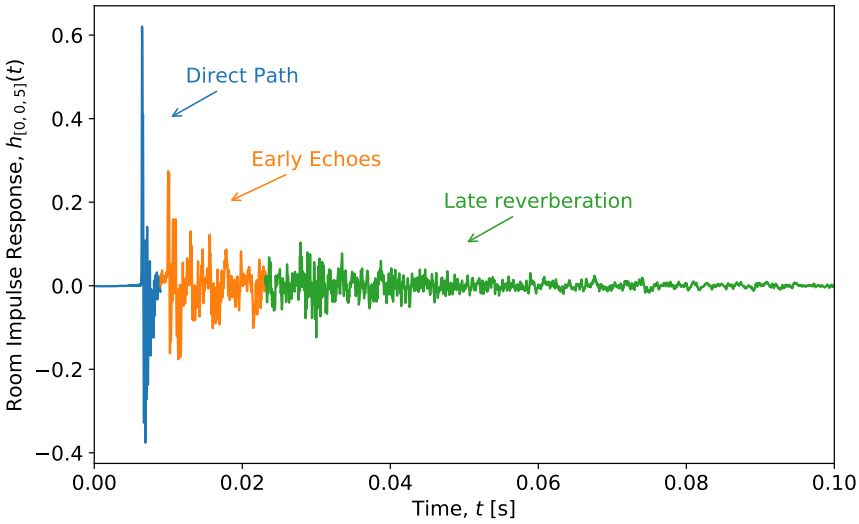


FIGURE 9.2: Example of a room impulse response from the dEchorate dataset.

software can be accessed from Diego Di Carlo's GitHub `https://github.com/Chutlhu/dEchorate`.

9.1.2 Dataset Annotation and Visualization

The main objective of this dataset is to feature annotations in the *geometrical space*, namely microphone, wall and source positions, that are fully consistent with annotations in the *signal space*, namely the (first order) echo times of arrival within the RIRs. This is achieved as follows:

(i) First, the ground-truth positions of the microphone array and source centers are acquired via a beacon-based indoor positioning system. This system consists in 4 stationary bases positioned at the corners of the ceiling and a movable probe used for measurements which can be located within errors of ± 2 cm.

(ii) The measured RIRs are then superimposed on RIRs simulated using the vanilla image source model of `pyroomacoustics`[42] (See Sec. 6.2) and the geometry obtained in the previous step. A Python graphical user interface[2] is then used to manually tune a peak finder and label all the echoes discovered this way with their timings and their corresponding image source position and room wall.

(iii) By solving a simple multidimensional scaling (MDS) problem as in [78], refined microphone and source positions are then computed from labeled echo timings. The non-convexity of the problem is alleviated by using a good initialization (obtained at the previous step), by the high signal-to-noise ratio (SNR) of the measurements and, later, by including additional image sources in the formulation. The prior information about the arrays' internal geometry reduces the number of variables of the problem, leaving as only unknown the 3D positions of sources and the arrays' barycenters and tilts on the azimuthal plane.

[2]This GUI is available in the dataset package `https://github.com/Chutlhu/dEchorate`.

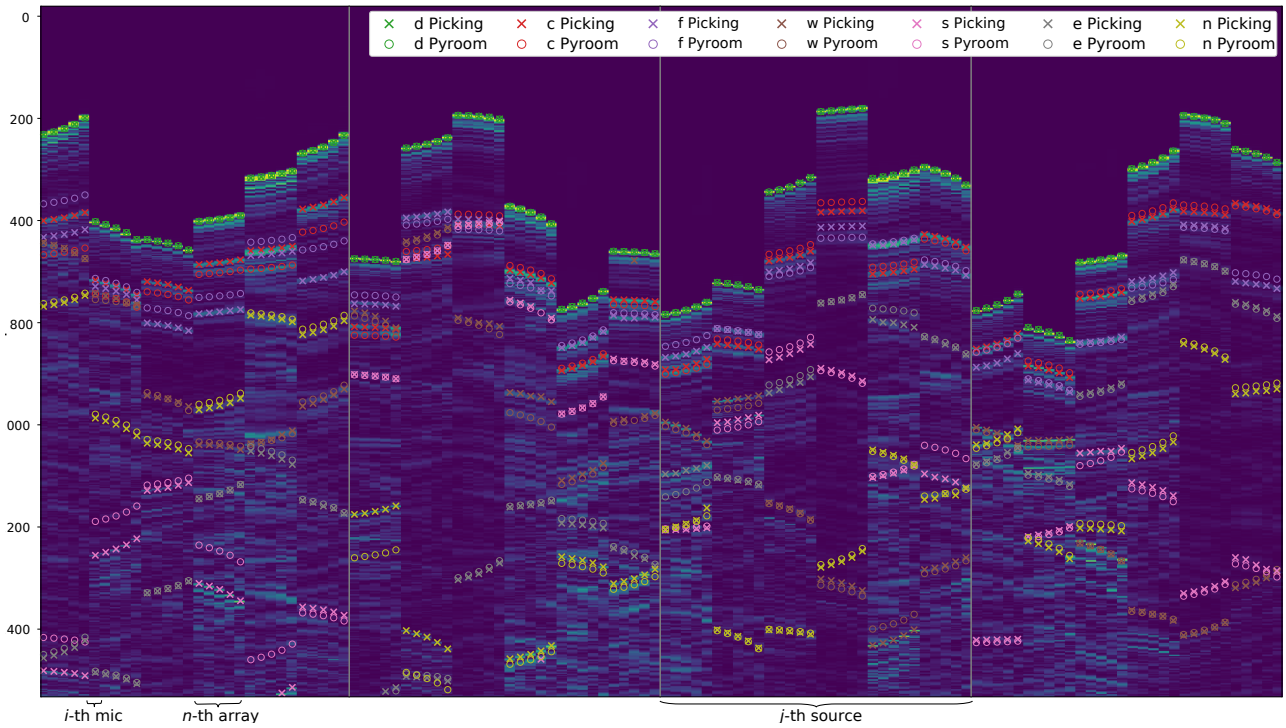


FIGURE 9.3: Example of a RIR *"skyline"* annotated with observed peaks (×, Picking) together with their geometrically-expected timing (○, Pyroom) computed with `pyroomacoustics` [42]. As specified in the legend, markers of different colors are used to indicate the room facets responsible for the reflection: direct path (`d`), ceiling (`c`), floor (`f`) and west (`w`), south (`s`), east (`e`) and north (`n`) walls.

(iv) Finally, by employing a multi-lateration algorithm [79], where the positions of one microphone per array serve as anchors and the times of arrival are converted into distances, additional image sources are localized alongside the real sources.

Knowing the geometry of the room, we calculate an initial estimate of the timings of echoes in the RIRs in step (i). Then, by iterating through steps (ii), (iii) and (iv), the echo timings are refined to be consistent under the image source model. The outcome is a set of "compromise" geometrical parameters and echo timings featuring a small mismatch in *both* signal space and geometrical space. The geometrical mismatch is of 0.4 cm on average and 1.86 cm max, with 98.1% of detected peaks matching first-order echo timings within a 0.5 ms threshold, which corresponds to errors below 1.7 cm on image source locations.

To visualize the annotated dataset, we developed the *"skyline"* tool. The idea is to represent the normalized absolute values of multiple RIRs as an image, such that the wave fronts corresponding to echoes can be highlighted. Let $x_m[n]$ be a discrete-time RIR from the dataset where $n \in [\![0, N-1]\!]$ and $m \in [\![0, M-1]\!]$ indexes source-microphone pairs in a fixed room configuration. Then, the *skyline* is the visualization of the $N \times M$ matrix created by stacking column-wise M normalized *echograms* defined by $|x_m[n]| / \max_n(|x_m[n]|)$. Fig. 9.3 shows an example of such skyline for 120 RIRs corresponding to 4 directional sources, 6×5 microphones, and the most reflective room configuration, stacked horizontally, preserving the order of microphones within each array. The theoretical echo timings computed using the image source model with the corrected geometry (Pyroom) and the labeled echo peaks in the RIRs (Picking) are superimposed on the image. As can be seen, an excellent

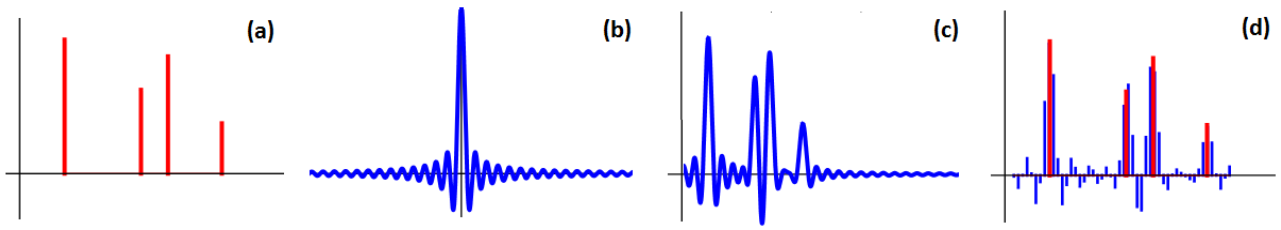


FIGURE 9.4: (a) Continuous-time stream of Diracs (b) ideal low-pass (sinc) kernel (c) stream filtered by the kernel (d) sampled version of the filtered stream (blue) overlayed with the orignal stream (red). This figure illustrates that discretely-sampled Dirac streams are generally not sparse, and that retrieving the Diracs' support is non-trivial.

match is obtained. One can notice several visually-pleasing clusters of 5 adjacent bins corresponding to reflected wavefronts arriving at 5-microphone arrays, nicely illustrating the overarching topic of this thesis: "hearing the walls of a room".

9.2 Blind Acoustic Echo Retrieval

Associated publication: [S12, S17, S21]

This section presents three methods to blindly estimate the times of arrival and amplitudes of early acoustic echoes from the two-channel recording of an unknown source. We refer to this task as *blind acoustic echo retrieval*. This problem is in fact quite general, and has also been referred to as *passive echolocation* or *sparse blind system identification* in the literature. Beyond its close connection to our familiar room-parameter estimation tasks [66, 80, 75, 76] it has applications in the fields of sonars [81], seismology [82], or ultrasounds [83].

Previously known approaches to this problem operated in the discrete-time domain, *e.g.* [84], and generally employed a sparsity criterion to retrieve a multichannel filter, to which peak-picking was subsequently applied, *e.g.*, [85, 86, 80, 76]. For context, the baseline approaches in [84] and [85] are briefly reviewed in Sec. 9.2.1. However, this *on-the-grid* paradigm suffers from intrinsic limitations. First, time-domain peak-picking fails when peaks are overlapping and distorted due to filtering effects. Second, the best achievable accuracy is fundamentally limited to half the discrete time step. Third, sparse optimization over a discrete grid is known to suffer from the so-called *basis-mismatch* problem [87]. These issues are illustrated in Fig. 9.4.

In contrast, we present here three methods that operate *off-the-grid* and directly retrieve times of arrival in the continuous-time domain. The methods in Sec. 9.2.2 and 9.2.3 are physics-driven and based on the *vanilla* image source model of Sec. 6.3. They borrow from the super-resolution literature and are, as such, related to the image-source retrieval method of Sec. 7.2.1, except that they are blind and operate in 1D time instead of 3D space. Meanwhile, the method of Sec. 9.2.4 is data-driven and tackles the task by virtually-supervised learning.

9.2.1 On-the-Grid Sparse Blind System Identification

Recall from Eq. 6.6 that when a source emits a signal $s(t)$ in a room, the resulting pressure signal $x_m(t)$ at a receiving point $m \in [\![1, M]\!]$ can be expressed as

$$x_m(t) = (h_m * s)(t) \tag{9.1}$$

where $h_m(t)$ is the continuous-time RIR from the source to point m. Assuming $s(t)$ is band-limited with maximum frequency $f_s/2$ and recalling Eq. 1.7 and the surrounding discussion, this can be approximated in discrete-finite time by:

$$x_m[n] = (h_m \circledast s)[n] \tag{9.2}$$

when $x_m[n]$, $h_m[n]$ and $s[n]$ are appropriately sampled and their respective number of samples[3] N, L and $N + L - 1$ is sufficiently large. This can be conveniently rewritten in matrix-vector form as

$$\boldsymbol{x}_m = \mathrm{Toep}_0(\boldsymbol{h}_m)\boldsymbol{s} = \mathrm{Toep}(\boldsymbol{s})\boldsymbol{h}_m \in \mathbb{R}^N, \tag{9.3}$$

where for any $\boldsymbol{h} \in \mathbb{C}^L$ and $\boldsymbol{s} \in \mathbb{C}^{N+L-1}$ we define the following Toeplitz matrices:

$$\mathrm{Toep}_0(\boldsymbol{h}) \overset{\text{def}}{=} \begin{bmatrix} h_L & \dots & h_1 & 0 & \dots & \dots & 0 \\ 0 & h_L & \dots & h_1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & h_L & \dots & h_1 \end{bmatrix}, \mathrm{Toep}(\boldsymbol{s}) \overset{\text{def}}{=} \begin{bmatrix} s_L & s_{L-1} & \dots & s_1 \\ s_{L+1} & s_L & \ddots & s_2 \\ \vdots & \ddots & \ddots & \vdots \\ s_{L+N-1} & s_{L+N} & \dots & s_N \end{bmatrix}. \tag{9.4}$$

In our setting, $M = 2$ and $\{\boldsymbol{h}_1, \boldsymbol{h}_2\}$ are discrete low-passed RIRs containing the desired information on echoes. A first natural step is then to estimate $\{\boldsymbol{h}_1, \boldsymbol{h}_2\}$ given $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ only. A classical approach to treat this blind system identification task relies on the observation that under noiseless conditions, the following *cross relation* holds:

$$\boldsymbol{h}_1 \circledast \boldsymbol{x}_2 - \boldsymbol{h}_2 \circledast \boldsymbol{x}_1 = \boldsymbol{0} \tag{9.5}$$

by associativity of the convolution. The problem can then be recast as the following minimization:

$$\underset{\boldsymbol{h},\, h_{1,1}=1}{\mathrm{argmin}} \|\mathrm{Toep}(\boldsymbol{x}_2)\boldsymbol{h}_1 - \mathrm{Toep}(\boldsymbol{x}_1)\boldsymbol{h}_2\|_2^2 = \left\| [\mathrm{Toep}(\boldsymbol{x}_2), \mathrm{Toep}(\boldsymbol{x}_1)] \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \end{bmatrix} \right\|_2^2 \tag{9.6}$$

which is a simple linear least-square problem in $\boldsymbol{h} \overset{\text{def}}{=} [\boldsymbol{h}_1^\top, \boldsymbol{h}_2^\top]^\top$, amenable to, *e.g.*, the conjugate gradient method [53]. This approach was already studied in the 90s [84]. The so-called *anchor constraint* $h_{1,1} = 1$ is used to avoid the trivial solution $\boldsymbol{h}_1 = \boldsymbol{h}_2 = \boldsymbol{0}$, but also to alleviate the global *shift-and-scaling* ambiguity that exists between $\boldsymbol{h}$ and $\boldsymbol{s}$, which is inherent to the problem.

