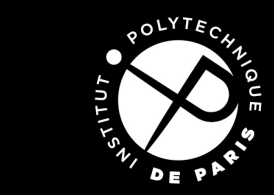
Acoustics-aware hybrid deep neural dereverberation

Louis Bahrman



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2025IPPAT052

Thèse de doctorat

Acoustics-aware hybrid deep neural dereverberation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Ecole Doctorale de l'Institut Polytechnique de Paris
(ED IP Paris)
Spécialité de doctorat : Signal, Images, Automatique et robotique

Thèse présentée et soutenue à Palaiseau, le 16 décembre 2025, par

LOUIS BAHRMAN

Composition du Jury :

Nelly Pustelnik Senior Research Scientist, ENS de Lyon	Présidente / Examinatrice
Emanuël Habets Professor, International Audio Laboratories Erlangen	Rapporteur
Axel Roebel Senior Research Scientist, IRCAM	Rapporteur
Augusto Sarti Professor, Politecnico di Milano	Examinateur
Gaël Richard Professor, Télécom Paris	Directeur de thèse
Mathieu Fontaine Associate Professor, Télécom Paris	Co-encadrant de thèse

Abstract

The aim of this thesis is to leverage room acoustics models in deep-learning-based approaches for dereverberation. Audio signals are often altered by reverberation effects induced by objects and walls of the room in which they propagate, leading to a loss in intelligibility. However, most deep learning methods developed to tackle this problem can be considered as black-box systems, as they are purely data-driven and not interpretable from a physical perspective. After studying whether neural dereverberators are consistent with physical reverberation models, we propose two hybrid approaches to train a dereverberation model in a physically realistic manner. The first one regularizes the training loss to encourage a deep neural network to produce realistic solutions, and the second is motivated by a maximum-likelihood formulation of the problem and consists in an unsupervised learning strategy that integrates a reverberation model into a deep learning framework.

Résumé

Cette thèse porte sur la modélisation de l'acoustique des salles dans les approches d'apprentissage profond pour la déréverbération. Les signaux audio enregistrés sont souvent altérés par des effets de réverbération dûs à la propagation du son dans l'espace, ce qui nuit à leur intelligibilité et leur qualité. Cependant, la plupart des approches d'apprentissage profond développées pour atténuer ces effets restent en grande partie opaques et difficilement interprétables d'un point de vue physique. Après avoir étudié la compatibilité des réseaux neuronaux profonds existants avec des modèles d'acoustique des salles, nous proposons deux méthodes d'hybridation afin d'introduire des contraintes physiques dans leur apprentissage. La première régularise la fonction de perte d'entraînement d'un réseau neuronal profond pour encourager des solutions physiquement plausibles, et la seconde s'appuie sur une formulation en maximum de vraisemblance du problème de déréverbération et consiste en une stratégie d'apprentissage non supervisée intégrant un modèle de réverbération dans un réseau neuronal.

Acknowledgements

This document concludes my PhD journey, and I'd like to first thank those who directly contributed to this experience.

Thank you Gaël Richard and Mathieu Fontaine, you were the duet of supervisors I could never have dreamt of. Gaël, ça a été un honneur d'avoir pu travailler sur ton projet [1]. Tu m'as donné cette formidable opportunité, m'as accueilli avec confiance, patiemment permis de développer mes idées, et de naviguer les deadlines par ta disponibilité. Ton expérience et ta vision de la belle recherche, ainsi que ton excellent encadrement, ont été et seront une inspiration pour les années à venir.

Mathieu, tu as été un encadrant toujours accessible. Ta sympathie et cette proximité ont rendu nos échanges toujours naturels et spontanés. Tu m'as fait découvrir que la beauté pouvait se cacher dans la subtilité d'une preuve mathématique, ou dans le style de rédaction habile qui te caractérise et que tu exprimais dans tes relectures. Tu as su apporter à ce projet une expertise et une rigueur remarquables combinées à la fraîcheur des idées du jeune chercheur.

I'd like to warmly thank the jury members for taking the time to read this document and take part to the defense. Thanks to my reviewers with whom I had valuable and fruitful exchanges.

My gratitude extends to those I had the chance to collaborate with, Marius Rodrigues, who was an exceptional intern and PhD student, Jonathan Le Roux for his perspicacity, and Roland Badeau, for his brilliant ideas and teaching skills.

Speaking of teachers, I'd like to also mention my former teachers and supervisors, who advised and inspired me to start this adventure. Thanks to the professors who entrusted me with teaching, which I greatly enjoyed.

I'd like to thank all members of the ADASP group, who started as "colleagues" and then became friends. Among them, I'd like to mention the Louis and Marius from the "Reverb" team, for the fascinating and inspiring discussions. Another special mention to the 5C01 office : Bernardo, Changhong, José, Manvi. José, you were always the

1. This work was funded by the European