In practice, the matrix $[\mathrm{Toep}(\boldsymbol{x}_2), \mathrm{Toep}(\boldsymbol{x}_1)]$ often has low row-rank, making (9.6) ill-posed with infinitely-many solutions. To counter this, a common approach that seems well-suited to the echo-retrieval task at hand is to assume that $\boldsymbol{h}$ is *sparse*. Sparsity is usually promoted using an ℓ_1-norm

[3] We use here the convention of *valid* convolution, *i.e.*, no zero-padding.

penalty term. For instance, in [85], the following LASSO problem [88] is considered:

$$\underset{\boldsymbol{h},\, h_{1,1}=1}{\operatorname{argmin}} \left\| [\operatorname{Toep}(\boldsymbol{x}_2), \operatorname{Toep}(\boldsymbol{x}_1)] \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \end{bmatrix} \right\|_2^2 + \lambda \|\boldsymbol{h}\|_1, \tag{9.7}$$

and a Bayesian-learning method for the automatic inference of λ is proposed. Many other approaches in the literature rely on a comparable scheme, *e.g.*, [86, 80, 76]. These are typically followed by peak picking to estimate the echo timings on the discrete-time grid. Four important bottlenecks of this framework can be identified in the context of acoustic echo retrieval:

1. Although these methods rely on sparsity-enforcing regularizers, discrete RIRs $\boldsymbol{h}$ are strictly-speaking non-sparse, due to low-pass sampling effects, even when $N \to \infty$ (see Fig. 9.4 for illustration).

2. The methods suffer from the problem called *basis mismatch* in the compressed sensing literature [87]. In particular, the true peaks of retrieved filters might not correspond to true echoes, and the heights of the peaks never correspond to amplitudes.

3. The methods' output is fundamentally *on-the-grid*, preventing sub-sample resolution, which may be important in applications such as room shape reconstruction (see Sec. 7.2.2.4 for an example). While the number of modeled RIR samples L for a fixed time-length can in-principle be arbitrarily increased by interpolation, the dimension of the search space is $2L - 1$. This is typically much larger than the actual number $4K$ of unknown variables (the timings and amplitudes of the first K echoes in each channel), and can become computationally prohibitive when considering RIRs with thousands of samples (the complexity of linear inverse solvers is at best quadratic in the search-space dimension).

4. These methods are known to be sensitive to the chosen time-length of the RIRs, which is never available in practice.

The methods presented in the following three sections attempt to alleviate these bottlenecks by directly estimating the echo times of arrival in continuous time.

9.2.2 MULAN: Blind Multichannel Annihilating

Associated publication: [S12]

9.2.2.1 Method

The first proposed method operates in the discrete Fourier domain, where the convolution theorem applied to (9.1) approximately gives[4], for large enough N:

$$\hat{x}_m[f] = \hat{h}_m[f]\hat{s}[f]. \tag{9.8}$$

[4] See Eq. 1.4 and the surrounding discussion.

Moreover, based on the *vanilla* image source method (6.16), the discrete Fourier transform of $\hat{h}_m[f]$ for large enough L has the form (up to normalizing factors):

$$\hat{h}_m[f] = \sum_{k=1}^{K} \frac{a_{m,k}}{c\tau_{m,k}} e^{-2\pi j f \tau_{m,k}} \tag{9.9}$$

where $\{a_{m,k}\}_{m,k=1}$ and $\{\tau_{m,k}\}_{m,k=1}$ denote the desired echo attenuations and times of arrival in seconds[5]. Recall from (1.2) that the discrete set of frequencies $\mathcal{F} = \{f_0, \dots, f_{F-1}\}$ is in arithmetic progression, and call its step Δ_f. We can rearrange (9.9) to make it take the form of a linear combination of *geometric series*:

$$\hat{h}_m[f_i] = \sum_{k=1}^{K} b_{m,k}(q_{m,k})^i, \text{ where } b_{m,k} = \frac{a_{m,k}}{c\tau_{m,k}(q_{m,k})^{f_0}} \text{ and } q_{m,k} = e^{-2\pi j \Delta_f \tau_{m,k}}. \tag{9.10}$$

This enables us to use the so called *annihilating filter trick*, also known as Prony's method [89, 90]. This technique is based on the remarkable observation that:

$$[1, -q] \circledast [q^0, q^1, q^2, \dots, q^{F-1}] = \mathbf{0}_{F-1}, \tag{9.11}$$

for any $q \in \mathbb{C}$ and $F \in \mathbb{N}^*$, where here $\circledast$ denotes *valid* convolution (no zero-padding). Now let us define the filter $\boldsymbol{\nu}_m \in \mathbb{C}^{K+1}$ by:

$$\boldsymbol{\nu}_m = [1, -q_{m,1}] \circledast [1, -q_{m,2}] \circledast \cdots \circledast [1, -q_{m,K}], \tag{9.12}$$

where this time a *full* zero-padding is used for each convolution. Using property (9.11) and the associativity and commutativity of convolution, we see that $\boldsymbol{\nu}_m$ must be an *annihilating filter* for $\hat{\boldsymbol{h}}_m \overset{\text{def}}{=} [\hat{h}_m[f_i]]_{i=0}^{F-1}$, namely,

$$\boldsymbol{\nu}_m \circledast \hat{\boldsymbol{h}}_m = \mathbf{0}_{F-K}. \tag{9.13}$$

Moreover, if $\hat{\boldsymbol{h}}_m \in \mathbb{C}^F$ has the form (9.10) and $F \geq 2K+1$, it can be shown that this is the *only* annihilating filter of size $K+1$. Finding this filter given $\hat{\boldsymbol{h}}_m$ can be done by solving the following simple minimization problem:

$$\underset{\boldsymbol{\nu}_m,\, \|\boldsymbol{\nu}_m\|_2^2 = 1}{\operatorname{argmin}} \left\| \operatorname{Toep}(\hat{\boldsymbol{h}}_m) \boldsymbol{\nu}_m \right\|_2^2, \tag{9.14}$$

whose solution is the minimal eigenvector of $\operatorname{Toep}(\hat{\boldsymbol{h}}_m)$. Then, it can be shown that the geometric ratios $q_{m,k}$ can be recovered from $\boldsymbol{\nu}_m$ as the *roots* of the following polynomial of degree K:

$$P_{\boldsymbol{\nu}_m}(X) = \sum_{k=0}^{K} \nu_{m,k} X^k. \tag{9.15}$$

Recovering the weights $b_{m,k}$ from (9.10) then amounts to a simple linear inverse problem involving a Vandermonde matrix containing the ratios. Finally, the ratios and weights can be used to recover the desired parameters $\{a_{m,k}, \tau_{m,k}\}_{k=1}^{K}$ using (9.10).

Note that this entire process assumed that the RIRs $\{\hat{\boldsymbol{h}}_m\}_{m=1}^{M}$ were known, which is obviously not the case. This is where our contribution comes in. We propose a simple iterative scheme to jointly

[5] Note that here *echoes* include the direct path.

estimate the annihilating filters $\{\boldsymbol{\nu}_m\}_{m=1}^M$ and $\hat{\boldsymbol{s}}$ from $\{\hat{\boldsymbol{x}}_m\}_{m=1}^M$:

- **Step 0:** Initialize $\hat{\boldsymbol{z}} \in \mathbb{C}$ as a random complex Gaussian vector. Throughout this algorithm, $\hat{\boldsymbol{z}}$ is to be understood as the current estimate of the *pointwise inverse* of $\hat{\boldsymbol{s}}$ so that $\hat{\boldsymbol{x}}_m \odot \hat{\boldsymbol{z}} \approx \hat{\boldsymbol{h}}_m$ according to (9.8). Then, alternate the following two steps until convergence:
  - **Step 1:** For each m, set $\hat{\boldsymbol{h}}_m$ to $\hat{\boldsymbol{x}}_m \odot \hat{\boldsymbol{z}}$ and estimate the annihilating filter $\boldsymbol{\nu}_m$ by solving (9.14).
  - **Step 2:** Update $\hat{\boldsymbol{z}}$ by fixing $\{\boldsymbol{\nu}_m\}_{m=1}^M$ and minimizing the combined losses of (9.14):

$$\underset{\|\boldsymbol{z}\|_2^2=1}{\operatorname{argmin}} \sum_{m=1}^{M} \left\| \operatorname{Toep}(\hat{\boldsymbol{x}}_m \odot \hat{\boldsymbol{z}}) \boldsymbol{\nu}_m \right\|_2^2 = \|\mathbf{Q}\hat{\boldsymbol{z}}\|_2^2 \tag{9.16}$$

    whose solution is given by the minimal eigenvector of $\mathbf{Q}$.
- **Step 3:** For each estimated $\boldsymbol{\nu}_m$, recover the desired $\{a_{m,k}, \tau_{m,k}\}_{k=1}^K$ as explained above.

We call this algorithm MULAN for *MULtichannel ANnihilation*. Note that it can be applied to an arbitrary number of channels $M \geq 2$. The normalization $\|\boldsymbol{z}\|_2^2 = 1$ in **Step 2** is to avoid potential divergence issues due to the multiplicative scalar ambiguity between $\hat{\boldsymbol{h}}$ and $\boldsymbol{z}$. Our experiments showed that it was not necessary to strictly enforce an *anchor constraint* here (see Sec. 9.2.1). An anchor is instead set *a posteriori* for evaluation purpose, by rescaling and shifting the estimated echoes so that $\tau_{1,1} = 0$ and $a_{1,1} = 1$.

9.2.2.2 Results

We evaluate MULAN on two-channel ($M = 2$) blind acoustic echo retrieval in a simple simulated setting, and compare its performance to the two baselines of Sec. 9.2.1, namely, the vanilla *cross-relation* (CR) approach (9.6) [84] and the LASSO approach (9.7) [85]. We simulate a dataset of 100 two-channel RIRs at 16 kHz using `pyroomacoustics` [42] and the random geometrical sampling scheme of Sec. 8.1.1.2 with room dimensions between $4 \times 6 \times 8$ and $5 \times 7 \times 9$ meters. Omnidirectional devices and a constant absorption coefficient of 0.2 on all surfaces are employed. Crucially, *only first-order reflections* are included in the simulation for this experiment, allowing to set $K = 7$ throughout. Each two-channel RIR is convolved with a random speech utterance from the TIMIT dataset [91] and the resulting signals are cropped to 250 ms, *i.e.*, $N = 4000$ samples.

For MULAN, we use a frequency grid $\mathcal{F}$ of $F = 401$ regularly spaced frequencies between 200 Hz and 2000 Hz. In practice, we run the algorithm on 20 random initializations and keep the one minimizing (9.16). Iterations are stopped when (9.16) changes by less than 0.1% or when they reach 1000. For the baselines, the true length L or the RIRs is used. For LASSO, the sparsity parameter λ is manually set to $\lambda = 10^{-3}$, which empirically showed best performance among $\{10^{-6}, 10^{-5}, \ldots, 10^{2}\}$.

To evaluate the estimation of echo timings, we count a test sample as *successful* if and only if the root mean squared error (RMSE) of the $7 \times 2 = 14$ echo timings is below 1 sample. This metric essentially measures *exact* channel recovery and fully penalizes any test where one echo is missed or completely off. To evaluate the estimation of echo attenuations, we compute the RMSE on them among *successful* timing estimations only. The results are reported in Table 9.2. As can be seen, the proposed approach strongly outperforms the baselines on these metrics, retrieving all 14 echoes

Method	Successful timing estimation	RMSE on attenuations
CR (9.6) [84]	1%	0.0442
LASSO (9.7) [85]	2%	0.0346
MULAN (proposed)	**70** %	**0.00048**

TABLE 9.2: Results obtained on a synthetic dataset of 100 two-channel speech signals with $K = 7$ echoes using MULAN and the two baselines of Sec. 9.2.1.

exactly in 70 out of 100 test signals. First, the poor results obtained using general-purpose, state-of-the-art blind system identification methods shows the inherent difficulty of the task in the context of room acoustics, even under an extremely simplified model. Second, this hints at the potential of using off-the-grid techniques for acoustic echo retrieval, which had not been tried prior to MULAN.

Unfortunately, further experiments revealed than neither MULAN nor the baselines were robust enough to handle more difficult scenarios that included echoes of higher order or additive noise. This calls for the development of more robust approaches. A potential lead to improve MULAN is to use more robust techniques for annihilating filter estimation, such as Cadzow denoising [92].

9.2.3 BLASTER: Off-the-Grid Sparse Cross-Relation

Associated publication: [S21]

9.2.3.1 Method

The second proposed method also operates in the Fourier domain, and can be viewed as a continuous-time expansion of the cross-relation+LASSO approach reviewed in Sec. 9.2.1. Closely paralleling our gridless image-source recovery method presented in Sec. 7.2.1, we will show that the non-convex problem at hand can be relaxed to an infinite-dimensional convex problem of BLASSO type [59], with the same form as (7.16) but in 1D instead of 3D.

Recall that under the vanilla image source model (6.16), the impulse response from the source to microphone m takes the form of a weighted combination of Dirac impulses:

$$h_m(t) = \sum_{k=1}^{K} b_{m,k}\delta(t - \tau_{m,k}) \text{ where } b_{m,k} = \frac{a_{m,k}}{c\tau_{m,k}} > 0. \tag{9.17}$$

Observe that $h_m(\cdot)$ lives in the space $\mathcal{M}(\mathbb{R})$ of (positive) Radon measures, *i.e.*, the topological dual of the space of continuous functions on $\mathbb{R}$ that vanish at infinity [55], and is a discrete such measure (see Sec. 7.2.1.1). Besides, let us write the *cross relation* (9.5) in the discrete Fourier transform (DFT) domain. We have:

$$\hat{\boldsymbol{x}}_2 \odot \hat{\boldsymbol{h}}_1 - \hat{\boldsymbol{x}}_1 \odot \hat{\boldsymbol{h}}_2 = \mathbf{0}_F \tag{9.18}$$

where once again $\hat{\boldsymbol{h}}_m \overset{\text{def}}{=} [\hat{h}_m[f_i]]_{i=0}^{F-1}$, $\hat{\boldsymbol{x}}_m \overset{\text{def}}{=} [\hat{x}_m[f_i]]_{i=0}^{F-1}$ and $\mathcal{F} = \{f_0, \dots, f_{F-1}\}$ is a discrete set of linearly-spaced frequency as in (1.2). We denote by $\Phi_{\mathcal{F}}$ the linear operator that maps a Radon measure to its *continuous-time* Fourier transform evaluated at the frequencies in $\mathcal{F}$ such that, with

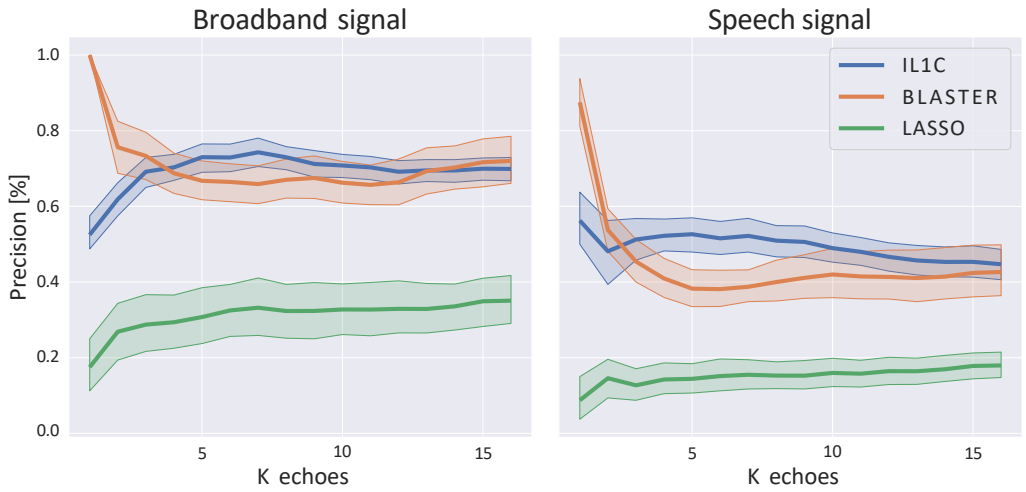


FIGURE 9.5: Comparing the precision of BLASTER (proposed), LASSO (9.7) [85] and IL1C [76] when recovering the times of arrival of K echoes over 1000 two-channel signals, using either a broadband or a speech source, and a recovery threshold of 2 samples. In this experiment, SNR=20 dB and RT_{60}=400 ms.

appropriate normalization and assuming the DFT is taken over sufficiently many points, we have

$$\hat{\boldsymbol{h}}_m = \Phi_{\mathcal{F}} h_m(\cdot). \tag{9.19}$$

We now introduce the following Radon measure on $\mathbb{R} \times \{1, 2\}$:

$$\mu \overset{\text{def}}{=} \begin{bmatrix} h_1(\cdot) \\ h_2(\cdot) \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} b_{1,k}\delta(\cdot - \tau_{1,k}) \\ \sum_{k=1}^{K} b_{2,k}\delta(\cdot - \tau_{2,k}) \end{bmatrix} \in \mathcal{M}(\mathbb{R} \times \{1, 2\}), \tag{9.20}$$

which can be interpreted as our two continuous-time RIRs vertically stacked. The left-hand side of (9.18) can now be viewed a *linear operator* applied to μ:

$$\hat{\boldsymbol{x}}_2 \odot \hat{\boldsymbol{h}}_1 - \hat{\boldsymbol{x}}_1 \odot \hat{\boldsymbol{h}}_2 = \left[\hat{\boldsymbol{x}}_2 \odot \Phi_{\mathcal{F}}, -\hat{\boldsymbol{x}}_1 \odot \Phi_{\mathcal{F}}\right] \begin{bmatrix} h_1(\cdot) \\ h_2(\cdot) \end{bmatrix} \overset{\text{def}}{=} \mathcal{A}\mu. \tag{9.21}$$

The cross relation (9.18) hence gives $\mathcal{A}\mu = \mathbf{0}_F$. To impose the *anchor constraints* $b_{1,1} = 1$ and $\tau_{1,1} = 0$ on μ (see Sec. 9.2.1), we can reparameterize μ as:

$$\mu = \begin{bmatrix} \delta \\ 0 \end{bmatrix} + \tilde{\mu} \tag{9.22}$$

where $\tilde{\mu}$ misses the first (direct-path) Dirac in the first channel, and $\tilde{\mu}_1(0) = 0$. Introducing

$$\boldsymbol{y} \overset{\text{def}}{=} -\mathcal{A} \begin{bmatrix} \delta \\ 0 \end{bmatrix}, \tag{9.23}$$

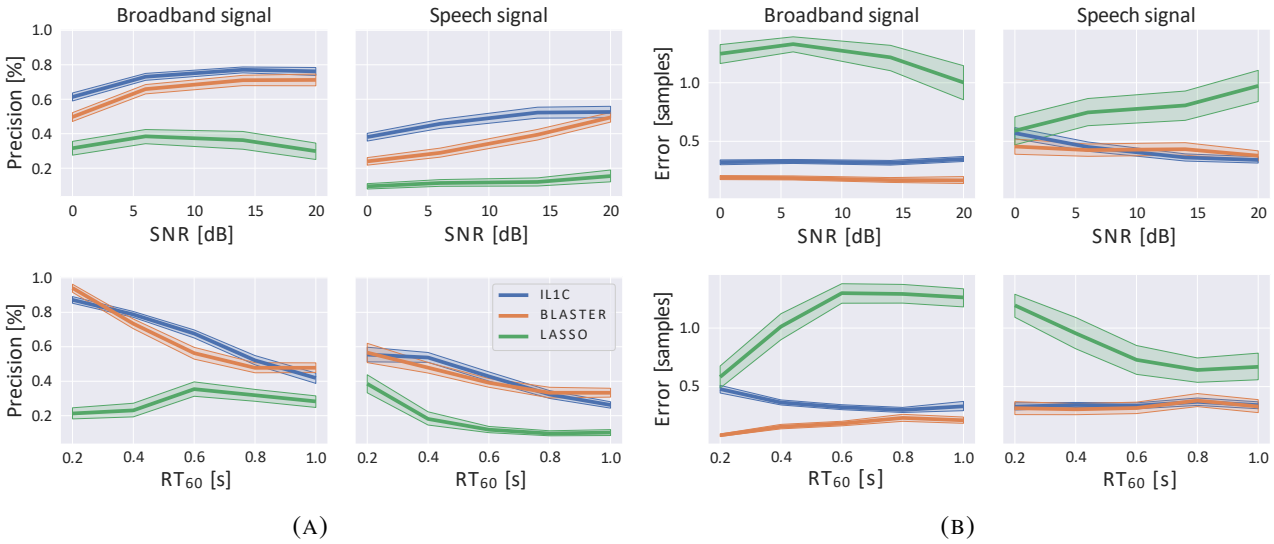


FIGURE 9.6: Comparing the precision (A) and RMSE (B) of BLASTER (proposed), LASSO (9.7) [85] and IL1C [76] when recovering $K = 7$ echoes over 1000 two-channel broadband or speech signals, using varying SNR and RT_{60} levels. Only errors below the recovery threshold of 2 samples are averaged in (B).

we are now ready to relax our original echo retrieval problem to a convex BLASSO problem over an infinite-dimensional space of Radon measures:

$$\underset{\tilde{\mu}\in\mathcal{M}(\mathbb{R}\times\{1,2\})}{\operatorname{argmin}} \quad \|\mathcal{A}\tilde{\mu} - \boldsymbol{y}\|_2^2 + \lambda\|\tilde{\mu}\|_{\mathrm{TV}} \quad \text{s.t.} \quad \tilde{\mu}_1(0) = 0. \tag{9.24}$$

As in Sec. 7.2.1, $\|\cdot\|_{\mathrm{TV}}$ denotes the total-variation norm on this space and acts as a sparsity promoter that *guarantees* that at least one minimizer of (9.24) is a weighted sum of Diracs [93]. As in Sec. 7.2.1, we particularize the sliding Frank-Wolfe algorithm proposed in [55] to the operator $\mathcal{A}$ in order to find such a sparse minimizer $\tilde{\mu}^*$ of (9.24). This minimizer is directly parameterized by $\{b_{m,k}, \tau_{m,k}\}_{m,k=1}^{2,K}$, yielding the quantities of interest. The resulting algorithm is called BLASTER for *BLind And Sparse Technique for Echo Retrieval*.

9.2.3.2 Results

We evaluate BLASTER on two-channel blind acoustic echo retrieval. Its performance is compared to the LASSO baseline (9.7) [85] reviewed in Sec. 9.2.1, as well as a more recent variant called IL1C [76]. IL1C enforces sparsity more strongly using an iteratively re-weighted ℓ_1-constraint scheme. We simulate multiple test datasets of 1000, two-channel, 16 kHz RIRs using `pyroomacoustics` [42], the random geometrical sampling scheme of Sec. 8.1.1.2, and omnidirectional devices. This time, all image sources (up to order 20) are simulated. Each two-channel RIR is convolved with either a random speech utterance from the TIMIT dataset [91] or broadband white noise. Reverberant signals are then further corrupted by additive white noise and cropped to 1 second. We generate test sets either with a fixed reverberation time (RT_{60}) of 400 ms and a signal-to-noise ratio (SNR) varying in the range $\{0, 6, 14, 20, \infty\}$ dB, or with a fixed SNR of 20 dB and RT_{60} varying in the range

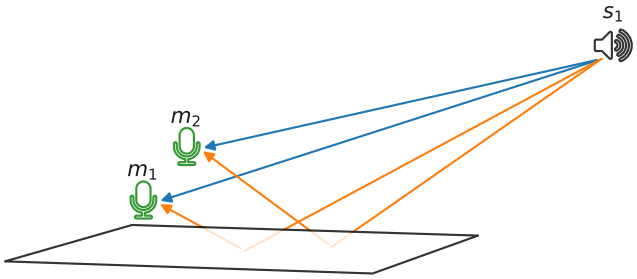


FIGURE 9.7: Example of the setup considered in Sec. 9.2.4 and 9.3.1, with one source recorded by two microphones near a reflective surface. The direct sound path (blue lines) and resulting first-order echoes (orange lines) are shown.

$\{200, 400, 600, 800, 1000\}$ ms. The RT_{60} values are reached by calculating a constant absorption coefficient for all surfaces in each room using Sabine's formula (6.28).

The three methods are evaluated on their ability to estimate the times of arrival of the K strongest echoes in each channel. A queried number of echoes K is picked in $[\![1, 16]\!]$, and each method is asked to retrieve exactly K echoes. The two metrics used are the precision[6] using a recovery threshold of 2 samples, and the root mean squared error (RMSE) in samples among successfully recovered echoes only. Results are reported in Fig. 9.5 and 9.6. As in Sec. 9.2.2.2, we observe that LASSO performs poorly on this task. Meanwhile, BLASTER performs on-par or slighlty worse than IL1C when recovering $K > 2$ echoes, but better when recovering 1 or 2 echoes. Overall, the performance of both methods is significantly worse on speech signals than on broadband signals. This interesting observation suggests that the frequency content of the source plays an important role for this task, calling for further investigation. Exploiting carefully-designed frequency sets $\mathcal{F}$ or some frequency-equalization schemes are potential leads to overcome this using BLASTER.

Zooming in on the $K = 7$ case in Fig. 9.6, and specifically looking at broadband signals, we see that both methods exhibit encouraging robustness to SNRs down to 5 dB and RT_{60} up to 400 ms, before quickly degrading. We also observe that the RMSE on successfully recovered echoes is 2-3 times lower using BLASTER than using IL1C, which could be explained by its off-the-grid nature.

9.2.4 Virtually-Supervised Blind Acoustic Echo Estimation

Associated publication: [S17]

9.2.4.1 Method

Our third and last method steps away from physics-driven approaches, and attempts to directly estimate the times of arrival of echoes using virtually-supervised learning on a carefully designed dataset. Formulating echo retrieval as a regression problem is a challenge on its own, due to the many shift, scale and permutation ambiguities inherent to the task. As a proof of concept, we hence restrict ourselves to a simple yet common *close-surface* scenario: two microphones, one source and one nearby reflective surface, as illustrated in Fig. 9.7. This may occur, for instance, when the sensors are mounted on a device placed on a table or on the floor, *e.g.*, a voice-based assistant device or a

[6] Note that precision is the same as recall in that case.

mobile robot. The reflective surface is assumed to be the most reflective and the closest one to the microphones in the environment, hence generating the strongest and earliest echo in each microphone. This removes permutation ambiguities, and we can now consider the task of estimating the timings of the direct path and of this one reflection in the two channels. To resolve the shift ambiguity, the four timings are parameterized by three quantities, illustrated in the right part of Fig. 9.8:

- the **TDOA**, *i.e.*, the time difference of arrival of the *direct path* at the two channels;
- the **iTDOA**, *i.e.*, the time difference of arrival of the *image source* at the two channels;
- the **TDOE**, *i.e.*, the time difference of arrival between direct path and first echo in one channel.

We virtually train a simple multilayer perceptron (MLP) model to estimate these three quantities blindly from a two-channel signal in the close-surface scenario. The following averaged *interaural-level-difference* (ILD) and *interaural-phase-difference* (IPD) features are used as input:

$$\begin{cases} \mathrm{ILD}_f = & \frac{1}{N}\sum_{n=1}^{N} \log\left|\frac{X_2[f,n]}{X_1[f,n]}\right| \\ \mathrm{IPD}_f = & \frac{1}{N}\sum_{n=1}^{N} \frac{X_2[f,n]/\mid X_2[f,n]\mid}{X_2[f,n]/\mid X_1[f,n]\mid}, f \neq 0 \end{cases} \tag{9.25}$$

where $\{X_m[f_i,n]\}_{i=1,n=1}^{F,N}$ denotes the short-time discrete Fourier transforms (STFT) of channel m at 16 kHz with 64 ms sliding Hann windows and 50% overlap, resulting in $F = 512$ non-negative frequency bins. Specifically, the network's input is the 1534-dimensional vector $[\mathbf{ILD}; \mathrm{Re}(\mathbf{IPD})]; \mathrm{Im}(\mathbf{IPD})]$. The network's output is the 3-dimensional vector $[\mathrm{TDOA}; \mathrm{iTDOA}; \mathrm{TDOA}]$. We use 3 fully connected hidden layers of respective size 500, 300 and 50. Rectified linear unit (ReLU) activation functions are used except at the output, and each hidden layer has a dropout probability of 0.3. We use the mean squared error loss function and the Adam optimizer [54] for training.

To build the training dataset, we first simulate 90k two-channel RIRs using `Roomsim` [41] with room sizes randomly drawn between $3\times3\times2$ and $9\times9\times4$ meters and omnidirectional devices. Random source-microphone positions and absorption coefficients are used, respecting the close-surface scenario. In particular, the microphones are at most 30 cm from the close surface, placed 10 cm from each other, the (frequency-independent) absorption coefficient of the close surface is sampled uniformly at random in $[0, 0.5]$ and the ones of the other walls in $[0.5, 1]$. Scattering coefficients as in Sec. 8.1.1.2 are used for all surfaces. As expected, the resulting RT_{60}'s are low, between 20 ms and 250 ms (see Sec. 8.1.2). An example of such RIR is shown in the left part of Fig. 9.8. Late reverberation effects are small while early effects are emphasized, which matches our scenario of interest.

Following up on the observation made in Sec. 9.2.3.2 that physics-driven echo-retrieval methods seem to be sensitive to the frequency content of the source, we design a training set to test whether this finding carries-over to the data-driven setting, despite using features that are approximately independent of the emitted signal (9.25). We also want to test the generalizability of a model trained on noiseless features. To this end, each two channel RIR is convolved with 1 second of broadband white noise signal, with no additive noise.

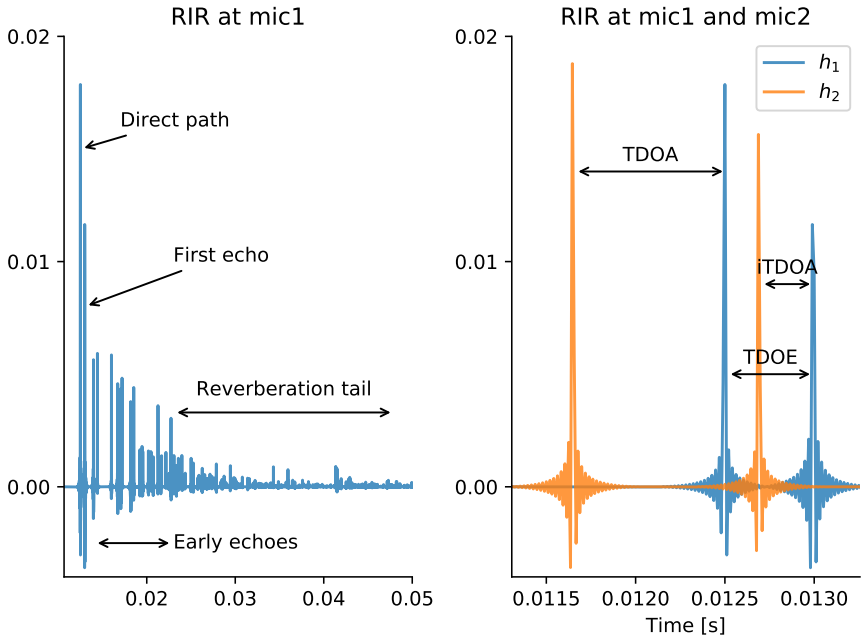


FIGURE 9.8: Left: Typical simulated RIR with annotated components. Right: Superposed graphs of a two-channel RIR annotated by the time differences of arrival between direct paths (TDOA), first echoes (iTDOA) and direct path and first echo (TDOE).

	Source signal	Additive noise (10 dB SNR)	nRMSE TDOA	nRMSE iTDOA	nRMSE TDOE
MLP (proposed)	white noise	✗	0.18	0.28	0.25
MLP (proposed)	white noise	✓	0.68	0.69	0.89
MLP (proposed)	speech	✗	0.31	0.34	0.56
MLP (proposed)	speech	✓	0.99	0.98	1.48
GCC-PHAT [94]	white noise	✗	0.21	-	-
GCC-PHAT [94]	white noise	✓	0.68	-	-
GCC-PHAT [94]	speech	✗	0.32	-	-
GCC-PHAT [94]	speech	✓	1.38	-	-

TABLE 9.3: Normalized root mean squared error (nRMSE) on echo timing estimation using the proposed virtually-supervised multi-layer preceptron (MLP) using noiseless or noisy, speech or white-noise, two-channel, simulated signals.

9.2.4.2 Results

The trained model is evaluated on the echo-timing estimation task using different test sets of 200 signals each. The source signal used is either white noise or a random speech excerpt from the TIMIT database [91], and we consider either the noiseless setting or additive white noise with an SNR of 10 dB. Table 9.3 reports the results in terms of normalized root mean squared error (nRMSE) on the three quantities. TDOA-estimation errors are compared the classical method of generalized cross-correlation with phase transform (GCC-PHAT, [94]) for reference. We first observe that for each of the 4 test sets, the two methods perform comparably on TDOA estimation. Meanwhile, the MLP errors on the three quantities are of comparable magnitudes, the TDOA ones being slightly lower.

This is expected, as this is the strongest of the 3 cues. We also reproduce the observation made in Sec. 9.2.3.2 that echo retrieval seems highly sensitive to the source content, with errors $1.5\times$ to $2\times$ larger using speech instead of white noise. Finally, the two methods are extremely sensitive to additive noise, both of them yielding unusably large relative errors at the 10 dB SNR level.

The weak performance of GCC-PHAT on this seemingly-simple single-source TDOA estimation task may be surprising at first, as it is usually considered to be a strong baseline. However, one should keep in mind that the considered scenario, though not uncommon in practice, has been very little studied, if at all, in the literature. Looking again at Fig. 9.8(left), the direct path and first echo, which are maximally-correlated signals, are extremely close to each other, which constitutes a worst-case situation for any correlation-based method, due to interference. The comparable results obtained by the MLP suggest that the network exploited similar correlation-based but noise-sensitive cues in the input feature.

Overall, the blind echo retrieval task is proven once again more challenging than one might initially expect, even in the restricted scenario of a single strong early reflection. While this scenario alleviates ambiguities inherent to the problem, the short delays between the direct and reflected paths make the task intrinsically hard, as witnessed by the relatively high errors even in the matched, broadband and noiseless setting. These results call for further investigation using data-augmentation strategies and higher-capacity architectures. This first attempt at the task using machine learning also puts forward the challenge of defining a meaningful echo retrieval task that is amenable to supervised regression, which constitutes an interesting and open research issue.

9.3 Using Echoes Beyond Room-Parameter Estimation

The goal of this last section is to investigate whether there is potential in leveraging the timings of early echoes, beyond the known application of room parameter estimation. This opens up a new research area which we refer to here as *echo-aware audio signal processing*. While multi-path signal processing is a widely studied topic since the late 50s due to its application to *rake-receivers* in wireless communication [95], very little work exists in the audio context, where reflections and reverberation are generally treated as *foes* rather than *friends*. To our knowledge, the first article attempting to bridge this gap is from 2015, proposing an acoustic beamforming framework inspired by rake receivers [16]. We present here three of our contributions to this endeavor, leveraging echoes for sound source localization (Sec. 9.3.1), speaker separation (Sec. 9.3.2) and beamforming (Sec. 9.3.3).

Since the techniques themselves are anchored in domains that are out of the scope of this thesis, the focus here is on key results, while most methodological details are omitted. Moreover, since the subtask of estimating echo timings from signals was the focus of Sec. 9.2, we investigate settings where they are readily available with high-enough precision, the main goal being to evaluate their potential usefulness in different applications. In addition, as explored at length throughout this thesis, the correspondence between echoes and room parameters is not one way. This means that conversely, knowledge on echoes can be inferred from external acoustical or geometrical information, as an alternative to signal-based estimation.

9.3.1 MIRAGE: Echo-Aware Sound Source Localization

Associated publication: [S17]

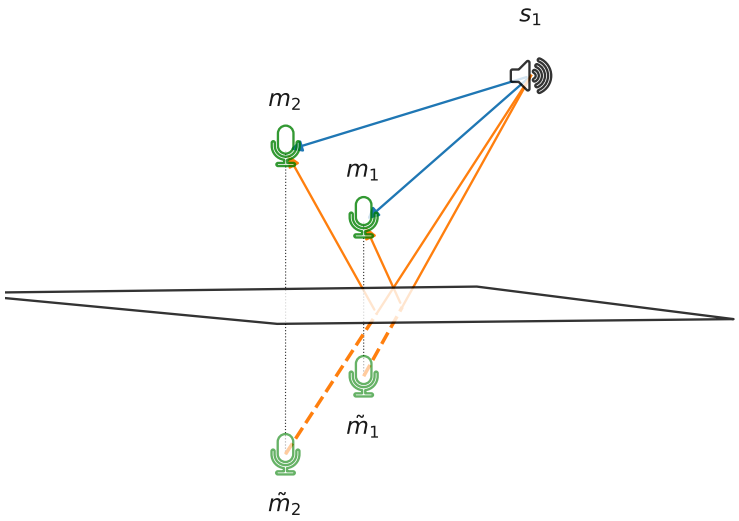


FIGURE 9.9: Illustration of MIcrophone aRray AuGmentation with Echoes (MIRAGE).

For this first example, we return to the *close-surface* scenario investigated in Sec. 9.2.4, and investigate whether the reflection timings of a nearby reflector can be exploited in the context of sound source localization. Let us start by considering a pair of microphones at a distance d and a sound source in free field. If the source is sufficiently far from the array, the sound it emits is received as an almost-planar wave by the microphone pair. Due to the obvious cylindrical symmetry around the axis formed by the pair, the source can only be localized in terms of the incidence angle between the plane wave and that axis, *i.e.*, 1D sound source localizaton. As already mentioned back in Sec. 5.2.1, there is then a one-to-one correspondence between this angle θ and the time difference of arrival (TDOA) τ of the wave at the microphones: $\tau/(cd) = \cos(\theta)$. To perform 2D localization on the sphere, however, at least a third microphone is needed. State-of-the-art physics-driven methods typically proceed by combining TDOA estimates for *multiple* microphone pairs in an array, using variants of *triangulation*. A computationally efficient implementation of this approach is SRP-PHAT [96, 97], which combines the *steered-response-power* (SRP) technique with the GCC-PHAT TDOA estimator discussed in Sec 9.2.4.2 [94].

The key idea explored in this section is illustrated in Fig. 9.9. Owing to the *reciprocity theorem* for the wave equation, a dual and strictly equivalent view to the image-source model is the *image-microphone* model, where echoes are interpreted as manifestations of an emitted source wave impinging on *virtual microphones*, corresponding to iterated geometrical reflections of the true microphones across the room boundaries. The RIR from a source to a true microphone is then the sum of all free-field RIRs from this source to the corresponding virtual microphones. From this viewpoint, a pair of microphone next to a reflective surface forms a *virtual array* of 4 microphones. If the TDOAs between the microphone pairs in this virtual array could be estimated, a method such as SRP could be directly used to achieve 2D localization, a task normally considered "impossible" using a single microphone pair. But these virtual TDOAS are *precisely* the TDOA, iTDOA and TDOE parameters estimated by the virtually-supervised multilayer perceptron (MLP) of Sec. 9.2.4, as shown in Fig. 9.8. We call this approach MIcrophone aRray AuGmentation with Echoes or *MIRAGE*.

To demonstrate the feasibility of MIRAGE, we use the SRP implementation of [97] on the TDOA, iTDOA and TDOE estimated by the MLP on the noiseless white-noise test set (Sec. 9.2.4.2). Both azimuth and elevation angles are calculated in a spherical coordinate frame placed at the barycenter of the virtual array. The metrics used are the accuracy with recovery thresholds of 10° and 20°, and

Accuracy ($< 10°$)		Accuracy ($< 20°$)	
θ	ϕ	θ	ϕ
$4.5°$ (59%)	$3.9°$ (71%)	$6.8°$ (79%)	$5.9°$ (88%)

TABLE 9.4: Mean angular errors of successful recoveries in $°$ (with accuracies in $\%$) for both azimuth (θ) and elevation (ϕ) angles with $10°$ and $20°$ thresholds, using the echo timings computed with the MLP of Sec. 9.2.4 and the SRP implementation of [97].

the mean absolute angular error across successful recoveries in degrees, calculated independently for the azimuth and the elevation. As can be seen in Table 9.4, "impossible" 2D localization is indeed achieved by the microphone pair, with both azimuth and elevation errors below $20°$ for $\approx 80\%$ of the 200 test signals.

9.3.2 Separake: Echo-Aware Sound Source Separation

Associated publication: [S13]

In this second example, we evaluate the usefulness of knowing the amplitudes and timings of $K \geq 0$ early echoes[7] in *non-blind* multichannel speaker separation. In the sound source separation literature, *non-blind* means that *some* information is assumed to be known on the RIR from each source to each microphone. The assumed information is often given in the form of *relative* transfer functions[8] (RTFs). Here, *relative* means that for each source, the transfer function to each microphone is divided by that to a reference microphone. The advantage of RTFs is that, as opposed to absolute transfer functions or RIRs, they alleviate the *shift-and-scale* ambiguity discussed in Sec. 9.2.1. Consequently, there exist methods to estimate them *blindly* from multichannel recordings, *i.e.*, without knowing the source signals [98]. Up to now, these methods fell into two categories:

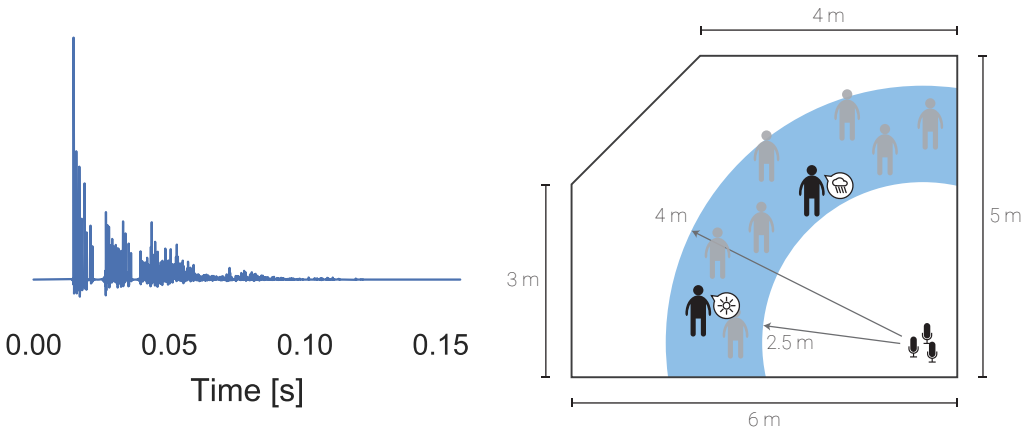
1. Assuming an *anechoic* sound propagation model ($K = 0$). In that case, RIRs only contain the direct path, and it is easily seen that estimating the RTFs is equivalent to estimating the time difference of arrival (TDOA) between each microphone and the reference one. As discussed in Sec. 9.3.1, TDOAs can either be estimated from signals or from geometrical knowledge on sources, since they are in one-to-one correspondence with their directions of arrival.

2. *Agnostic* to any sound propagation model. These methods attempt to estimate RTFs as accurately as possible purely from signals. This can be viewed as an instance of the data-driven/learning-based approach.

We now open up a third, new category, which we call ***Separake***, in reference to rake-receivers:

3. Assuming a sound propagation model involving $K \geq 1$ echoes per channel. RTFs can then be computed, given the amplitudes and timings of these echoes up to a global scale and shift (see *anchor constraints* in Sec. 9.2). Analogously to the second category, these parameters can be obtained in two ways: either blindly from signals, which is the topic of Sec. 9.2, or based on acoustical and geometrical parameters (see Chap. 6).

[7] In this section, the direct path is not counted in the number K of echoes.
[8] Recall that transfer functions are defined as the Fourier transforms of impulse reponses.



(A) (B)

FIGURE 9.10: (A) A typical simulated RIR. (B) The scenario used to evaluate *Separake*.

To assess the potential of echo-informed RTFs in non-blind multichannel speaker separation, we use a versatile framework called multichannel non-negative matrix factorization (NMF)[9] [102]. Without spelling details, multichannel NMF hinges on the following observation model in the short-term Fourier transform (STFT) domain[10]:

$$\boldsymbol{X}[f,n] = \widetilde{\mathbf{H}}[f]\boldsymbol{S}[f,n] + \boldsymbol{E}[f,n] \tag{9.26}$$

where $\boldsymbol{X}[f,n] \in \mathbb{C}^M$ contains the observed spectrograms at M microphones, $\boldsymbol{S}[f,n] \in \mathbb{C}^J$ contains the J source spectrograms to be estimated, $\widetilde{\mathbf{H}}[f] \in \mathbb{C}^{M\times J}$ contains the RTFs from the J sources to the M microphones, and $\boldsymbol{E}[f,n] \in \mathbb{C}^M$ captures noise and modeling errors. The core idea is to assume that the entries in $\boldsymbol{S}$ and $\boldsymbol{E}$ are independent zero-mean circular-symmetric complex Gaussian variables, and that the variances for each source are tied by a non-negative matrix factorization model. The model parameters are estimated by maximization of the observed-data log-likelihood, either using the expectation-maximization algorithm (EM-NMF) as in [102], or using multiplicative updates (MU-NMF) as in [103]. The framework is versatile in that the different parameters can either be fixed, learned at inference time, or pre-learned during a training phase. We consider the following variations:

- **Speaker dependent.** The identities of the speakers are known. Speaker-specific NMF dictionaries are pre-learned on an anechoic training set built from the TIMIT database [91]. The dictionaries are fixed to their respective source at inference time.
- **Universal speaker.** The identities of the speakers are unknown. A *universal-speaker* dictionary [104] is pre-learned on an anechoic training set built from the TIMIT database [91]. The dictionary is fixed and used for all sources at inference time.
- **Learned RTF.** The RTFs are learned at inference time, jointly estimated with the source spectrograms.

[9] At the time of working on [S13], NMF-based sound source separation methods were still considered state-of-the art, DNN-based methods such as [99] being in their infancy. They have since then been largely outperformed by the latter. Nonetheless, their low complexity, versatility and statistical interpretability preserved their relevance in some contexts, notably in the development of hybrid NMF-DNN approaches, e.g. [100, 101].

[10] In the experiments, 128 ms cosine windows with 50% overlap at 16 kHz are used for analysis and synthesis.

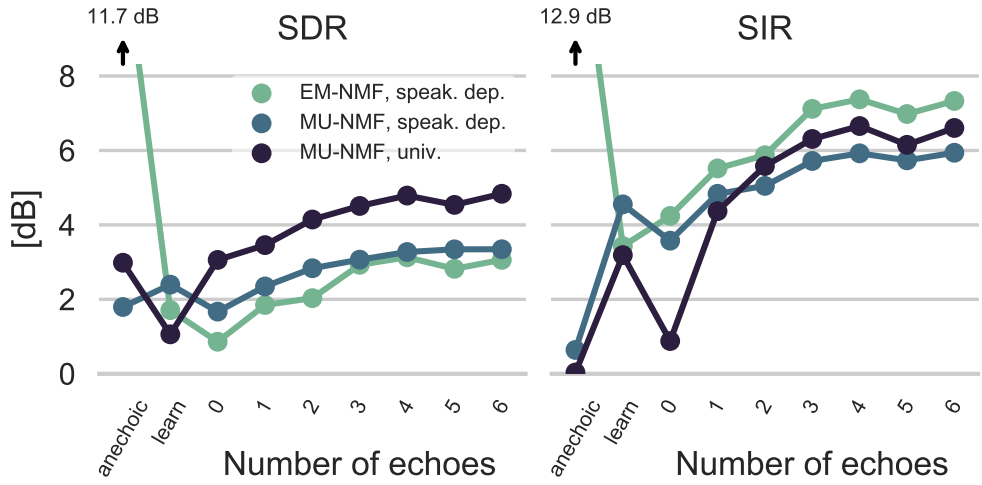


FIGURE 9.11: Median SDR and SIR for the different multichannel speaker separation algorithms and scenarios used to evaluate *Separake*.

- **Echo-aware RTF.** The RTFs are fixed and known at inference time, calculated using the vanilla image source model truncated to $K \in [\![0, 6]\!]$ echoes + direct path using `pyroomacoustics` [42]. The true parameters of the first K echoes (up to global scale and shift) are used for this purpose.

Three multichannel NMF variants under 8 different RTF models each are evaluated on simulated reverberant speech mixtures. We use `pyroomacoustics` [42] with echoes up to order 20, omnidirectional devices and the geometry depicted in Fig. 9.10. An array of $M = 3$ microphones is placed in the corner of a non-shoebox room with 5 walls. We select 40 sources at random locations at a distance ranging from 2.5 m to 4 m from the microphone array. Pairs of sources ($J = 2$) are chosen so that they are at least 1 m apart. The scenario is repeated for every two active sources out of the 780 possible pairs. The wall absorption factor is set to 0.4, leading to a RT_{60} of approximately 100 ms. The three-channel RIRs are convolved with out-of-training speech excerpts from the TIMIT database [91]. In addition, the same data are generated in anechoic conditions (absorptions set to 1) and fed to the anechoic ($K = 0$) RTF model, to examine results under a perfectly-matched RTF model.

Performance is evaluated in terms of the standard signal-to-distortion ratio (SDR) and source-to-interference ratio (SIR) metrics, as defined in [105], which are computed using the `mir_eval` toolbox [106]. The median SDR and SIR for all 3×9 variants are shown in Fig. 9.11. We observe that introducing echoes in RTF models progressively and significantly improves performance of all 3 source-separation approaches, validating the potential of echo-awareness. The *Separake* approach even outperforms the learned RTF models with only 1 or 2 echoes. With up to six echoes, gains are +[1,2] dB SDR and +[2,7] dB SIR compared to the learned model, depending on the considered variant. It is interesting to note that in all experiments, the first three echoes nearly saturate the metrics. This is good news, since higher order echoes are harder to estimate (Sec. 9.2).

In the anechoic setting with a perfectly-matched RTF model, EM-NMF has near-perfect performance. This is in fact expected because the non-blind separation problem is then over-determined, due to $M > J$. However, the same is not true for the MU-NMF variant, despite its better performance otherwise. This could be explained by its use of an ℓ_1-enforced sparse prior in the NMF weights, pushing the solution away from the global optimum in over-determined conditions, but acting as

Acronym	RTF model	Noise Model
Delay-and-sum (DS) [112]	Anechoic ($K = 0$)	Spatially white n.
MVDR-DP [112]	Anechoic ($K = 0$)	Diffuse n.
MVDR-ReTF [113]	Agnostic/Learned	Diffuse n.
MVDR-RAKE [16]	Echo-aware ($K = 3$)	Diffuse n.
MVDR-DP-Late [114]	Anechoic ($K = 0$)	Spatially white n. + lr.
MVDR-ReTF-Late [108]	Agnostic/learned	Diffuse n. + lr.
MVDR-RAKE-Late [114]	Echo-aware ($K = 3$)	Diffuse n. + lr.

TABLE 9.5: Summary of the considered beamformers. "n." and "lr." are used as shorthand for noise and late reverberation while RAKE denotes *echo-aware* beamformers.

a beneficial regularizer otherwise. Using this variant, it is remarkable that performance are actually *better* in reverberant, echo-aware conditions than in anechoic conditions. While surprising at first, this is plausibly explained by the *virtual-microphone* viewpoint used in MIRAGE (Sec. 9.3.1). Knowing the relative echo timings can augment the spatial resolution of the array by virtually increasing its size, improving its ability to separate nearby sources. Although further experiments are needed to confirm this finding, it could reveal another instance of echoes being *friends* rather than *foes*.

9.3.3 Echo-Aware Beamforming

Associated publication: [S24]

We now evaluate whether echo-awareness may help beamforming. We consider the same STFT observation model as for *Separake* (9.26), but with a single source:

$$\boldsymbol{X}[f,n] = \widetilde{\boldsymbol{h}}[f]S[f,n] + \boldsymbol{E}[f,n] \tag{9.27}$$

where $\widetilde{\boldsymbol{h}}[f] \in \mathbb{C}^M$ contains the (early) RTFs and the noise term $\boldsymbol{E}[f,n]$ captures both background noise and late reverberation. Beamforming is one of the most widely used techniques for enhancing multichannel microphone recordings, the literature on this topic spanning several decades [107]. In a nutshell, in the frequency domain, the goal of beamforming is to estimate a set of coefficients $\boldsymbol{w}[f] \in \mathbb{C}^M$ that are applied to $\boldsymbol{X}[f,n]$, such that $\boldsymbol{w}[f]^{\mathrm{H}}\boldsymbol{X}[f,n] \approx S[f,n]$. We consider eight variants of minimum-variance distortionless response (MVDR) beamformers [107] that combine different RTF and noise models, as summarized in Table 9.5. The three categories of RTF models defined in Sec 9.3.2 are represented again here: *anechoic* (marked "DP" for direct path), *agnostic/learned* (Marked "ReTF", where the approach in [108] is used for estimation), and *echo-aware* with $K = 3$ echoes + direct path per channel (marked RAKE). In addition, different statistical noise models are considered: spatially white, diffuse (*i.e.* the *Capon* filter), or diffuse *plus* late reverberation [109]. For the latter, the late-reverberation statistics are modeled by a spatial coherence matrix [110] weighted by the late reverberation power, which is estimated off-line using the procedure described in [111].

The performance of the designs are compared on the task of enhancing a target speech signal in a 5-channel mixture using the linear microphone arrays in the dEchorate dataset (see Sec. 9.1.1). They are tested in scenarios featuring high reverberation and diffuse babble noise, appropriately scaled to pre-defined signal-to-noise ratios (SNR) $\in \{0, 10, 20\}$ dB. Using the dEchorate data, we consider the room configuration `011111` ($\mathrm{RT}_{60} \approx 730$ ms) and all possible combinations of target-source and

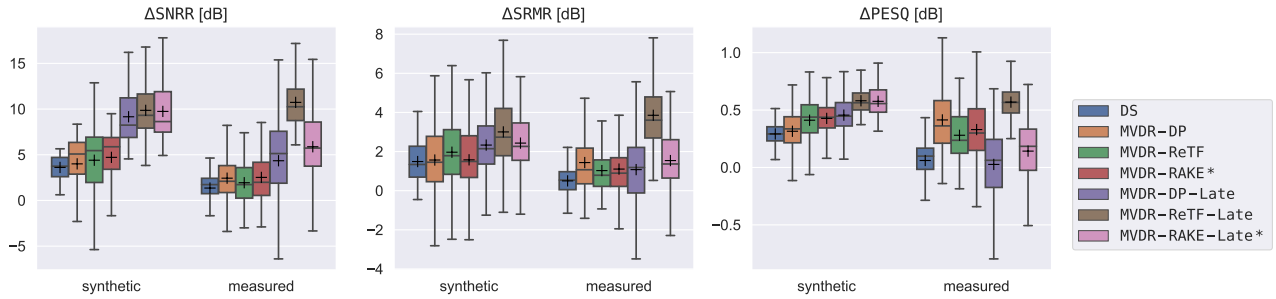


FIGURE 9.12: Box plots comparing different echo-agnostic and echo-aware (*) beamformer designs (Table 9.5) on both measured and simulated data corresponding to all source-array combinations in room `011111` ($\mathrm{RT}_{60} \approx 730$ ms) of the dEchorate dataset (Sec. 9.1).

array positions. Both real and corresponding simulated RIRs are used, which are then convolved with anechoic utterances from the wall street journal corpus [115] and corrupted by recorded diffuse babble noise. For the dEchorate RIRs, the annotated echo times of arrival are used (see Sec. 9.1.2). The simulated RIRs are computed using `pyroomacoustics` [42] under the *vanilla* image source model (6.26), and the true echo parameters are used to test echo-aware beamforming in the absence of model mismatch. The evaluation is carried out similarly to the one in [114], with the following metrics:

- the signal-to-noise plus reverberation-ratio improvement (ΔSNRR) in dB, computed as the difference between the input SNRR at the reference microphone and the SNRR at the beamformed output;
- the speech-to-reverberation energy-modulation ratio improvement (ΔSRMR) in dB [116] to evaluate dereverberation;
- the perceptual evaluation of speech quality improvement (ΔPESQ) score [117] to assess the perceptual quality of the signal and indirectly the amount of artifacts.

Both ΔSNRR and ΔPESQ are relative metrics, meaning they require a target reference signal. Here we consider the target to be the dry source signal convolved with the early part of the RIR (up to the K-th echo) at the reference (first) microphone. On the one hand, this choice numerically penalizes both anechoic and agnostic beamformers, since they respectively aim at extracting the direct-path signal and the full reverberant signal in the reference microphone. On the other hand, considering only the direct path or the full reverberant signal would be equally unfair for the echo-aware beamformers. Moreover, including early echoes in the target signal is perceptually motivated since they are known to contribute to speech intelligibility [118]. The late-reverberant signal used to compute the SNRR is the dry source signal convolved with the late part of the RIR, assumed here to start 70 ms after the direct path's time of arrival. This value corresponds to the average timing of second-order echoes, and was found to make the late reverberation well-approximated by the late diffusion model of [110].

Numerical results are reported in Fig. 9.12. On simulated data, one can see that performance consistently improves as more and more information is used on either RTF or noise models. Including late reverberation statistics considerably boosts performance in all cases. Both learned (ReTF) and echo-aware (RAKE) RTF models significantly outperform the simple anechoic design. While the two

designs perform comparably in terms of ΔSNRR and ΔPESQ, the learned model has a slight edge over the latter in terms of median ΔSRMR, which contrasts with results obtained in the *Separake* scenario of Sec. 9.3.2. A possible explanation is that the off-line RTF estimation method [108] tends to consider the stronger and more stable components of the RTF, which in the considered scenario may identify with early reverberation. Moreover, since it is not constrained by a fixed echo model, it can capture more information such as frequency-dependent attenuation coefficients. Finally, one should consider the much smaller number of parameters in the $K = 3$ echo-aware model with respect to the learned one (8 vs. several hundreds), which should in-principle be favorable in the context of on-line estimation.

When it comes to measured RIRs, however, the errors in echo timings due to annotation mismatch and the richness of real acoustic propagation seems to lead to a drop in performance for explicit echo-aware methods. While the anechoic model still performs worst across all metrics, combining an *a-priori* learned RTF model with a late-reverberation model (ReTF-Late) clearly outperforms the echo-aware approaches, which sometimes even yield a significant portion of negative metrics. As already observed in [114], this is probably due to tiny echo-timing annotation mismatch in dEchorate, as well as the fact that their frequency-dependent amplitudes are not modeled. This strong sensitivity of echo-aware methods to geometrical and timing errors mirrors the one observed in Sec. 7.1.2 of this thesis in the context of wall absorption estimation, further reinforcing the point that precise off-the-grid methods to correct timings based on signals are needed, such as the ones explored in both Chap. 7 and Sec. 9.2 of this thesis.

9.4 Conclusion on Echo-Aware Audio Signal Processing

While the primary targets of Chap. 7 and 8 were acoustical and geometrical parameters directly attached to the walls of a room, this chapter focused on their signal-domain manifestation at the receivers, namely, *echoes*. First, as a mean to foster research on echoes in the audio signal processing community, we presented dEchorate, a unique dataset of RIR measurements under variable acoustic conditions, annotated with early echo timings in a geometrically-consistent way. This effort notably yielded the *Skyline* (Fig. 9.3), a visually-pleasing representation of this thesis' overarching topic: "hearing the walls of a room".

We then proposed three methods to estimate the timings and/or amplitudes of early echoes from two-channel recordings of an unknown source, two of which are physics-driven, and the last one data-driven. The approaches distinguish themselves from prior work in the field by estimating timings "off-the-grid". Experiments on simulated data revealed that estimating the timings of up to 6 early reflections with very high, sub-sample accuracy was possible with such methods. They also revealed the high sensitivity of both physics-driven and data-driven approaches to model mismatch. This suggests one more time that further increasing the realism of the underlying forward models is necessary for real-world applicability. Another interesting observation calling for further research is the sensitivity of the methods to the frequency content of the source, speech signals proving significantly more challenging than broadband signals on this task.

Finally, we investigated whether knowing echo parameters could benefit methods in the broader field of audio signal processing, beyond their conventional application to room parameter estimation. We proposed the microphone-array augmentation with echoes (MIRAGE) framework, built on the dual *image-microphone* viewpoint to the image-source model, and used it to perform an "impossible" 2D sound source localization task using only two receivers near a reflective surface. We then

presented the *Separake* approach, *i.e.*, explicitly modeling early echoes (assumed known up to global scale and shift) in relative transfer functions to improve multichannel source separation. Applying this idea to different existing speaker separation techniques, we demonstrated on simulations that using an increasing number of early echoes (up to 3 per channel), could provide a boost in performance compared to anechoic or echo-agnostic (learned) models. We then applied the same idea to beamforming, where on simulations, echo-aware models showed to outperform anechoic ones while being on-par with *a priori* learned models, using a fraction of the number of parameters. However, testing echo-aware beamforming on real echo-annotated RIRs from the dEchorate dataset revealed that the approach was highly sensitive to even small mismatches in echo timings.

Overall, these combined findings allow one to conclude that (i) echo parameters form a potentially useful intermediate representation of "walls", being both amenable to blind estimation *and* usable in a broader audio signal processing context; (ii) subsample accuracy for up to 6 early-echo timings is achievable using both physics-driven and data-driven off-the-grid approaches, with as little as two channels; (iii) existing echo-estimation and echo-aware methods are too sensitive to model-mismatch to be applicable to real signals as of yet, calling for the development of more robust inverse methods built on more realistic forward models.

10 Conclusion

10.1 Summary and Outlook

Throughout this thesis, we formalized and explored different facets of the question *"Can one hear the walls of a room?"*, introducing a number of physics- and data-driven inverse methods that can estimate or exploit the geometrical and acoustical characteristics of sound reflectors in a room using acoustic measurements. After this synthesis of 12 articles disseminated along 7 years of research, what partial answers can we give to our original question?

Well, if one is allowed to *look* at a set of room impulse responses, our skyline visualization of the dEchorate dataset certainly suggests that one can *see* the walls of a room (Fig. 9.3, Sec 9.1)! Alternatively, from the combined results of Sec. 7.2, it is tempting to answer that "yes, one can hear the walls of a room *IF* it is a perfect empty shoebox made of perfectly reflective walls within which a known source signal is measured with a large enough microphone array using perfectly omnidirectional calibrated devices." Under such assumptions, the methods in Sec. 9.2 even suggest that the timings of early echoes from the walls could be retrieved from a single microphone pair attending an *unknown* source. But this is, arguably, a big *IF*. While a long road full of obstacles seems to lie ahead between this drastically idealized setting and reality, we delineated and undertook plausible independent paths to overcome some of these barriers, from different angles.

Among *physics-driven* inverse methods, the one using our most realistic source-receiver-wall model (Sec. 7.1.2) suggests that retrieving wall parameters is within reach as long as initial geometrical knowledge sufficiently-close to the truth is given. Coming from the opposite direction, the absorption-profile estimation technique of Sec. 7.1.1 illustrates that a much coarser model can be useful, when inverted using a more robust technique. Meanwhile, using the different paradigm of implicitly building inverse regression models by *virtually supervised learning* on simulated data, we were able to identify a number of approachable intermediate tasks. These include mapping a single RIR to area-weighted wall absorption profiles (Sec. 8.2.1); mapping a single RIR to individual absorption profiles under a fixed geometry (Sec. 8.2.2); mapping a set of two-channel noisy speech recordings to the room's volume, total surface area and reverberation time (Sec. 8.3); and mapping a two-channel broadband-source recording to the times of arrival of a reflection caused by a nearby surface (Sec. 9.2.4). Encouragingly, the results in Sec. 8.2.1 and 8.3 demonstrated the generalizability of this approach to real data, with increasing success using increasingly-realistic models for the source, receivers and walls.

Finally, the last contributions synthesized in Sec. 9.3 could be viewed as a broadening of our initial question to whether one *should* hear the walls of a room, if what one only cares about are sound sources inside the room. Preliminary results obtained on simulations encouragingly suggest that knowing the timings of as little as one to three echoes may benefit sound source localization, boost multichannel speaker separation, or improve beamforming capacity, at a low parametric cost. These findings are however yet to successfully generalize to real data, due to their sensitivity to model-mismatch. This brings further support to the relevance of further developing highly-accurate, off-the-grid reflector identification techniques, such as as the ones studied in Chap. 7 and Sec. 9.2.

10.2 Directions for Future Research

An important limitation of all the physics-driven inverse methods presented in this thesis (Chap. 7 and Sec. 9.2, 9.3) is their high-sensitivity to *model mismatch*[1], which has so far resisted attempts to successfully apply them to real data. This calls for further increasing the realism of the models and the robustness of the methods. Meanwhile, although encouraging real-data generalizability has been achieved using virtually-supervised data-driven models (Sec. 8.2.1, 8.3), further investigation is needed to understand the joint impact of surface scattering, directivity and noise on these results. Moreover, known limitations of the image source model will need to be addressed to obtain satisfying performance below the Schroeder frequency barrier (500 Hz and below). In this last section, we outline a number of identified research directions that could help bridging this gap, and bring us closer to our goal of "hearing the walls of a room" using real-world acoustic measurements.

10.2.1 Geometry-Corrected Wall Impulse Response Estimation

A natural direction for future research, building on the results of Chap. 7, would be to combine the delay-correcting wall impulse response (WIR) estimator of Sec. 7.1.2 with the gridless image-source retrieval method of Sec. 7.2.1, since each have complementary strengths and weaknesses. On the one hand, the former uses our most realistic source-receiver-wall model, but crucially relies on accurate-enough geometrical knowledge and can only successfully disentangle a limited number of reflections. On the other hand, the latter can recover hundreds of image source locations with near-exact accuracy, but crucially relies on the "vanilla" image source model. The two ideas could be combined by correcting geometrical errors using "sliding steps" on image sources in continuous 3D-space as in Sec. 7.2.1.2, instead of correcting echo timings independently inside each RIR. This would drastically reduce the dimension of the search space. The geometry-correction step and WIR estimation step could then be interleaved and iterated to further reduce errors and retrieve more reflections, even when initializing further from the true geometry. One challenge is that the degrees of freedom added by the use of *directive* sources would require optimizing the image-source positions in 5D space. This should however remain approachable by gradient descent, starting from the original geometrical estimate, combined with greedy optimization over the image source point-cloud. The recovery of higher-order echoes could also be used to further refine WIR estimates by leveraging the multi-linear constraints imposed by multi-wall reflection paths, *i.e.*, (7.2).

10.2.2 Hearing the Shape of a Polytope Room

An interesting problem is to generalize the room-geometry estimation method of Sec. 7.2 to non-shoebox, polytope rooms. Assuming that scattering effects induced by non-rectangular edges and corners can be neglected (which should hold, at least at high-enough frequencies), the first issue is to upgrade the image-source localization task to handle *partial occlusions*, *i.e.*, image sources that are only audible by a subset of microphones. One way to fight combinatorial explosion when identifying occlusions would be to use an ℓ_1- instead of ℓ_2-loss, which would be more tolerant to spurious echoes in RIRs, combined with a thresholding step. To alleviate the increased difficulty of the inverse problem, one could extend the approach of Sec. 7.2.2 to leverage *multiple* (unknown) source-receiver

[1]This is in fact a well-known limitation of inverse methods in general, and surprisingly little is known on how to generally overcome this issue. This is typically addressed using domain-specific, ad-hoc techniques in the literature.

placements in the room. This would imply solving an interesting point-cloud alignment problem, where Procrustes analysis could be used to align receiver-dependent point clouds, while more general optimal transport schemes could be used to align source-dependent point clouds. Finally, assuming a sufficiently precise image-source point cloud can be recovered in this way, inferring a polytope shape from it forms an original geometrical problem which has not been studied before (to the best of our knowledge), even under simplifying constraints. Advanced mathematical tools from tiling theory or higher-order Fourier analysis could be potentially useful to exploit the symmetries and periodic patterns in the point cloud.

10.2.3 Bridging the Real-to-Simulation Gap using Diffusion Models

Even assuming the problems posed in Sec. 10.2.1 and 10.2.2 can be fully solved, we would likely reach the limit of what can be achieved using explicit physics-driven inverse methods. Handling more complex boundary features, including curvature, or the presence of arbitrary complex objects in the room seems out of reach. Given the sensitivity of these methods to model-mismatch, does it mean the gap to real data will remain hopelessly open? Not necessarily, if we simultaneously attempt to bridge it from the other side! One attractive research direction would be to leverage the recent success of *diffusion-based generative models* in transporting signals from one distribution (say, white noise) to another (say, natural images) [119]. In our case, a diffusion model could be trained to transform a real/realistic RIR to one following a simpler ISM model, while preserving its key geometrical and acoustical parameters. A training set consisting of (realistic, simple) RIRs simulated using two different forward models under matched parameters would be used to train a *real2sim* diffusion model. The remarkable ability of these models to generalize to out-of-distribution samples in practice, due to their inherent stochastic nature, gives hope that this approach could finally close the gap, transforming real measurements into signals that are amenable to sufficiently-robust physics-driven inverse methods. A promising candidate for this could be the absorption-profile estimation technique of Sec. 7.1.1, because it operates in the magnitude Fourier domain and makes use of a robust optimization scheme, both of which could be forgiving to potential artifacts created by the diffusion model.

10.2.4 Data Augmentation with Sim-to-Real Diffusion Models

In the opposite direction from Sec. 10.2.3, diffusion-based generative models could be used to *improve the realism* and diversity of training sets for virtually-supervised learning. First, a high-end wave-based simulator could be used to obtain one-to-many matched pairs of (simplistic, realistic) RIRs in order to train a *sim2real* diffusion model. For example, furniture or complex geometrical features could be added to shoebox rooms. Then, the sim2real model could be applied to a simplistic simulated dataset to vastly expand its diversity at a lower computational cost than by using the high-end simulator all the way. The data could be generated on-the-fly, to train an inverse model on an essentially "infinite" set. The benefit of starting from a dataset of simplistic "anchor" rooms is that their acoustical and geometrical parameters (including device positions) can be represented in a compressed form, allowing to keep full control over the distribution of these parameters while using them as labels if needed.

10.2.5 Geometric Conditioning of Virtually-Supervised Models.

In this thesis, the two virtually-supervised models that achieved best generalization performance were concerned with the estimation of *global* geometrical and acoustical parameters, *i.e.*, attached to the entire room, without depending on the detailed room-and-device geometry (Sec. 8.2.1 and 8.3). In contrast, the models in Sec. 8.2.2 and Sec. 9.2.4, targeting the properties of one or multiple *individual* reflectors, were trained on datasets obeying strong geometrical constraints, with no hope of generalizing beyond them. Overcoming this limit will require the development of neural network architecture that can be *conditioned* by any geometry, on top of input signals. In principle, such conditioning can be achieved by general-purpose modifications of their layers, such as *feature-wise linear modulation* (FiLM, [120]). While this may work out-of-the-shelf for simple geometrical information such as, *e.g.*, the 3D positions of one source and one microphone, properly encoding more advanced geometrical knowledge such as the internal geometry of a varying number of microphone arrays or arbitrary room boundaries could be trickier. Adequate tools respecting the inherent spatial symmetries at hand could include *graph neural networks* [121] or parameterized *signed distance functions* [122].

10.2.6 Learning with Little-to-No Labels.

The data-driven methods presented in this thesis only used *supervised machine learning*, which requires labeled examples. To tap into the vast potential of unlabeled measured acoustic data, a possible approach would be to make use of *self-supervised learning*. In our context, one could for example leverage *contrastive losses* to learn useful RIR representations that are invariant to, *e.g.*, the positions or the responses of the devices. Subsequently training a virtually-supervised, geometry-conditioned model on those features could plausibly make it more robust to geometrical errors or mismatched devices. Another approach to compensate the lack of labels would be to supervise models not only on data, but also by imposing physical constraints on the learned neural-network functions, by means of partial differential equations. This is known under the name of *physics-informed neural networks* (PINNs) [123], an increasingly-active research area which departs from the data-driven paradigm. For example, this idea could be beneficial to the diffusion-based approaches proposed in Sec. 10.2.4 and 10.2.3, by ensuring that generated data are physically valid. Alternatively, room parameter estimation could be achieved by pre-training a PINN to solve a room-acoustic forward model that overcomes some of the limitations of the image-source method, *e.g.*, handling arbitrary boundary impedance and low-frequency effects. The PINN model could then be inverted by minimizing a loss between an observed signal and its output with respect to the input parameters of interest.

10.2.7 New Measurements for Room Acoustic Analysis

The findings of both Chap. 7 and 8 revealed the crucial impact of the directivity of devices in room parameter estimation tasks. While the results in Sec. 8.3.3 suggest that including multiple directivity profiles at train time may be sufficient to build robustness against this effect for *global* room parameter estimation, directivities must be explicitly known when estimating the parameters of *individual* walls, because of the ambiguity created by the convolution of source, wall and receiver responses inside early echoes, *i.e.*, (6.27). Unfortunately, datasets containing measured room impulse responses together with the directivity profiles of involved devices currently do not exist, partially because of the difficulty of acquiring such data, and partially because the importance of this effect has only be

brought recently to the attention of the community. Acquiring a large, diverse and geometrically-annotated RIR dataset using source-microphone pairs whose directivity profiles have been measured beforehand would represent an invaluable research contribution to the field.

10.2.8 Hearing the What?

Following the standards of room and building acoustics, this thesis has used *absorption coefficients* as the main acoustic parameter of interest pertaining to walls. However, this may be called into questions in the context of *in-situ* room acoustic diagnosis. Indeed, a standard approach used by manufacturers to measure absorption coefficients is to place an isolated piece of the considered material in a reverberant chamber, and to use reverberation-theory formulas to obtain averaged values over incidence angles from the diffuse field (Sec. 6.4). In this lab setting, scattering effects, directive effects, as well as the potential effect of mounting the material on other structures cannot be accounted for. More broadly, to the best of our knowledge, theory is still lacking to fully understand the precise intertwining between absorption, transmission and scattering effects on the one hand, and general impedance boundary conditions on the other hand, and their combined effects on reverberation fields. Integrating these factors into well-established room and building acoustic standards would realistically require an inter-disciplinary dialogue that goes beyond the sheer development of techniques. In the meantime, the retrieval of *wall impulse responses* (WIRs) from RIRs, as investigated in Sec. 7.1.2 and prior to that in [21], may constitute an interesting alternative research avenue, which is currently still at its infancy. Indeed, WIRs summarize all the acoustic effects reflectors may have on impinging waves, with the caveat of being potentially dependent on the source and receiver placement in the room. The strength of this dependency in practical room acoustic conditions is currently unknown. Continuing efforts to develop *in-situ* wall-impulse-response estimators and using them to gather substantial datasets of wall acoustic characteristics in real-world conditions could help quantifying these effects, and further facilitate future research towards "hearing the walls of a room".

Part III

Other Contributions in Research Snippets

11 Data-Driven Sound Source Localization

Associated publications: [S2, S4, S8, S9, S34]

Context

The task of localizing a sound source using microphone signals has likely been studied for as long as microphones exist. It can either be used directly in applications such as video-conferencing or interactive robotics (Snippet 15), or indirectly as a pre-processing step for multichannel audio processing (Sec 9.3.2 and 9.3.3). The main approach, still in use today, is physics-driven and recast the problem as that of estimating time differences of arrival at microphone pairs (Sec. 9.3.1). In the early 2010s, data-driven approaches began to emerge, initially using locally-linear regression models on real recorded data [124, 125]. Since then, the deep learning wave has taken over and hundreds of models have been published over the 2019-2021 period only [126]. A common theme in the field is that of crafting appropriate input features to feed data-driven models, which should ideally be as independent as possible from the sound source signal while preserving relevant spatial information. The most commonly used ones are so-called *interaural level and phase differences*, which can be computed from the relative transfer function (RTF) at a pair of microphones (Sec. 9.3.1), and were already identified as relevant for human audition by Lord Raighley in 1907 [127].

Contributions[1]

We proposed two new sound-source localization (SSL) features. The first one, called *rectified binaural ratio* (RBF, [S4]), is a statistically-principled way to robustly compute two-channel RTFs assuming a Gaussian model on each signal. It involves a natural generalization of the *t-distribution* to complex values, and was shown to be more robust than averaged ratios such as (9.25), in particular for sources with sparse spectrograms. The second one generalizes the RTF to more than one source and two microphones, in the sense that it is independent of source signals while preserving spatial information [S2]. It involves computing the *Plücker embedding* of the linear subspace on which the multichannel signal lies, at individual frequencies of its short-time Fourier transform. It yielded good SSL performance on real data when used with a data-driven, 3-source, 4-microphone model based on nearest-neighbor search under mild noise, but showed poor robustness to higher noise levels.

We also contributed three studies on virtually-supervised SSL [S8, S9, S34]. The first two use a locally-linear regression model and are precursors of the works in Chap. 8 and Sec. 9.3.1. Both perform 3D binaural SSL by leveraging reverberation cues, one in arbitrary shoebox rooms [S8], the other one under a fixed room-receiver geometry by jointly estimating mean wall-absorption coefficients [S9] and leveraging multi-task learning. They yielded satisfying results on out-of-training simulated data. Finally, [S34] studies the impact of simulation realism on the real-world generalizability of virtually-supervised 2D SSL, in the same fashion as Sec. 8.3.3, using a state-of-the-art model [128]. Likewise, both increasing the realism of devices (Sec. 8.1.3) and walls (Sec. 8.1.2) at train-time were found to positively contribute to performance on three test sets involving human speakers in real rooms.

[1] I initiated, led and co-supervised these contributions. The methods and experiments were implemented by me in [S2, S4] and by the master and PhD students involved in [S8, S9, S34].

12 Phase Retrieval in Audio

Associated publications: [S3, S6, S7, S32]

Context

Phase retrieval is a long-standing and broadly-studied research topic, that consists in retrieving a signal given a linear observation from which the phases are missing. Mathematically, one wants to retrieve $\boldsymbol{x} \in \mathbb{C}^K$ given $\boldsymbol{b} = |\mathbf{A}\boldsymbol{x}| \in \mathbb{R}_+^M$ and $\mathbf{A} \in \mathbb{C}^{M\times K}$, where the complex modulus is taken element-wise. A vast set of algorithms have been developed to tackle this problem since the 1970s [129, 130], including recent convex relaxations (*PhaseCut*, [131]) or non-convex optimization schemes [132]. It has applications in fields such as adaptative optics and X-ray crystallography, where the phases of the Fourier transform are intrinsically lost during measurement. In audio, signals are directly observed in the time-domain, and their phases are hence available. However, it is well known that natural sounds including speech or music typically have *magnitude* spectrograms that are amenable to simple, low-parametric models (*e.g.*, NMF, see Sec. 9.3.2), whereas their *phase* spectograms feature more intricate, non-linear, long-range dependencies. Phase estimation is hence relevant to complement techniques operating in the magnitude-spectrogram domain, forming a research area that has been referred to as *phase-aware audio signal processing* [133]. Despite this, the most recent methodological developments in phase retrieval had been largely unnoticed by this community.

Contributions[1]

In [S3] and [S6], we established a bridge between magnitude-informed multichannel sound source separation and phase retrieval, by showing that the problem of estimating $\boldsymbol{x} \in \mathbb{C}^K$ given $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} \in \mathbb{C}^M$, $\mathbf{A} \in \mathbb{C}^{M\times K}$ and $\boldsymbol{b} = |\boldsymbol{x}| \in \mathbb{R}_+^K$, which we coined *phase unmixing*, was closely related. In [S6], we built on this connection to propose a convex relaxation of the problem inspired by PhaseCut. The approach outperformed conventional Wiener filtering on synthetic under-determined ($M < K$) noisy speech mixtures, even yielding exact solutions in the noiseless case. In [S3], a coordinate-descent scheme for phase unmixing combined with K-SVD dictionary-learning [134] was successfully applied to multichannel sound source separation on real audio recordings. In [S7], we developed an extension of PhaseCut allowing for *Von Mises* probablisitic priors on phases, with potential application to under-water sound-source localization [135]. Finally, in [S32], we proposed an alternate-minimization and a convex-relaxation scheme for phase unmixing when $\boldsymbol{y}$ is exact and $\boldsymbol{b}$ is noisy. The combined schemes were shown to outperform a state-of-the-art, sparsity-based audio inpainting method [136] when the Fourier-magnitude of the target signal was assumed known with higher than 10 dB SNR, achieving exact recovery in the noiseless case.

[1] I initiated and implemented the work in [S3] during my postdoc. I initiated, led and implemented the work in [S6]. The work in [S7] was equally shared with my co-author. I initiated, led, and co-supervised the work in [S32], which was implemented by the student and postdoc involved.

13 Blind Audio Source Separation: Probabilistic Models

Associated publications: [S10, S11, S18]

Context

Prior to the advent of end-to-end deep-learning-based methods that started dominating the state of the art around 2019 [137, 71], a widely used probabilistic model for maximum-likelihood-based audio processing, and in particular sound source separation, was the so-called *local-Gaussian* model. It consists in modeling the entries of the short-time Fourier transform (STFT) of audio signals as independent, circular-symmetric complex Gaussians, *e.g.* [102] (see Sec. 9.3.2). The combined simplicity and versatility of this model made it extremely successful in the pre-deep-learning era, and it is still in use today, in combination with deep variational auto-encoders (VAEs), *e.g.*, [100, 101, 138]. Despite this, the model is not without certain limitations, *e.g.*, complex Gaussian variables may fail to capture high-amplitude or sparse signal dynamics, and they prevent the combination of strong priors on magnitudes with weak priors on phases, the statistics of the two being intertwined.

Contributions[1]

In [S10], we proposed a blind M-channel, K-source separation method based on independent K-mixtures of α-*stable distributions* at each frequency in the STFT domain. This family of distributions generalizes Gaussian ($\alpha = 2$) and Cauchy ($\alpha = 1$) distributions, allowing the modeling of heavier-tail or sparser statistics by setting $\alpha \in (0, 2]$. Since the likelihood is not available for $\alpha \notin \{1, 2\}$, the model was fitted using *generalized moment matching* and a greedy approach. A different stability α_k, variance σ_k^2, and mixing weight π_k was estimated for each source k. The approach was shown to outperform three similar Gaussian-based variants on two-channel speech ($K = 3$) and music ($K = 4$) mixtures, where an oracle was used by all methods to solve the frequency-permutation problem.

In [S11], we proposed to model STFT-domain source signals using independent mixtures of circular-symmetric complex Gaussians, whose means are fixed and equally spread on a zero-mean circle. For a fixed number of mixture components C, the only two parameters of the models are the common variance of the Gaussians σ^2 and the circle's radius b. This allows enforcing a more flexible magnitude prior, where b controls the mean and σ^2 the strength of the prior. This is relevant in the context of *informed* source separation, where information is injected in the form of highly-compressed versions of the magnitude spectrogram of sources. In this context, using $C \in \{8, 16\}$ showed to strike better separation/bit-rate trade-offs than competing Gaussian-based approaches.

Finally, in [S18], we contributed to the then-nascent idea of using local-Gaussian models parameterized by VAEs. The proposed approach builds on [100], combining an NMF-based noise model (Sec. 9.3.2) with a VAE-based speech model, and maximizing the likelihood by means of an *expectation-maximization* (EM) algorithm. The computational bottleneck of [100] is the need for Metropolis-Hastings sampling steps at each iteration. We were able to replace these steps by a fast computation using a variational approximation. The approach was shown to perform on par with [100] on single-channel mixtures of real speech and noise signals, with a $36\times$ computational speedup.

[1] I co-initiated and supervised the work in [S10], which was implemented by the involved PhD student. The work in [S11] was equally shared with my two co-authors. I initiated and co-supervised the work in [S18], which was implemented by the involved PhD student.

14 Blind Audio Source Separation: End-to-End Approaches

Associated publications: [S19, S22, S23]

Context

As mentioned in Snippet 13, the state of the art in sound source separation has been dominated by *end-to-end* approaches since around 2019, *i.e.*, models trained to map a mixture to its separated components in a supervised fashion. Most existing architectures, as pioneered by [137, 71], are made of three blocks: (i) an *encoder* block that transforms the 1D input signal into a 2D, "time-frequency-like" representation, (ii) a *masking* block that estimates one multiplicative *mask* of the same dimension for each desired separated source and (iii) a *decoding* block that transforms the masked input back to the time domain. The models are then trained using a permutation-invariant loss. As is typically the case at the start of a new research area, a breadth of end-to-end models quickly followed in the steps of [137, 71], using independent software and evaluated on disparate datasets, not all of which were necessarily made accessible to the community, hindering reproducibility and progress.

Contributions[1]

In [S22], we proposed a *filter-bank* formulation of the encoder and decoder blocks. Based on this view, we unified and generalized a number of approaches that either used *freely learned* or *fixed* filters by introducing *parametric* filters. We proposed to further constrain the filter design by making them *analytic*, *i.e.*, with real and imaginary parts forming Hilbert-transform pairs in the discrete-Fourier domain. We compared different variations of the encoding-masking-decoding scheme, using different window sizes and different representations of complex numbers. Extensive experiments on clean and noisy single-channel separation tasks notably showed that the analytic constraint on filters offered a consistent performance boost.

The works in [S19, S23] represent more direct contributions to the end-to-end audio-source-separation research community. *Asteroid* [S23] is an open-source versatile toolkit facilitating the implementation of source-separation methods following the encoder-masking-decoder framework, in PyTorch. *LibriMix* [S19] represents the first fully open-source, creative-common dataset aimed at training and benchmarking (noisy) speech source separation methods. The dataset features different levels of noise and source overlaps, and models trained on it were shown to exhibit better generalizability to out-of-distribution mixtures than on a previously existing dataset. Both endeavors seem to have significantly benefited the community since then, their respective GitHub repositories cumulating ∼3k stars and ∼500 forks as of June 2025.

[1] The work presented here was initiated and led by M. Pariente, under my (co-)supervision. [S22] was implemented by M. Pariente, while [S23] and [S19] were co-implemented by M. Pariente, J. Cosentino and other collaborators.

15 Robot Audition

Associated publications: [S1, S3, S5, S14–S16, S20, S27]

Context

Robot audition is concerned with endowing robots with auditory capabilities, including the recognition, detection and localization of speech or other sound events in their environment. Audio can be a useful modality on its own or complementarily to other ones in many applications, including interactive robotics [S1] or search-and-rescue with drones [S20]. As indicated in our surveys [S1, S20], a number of common challenges specific to robot audition can be identified. These include (i) the noise produced by the robots themselves, called *egonoise*, which can be loud due to its proximity to microphones, and hard to model due to transient, harmonic and random components; (ii) dynamic scenarios where target sources, interferers and receivers may all be moving; and (iii) the sound emitted by embedded loudspeakers, whose response at microphones can be non-linear due to mechanical vibrations or clipping. But as identified in our survey on audio-motor integration [S16], robot audition also presents specific opportunities. Indeed, noises induced by robot movements are reproducible and strongly structured spatially and spectrally, making them amenable to physics- or data-driven modeling. Moreover, the mobility of robots may enable the active or passive acquisition of multiple viewpoints in an audio scenes, for sensing or training purposes.

Contributions[1]

The method proposed in [S3] and already reviewed in Snippet 12 was successfully used to enhance real speech-recordings acquired by the humanoid robot *Nao* while waving the arm of walking. The spatial and spectral characteristics of these egonoises were modeled as phase-corrected combinations of multichannel *atoms* contained in a *dictionary*, learned beforehand from recordings of repeated executions of these actions in the absence of other sources. The approach was further improved in [S5] by direcly mapping *motor signals* from the robot's actuators to the indices and weights of their corresponding audio atoms, using a supervised classifier.

The series of papers [S14, S15, S20] detail the outcomes of a three-year research effort centered on the task of drone-embedded sound source localization in the context of search and rescue. These include a literature survey [S20], the acquisition [S14] and crowd-sourcing [S20] of annotated data, the implementation and benchmarking of open-source software [S14] and the organization of an international IEEE student competition [S15]. Data, code and more information on the *DREGON* project can be found at `dregon.inria.fr`. Finally, [S27] establishes a link between this line of research and the core topic of this thesis, by proposing a correlation-based presence-probability estimator for nearby *acoustic reflectors*, using one embedded microphone and a drone's egonoise. We assumed that the direct-path signal from the egonoise source could be obtained, *e.g.*, from its known location and another close-range microphone. Using simulated data incorporating measured drone noise from DREGON, the method was shown to reliability detect whenever a reflector was less than one meter away, even in the presence of diffuse background noise at down to 0 dB SNR.

[1] I contributed to some experiments in [S1] and initiated and implemented the work in [S3] during my postdoc. I initiated and co-supervised the work in [S5]. I initiated, led, supervised and co-implemented the work in [S14, S15]. I am the main/only author of [S16, S20]. I co-supervised the work in [S27].

16 Diffusion-Based Audio Generative Models

Associated publications: [S33, S35]

Context

The rapidly expanding capabilities of generative models driven by deep-learning methodologies have recently enabled the generation of long-form audio signals, *e.g.*, entire music tracks, from extremely compressed representations or weak conditioning such as short text prompts. This is made possible by means of *discrete* short-time representations obtained by *quantized* convolutional auto-encoders with large receptive fields, *e.g.*, the *EnCodec* model [139]. These discrete *audio tokens* can in turn be generated auto-regressively using a so-called *audio language model* (ALM), leveraging a transformer-based architecture trained on a large-scale unlabeled audio dataset. Decoding such discrete tokens is however prone to generating audible artifacts when the conditioning is flawed or imperfect. Besides, ALMs raise ethical concerns due to potential misuses allowed by the undetectability of their output.

Contributions[1]

In [S33], we proposed to replace the decoding block in *EnCodec* [139] by a diffusion model to synthesize high-fidelity (24 kHz) audio signals from audio-token series. The generative nature of diffusion models enabled a decrease of audible artifacts, achieving state-of-the-art subjective perceptual-quality scores across all tested modalities (speech, music, environmental sounds) according to MUSHRA tests. The proposed diffusion process operates in the complex short-time Fourier transform (STFT) domain, with real and imaginary parts concatenated. The key idea is to split the STFT into 4 mel-scale frequency bands, and to train a distinct diffusion model on each band, using the same custom power-law noise schedule for each.

In [S35], addressing the challenge of detecting audio signals created by generative models, we proposed a framework to add imperceptible *watermarks* to an ALM's output, which can be detected with high accuracy but cannot be removed easily, even in a scenario where both the ALM weights and the watermarker's source code are public. It proceeds in three steps, (i) the watermark generator and detector are jointly trained to be robust to auto-encoding by EnCodec while minimizing perceptual impact; (ii) the generator is used to watermark an entire (proprietary) audio training set; (iii) an ALM decoded by EnCodec is trained on the watermarked set. Experiments showed that classical attacks to reduce watermark detection failed to decrease the detector's accuracy below 95%, *including fine-tuning the ALM* on a non-watermarked set. More precisely, to decrease detection accuracy below 80%, the amount of fine tuning required was such that the performance of the resulting model were reduced close to that of a model trained from scratch on the non-watermarked set.

[1] I co-supervised this work, which was initiated and implemented by R. San Roman and collaborators at Meta.

Bibliography

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

[2] D. Habault and P. Filippi, "Ground effect analysis: Surface wave and layer potential representations," *Journal of Sound and Vibration*, vol. 79, no. 4, pp. 529–550, 1981.

[3] R. van denBoomgaard and R. van derWeij, "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, 2001, pp. 205–214.

[4] M. Kac, "Can one hear the shape of a drum?" *The american mathematical monthly*, vol. 73, no. 4P2, pp. 1–23, 1966.

[5] C. Gordon, D. L. Webb, and S. Wolpert, "One cannot hear the shape of a drum," *Bulletin of the American Mathematical Society*, vol. 27, no. 1, pp. 134–138, 1992.

[6] E. Brandão, A. Lenzi, and S. Paul, "A review of the in situ impedance and sound absorption measurement techniques," *Acta Acustica united with Acustica*, vol. 101, no. 3, pp. 443–463, 2015.

[7] J.-M. Jot and K. S. Lee, "Augmented reality headphone environment rendering," in *Audio Engineering Society Conference: AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society, 2016.

[8] S. Koyama, E. De Sena, P. Samarasinghe, M. R. Thomas, and F. Antonacci, "Past, present, and future of spatial audio and room acoustics," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

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

[10] B. F. Katz, D. Murphy, and A. Farina, "The past has ears (PHE): XR explorations of acoustic spaces as cultural heritage," in *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, Springer, 2020, pp. 91–98.

[11] M. Cairoli and L. C. Tagliabue, "Digital twin for acoustics and stage craft facility management in a multipurpose hall," in *Acoustics*, MDPI, vol. 5, 2023, pp. 909–927.

[12] M. Jälmby, F. Elvander, and T. van Waterschoot, "Compression of room impulse responses for compact storage and fast low-latency convolution," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 45, 2024.

[13] H. Rosseel and T. van Waterschoot, "A state-of-the-art review of acoustic preservation of historical worship spaces through auralization," *Signal Processing*, p. 109 992, 2025.

[14] C. Evers and P. A. Naylor, "Acoustic slam," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[15] U. Saqib and J. R. Jensen, "A framework for spatial map generation using acoustic echoes for robotic platforms," *Robotics and Autonomous Systems*, vol. 150, p. 104 009, 2022.

[16] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the cocktail party," *IEEE journal of selected topics in signal processing*, vol. 9, no. 5, pp. 825–836, 2015.

[17] G. P. Nava, Y. Yasuda, Y. Sato, and S. Sakamoto, "On the in situ estimation of surface acoustic impedance in interiors of arbitrary shape by acoustical inverse methods," *Acoustical science and technology*, vol. 30, no. 2, pp. 100–109, 2009.

[18] N. Antonello, T. van Waterschoot, M. Moonen, and P. A. Naylor, "Evaluation of a numerical method for identifying surface acoustic impedances in a reverberant room," in *Proc. of the10th European Congress and Exposition on Noise Control Engineering*, 2015, pp. 1–6.

[19] N. Bertin, S. Kitić, and R. Gribonval, "Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 6340–6344.

[20] Y. Okawa, Y. Watanabe, Y. Ikeda, and Y. Oikawa, "Estimation of acoustic impedances in a room using multiple sound intensities and fdtd method," in *27th International Congress on Sound and Vibration, ICSV 2021*, Silesian University Press, 2021.

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

[22] R. Badeau, "Statistical wave field theory," *The Journal of the Acoustical Society of America*, vol. 156, no. 1, pp. 573–599, 2024.

[23] D. Markovic, F. Antonacci, A. Sarti, and S. Tubaro, "Estimation of room dimensions from a single impulse response," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[24] L. Zamaninezhad, P. Annibale, and R. Rabenstein, "Localization of environmental reflectors from a single measured transfer function," in *6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2014, pp. 157–160.

[25] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1539–1551, 2015.

[26] T. Shlomo and B. Rafaely, "Blind localization of early room reflections using phase aligned spatial correlation," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1213–1225, 2021.

[27] T. J. Sejnowski, "The unreasonable effectiveness of deep learning in artificial intelligence," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 033–30 038, 2020.

[28] A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Blind room volume estimation from single-channel noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 231–235.

[29] C. Ick, A. Mehrabi, and W. Jin, "Blind acoustic room parameter estimation using phase features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[30] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2018, pp. 136–140.

[31] H. N. Bicer, C. Tuna, A. Walther, and E. A. Habets, "Data-driven joint detection and localization of acoustic reflectors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, IEEE, 2024, pp. 745–749.

[32] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2020.

[33] P. Götz, C. Tuna, A. Walther, and E. A. Habets, "Blind reverberation time estimation in dynamic acoustic conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 581–585.

[34] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 1–5.

[35] J. F. Claerbout, "Fundamentals of Geophysical Data Processing, 2nd edn," *Geophysical Journal*, vol. 86, no. 1, pp. 217–219, Jul. 1986.

[36] R. Lang, "On the eigenvalues of the non-self-adjoint robin laplacian on bounded domains and compact quantum graphs," Ph.D. dissertation, Dissertation, Stuttgart, Universität Stuttgart, 2020, 2021.

[37] T. Sprunck, "Can one hear the shape of a room?" Ph.D. dissertation, Université de Strasbourg, 2024.

[38] B. Mondet, J. Brunskog, C.-H. Jeong, and J. H. Rindel, "From absorption to impedance: Enhancing boundary conditions in room acoustic simulations," *Applied Acoustics*, vol. 157, p. 106 884, 2020.

[39] F. P. Mechel, *Formulas of acoustics*. Springer Science & Business Media, 2004, vol. 2.

[40] Z. Xu, A. Herzog, A. Lodermeyer, E. A. Habets, and A. G. Prinn, "Simulating room transfer functions between transducers mounted on audio devices using a modified image source method," *The Journal of the Acoustical Society of America*, vol. 155, no. 1, pp. 343–357, 2024.

[41] S. M. Schimmel, M. F. Muller, and N. Dillier, "A fast and accurate "shoebox" room acoustics simulator," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2009, pp. 241–244.

[42] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 351–355.

[43] H. E. Bass, L. C. Sutherland, A. J. Zuckerwar, D. T. Blackstock, and D. Hester, "Atmospheric absorption of sound: Further developments," 1995.

[44] R. Badeau, "Statistical wave field theory: Special polyhedra," *The Journal of the Acoustical Society of America*, vol. 157, no. 3, pp. 2263–2278, 2025.

[45] J. Gómez Bolaños, "Features and applications of the laser-induced spark as a monopole source for acoustic impulse response measurements," 2017.

[46] D. Schröder, *Physically based real-time auralization of interactive virtual environments*. Logos Verlag Berlin GmbH, 2011, vol. 11.

[47] L. Hormander, "The analysis of linear partial differential operator i," *I, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]/Springer-Verlag*, vol. 256, 1983.

[48] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Elsevier, 1999.

[49] M. Brandner, M. Frank, and D. Rudrich, "DirPat—database and viewer of 2D/3D directivity patterns of sound sources and receivers," in *Audio Engineering Society Convention 144*, Audio Engineering Society, 2018.

[50] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[51] S. Dilungana, "Apprentissage automatique et optimisation pour la détermination des propriétés acoustiques d'une salle à partir de signaux audio," Ph.D. dissertation, Université de Strasbourg, 2024.

[52] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[53] M. R. Hestenes, E. Stiefel, *et al.*, "Methods of conjugate gradients for solving linear systems," *Journal of research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, 1952.

[54] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[55] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, "The sliding frank–wolfe algorithm and its application to super-resolution microscopy," *Inverse Problems*, vol. 36, no. 1, p. 014 001, 2019.

[56] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on pure and applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[57] Y. Traonmilin and J.-F. Aujol, "The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem," *Inverse Problems*, vol. 36, no. 4, p. 045 003, 2020.

[58] P.-J. Bénard, Y. Traonmilin, and J.-F. Aujol, "Fast off-the-grid sparse recovery with over-parametrized projected gradient descent," in *30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022.

[59] Y. De Castro and F. Gamboa, "Exact reconstruction using beurling minimal extrapolation," *Journal of Mathematical Analysis and applications*, vol. 395, no. 1, pp. 336–354, 2012.

[60] C. Boyer, A. Chambolle, Y. D. Castro, V. Duval, F. De Gournay, and P. Weiss, "On representer theorems and convex regularization," *SIAM Journal on Optimization*, vol. 29, no. 2, pp. 1260–1281, 2019.

[61] F. Gerber, *optimparallel - A parallel version of scipy.optimize. minimize(method='L-BFGS-B')*, version v0.0.6-2, Jun. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3888570.

[62] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[63] H. P. Tukuljac, V. Pulkki, H. Gamper, K. Godin, I. J. Tashev, and N. Raghuvanshi, "A sparsity measure for echo density growth in general environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 1–5.

[64] T. Shlomo and B. Rafaely, "Blind localization of early room reflections using phase aligned spatial correlation," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1213–1225, 2021.

[65] Y. Traonmilin, J.-F. Aujol, and A. Leclaire, "Projected gradient descent for non-convex sparse spike estimation," *IEEE Signal Processing Letters*, vol. 27, pp. 1110–1114, 2020.

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

[67] M. Vorländer, "Computer simulations in room acoustics: Concepts and uncertainties," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1203–1213, 2013.

[68] Z. Tang, R. Aralikatti, A. J. Ratnarajah, and D. Manocha, "Gwa: A large high-quality acoustic dataset for audio processing," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–9.

[69] M. Vorländer and E. Mommertz, "Definition and measurement of random-incidence scattering coefficients," *Applied acoustics*, vol. 60, no. 2, pp. 187–199, 2000.

[70] F. Zotter, *Analysis and synthesis of sound-radiation with spherical arrays*. na, 2009.

[71] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[72] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[73] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[74] F. B. Gelderblom, Y. Liu, J. Kvam, and T. A. Myrvoll, "Synthetic data for dnn-based doa estimation of indoor speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 4390–4394.

[75] M. Crocco, A. Trucco, V. Murino, and A. Del Bue, "Towards fully uncalibrated room reconstruction with sound," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, IEEE, 2014, pp. 910–914.

[76] M. Crocco and A. Del Bue, "Estimation of TDOA for room reflections by iterative weighted L1 constraint," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 3201–3205.

[77] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, and J. H. Cernocky, "Building and Evaluation of a Real Room Impulse Response Dataset," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[78] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration," *IEEE Signal Processing Magazine*, no. July, pp. 14–28, 2016.

[79] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1770–1778, 2008.

[80] K. Kowalczyk, E. A. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for tdoa estimation of room reflections," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 653–656, 2013.

[81] L. Kleeman and R. Kuc, "Sonar sensing," in *Springer handbook of robotics*, Springer, 2016, pp. 753–782.

[82] H. Sato, M. C. Fehler, and T. Maeda, *Seismic wave propagation and scattering in the heterogeneous earth*. Springer, 2012, vol. 496.

[83] A. Achim, B. Buxton, G. Tzagkarakis, and P. Tsakalides, "Compressive sensing for ultrasound RF echoes using a-stable distributions," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, IEEE, 2010, pp. 4304–4307.

[84] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[85] Y. Lin, J. Chen, Y. Kim, and D. Lee, "Blind channel identification for speech dereverberation using l1-norm sparse learning," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[86] A. Aissa-El-Bey and K. Abed-Meraim, "Blind SIMO channel identification using a sparsity criterion," in *IEEE 9th International Workshop on Signal Processing Advances in Wireless Communications(SPAWC)*, IEEE, 2008, pp. 271–275.

[87] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, 2011.

[88] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[89] P. Stoica and R. L. Moses, *Introduction to spectral analysis*. Prentice Hall Upper Saddle River, NJ, 1997.

[90] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1417–1428, 2002.

[91] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[92] L. Condat and A. Hirabayashi, "Cadzow denoising upgraded: A new projection method for the recovery of dirac pulses from noisy linear measurements," *Sampling Theory in Signal and Image Processing*, vol. 14, no. 1, pp. 17–47, 2015.

[93] K. Bredies and M. Carioni, "Sparsity of solutions for variational inverse problems with finite-dimensional data," *Calculus of Variations and Partial Differential Equations*, vol. 59, no. 1, p. 14, 2020.

[94] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[95] R. Price and P. E. Green, "A communication technique for multipath channels," *Proceedings of the IRE*, vol. 46, no. 3, pp. 555–570, 1958.

[96] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*, Springer, 2001, pp. 157–180.

[97] R. Lebarbenchon, E. Camberlein, D. Di Carlo, C. Gaultier, A. Deleforge, and N. Bertin, "Evaluation of an open-source implementation of the srp-phat algorithm within the 2018 locata challenge," in *LOCATA Challenge Workshop, a satellite event of IWAENC 2018*, 2018.

[98] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[99] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[100] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *2018 IEEE 28th international workshop on machine learning for signal processing (MLSP)*, IEEE, 2018, pp. 1–6.

[101] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2993–3007, 2022.

[102] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

[103] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000.

[104] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 141–145.

[105] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[106] C. Raffel *et al.*, "Mir_eval: A transparent implementation of common mir metrics.," in *ISMIR*, vol. 10, 2014, p. 2014.

[107] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[108] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2499–2503.

[109] O. Schwartz, S. Gannot, and E. A. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2014.

[110] M. Kuster, "Objective sound field analysis based on the coherence estimated from two microphone signals," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 3284–3284, 2012.

[111] O. Schwartz, S. Gannot, and E. A. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm," in *24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1123–1127.

[112] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. United States: John Wiley & Sons, 2004.

[113] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[114] K. Kowalczyk, "Raking early reflection signals for late reverberation and noise reduction," *The Journal of the Acoustical Society of America (JASA)*, vol. 145, no. 3, pp. 257–263, 2019.

[115] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362.

[116] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[117] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal (ICASSP)*, vol. 2, 2001, pp. 749–752.

[118] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, 2003.

[119] L. Yang *et al.*, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[120] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[121] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[122] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[123] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[124] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *2011 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, IEEE, 2011, pp. 245–248.

[125] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, IEEE, 2012, pp. 1–6.

[126] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[127] L. Rayleigh, "Xii. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[128] W. He, P. Motlicek, and J.-M. Odobez, "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1303–1317, 2021.

[129] R. W. Gerchberg, "A practical algorithm for the determination of plane from image and diffraction pictures," *Optik*, vol. 35, no. 2, pp. 237–246, 1972.

[130] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[131] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, pp. 47–81, 2015.

[132] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[133] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE signal processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[134] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[135] G. Beaumont, "Traitements correctifs des effets de décohérence acoustique induits par les fluctuations du milieu de propagation: Algorithmes d'estimation bayésienne des directions d'arrivée en milieu fluctuant," Ph.D. dissertation, Ecole nationale supérieure Mines-Télécom Atlantique, 2020.

[136] O. Mokrỳ, P. Záviška, P. Rajmic, and V. Veselỳ, "Introducing spain (sparse audio inpainter)," in *2019 27th European Signal Processing Conference (EUSIPCO)*, IEEE, 2019, pp. 1–5.

[137] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 696–700.

[138] P. Wang and X. Li, "Rvae-em: Generative speech dereverberation based on recurrent variational auto-encoder and convolutive transfer function," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 496–500.

[139] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023, ISSN: 2835-8856.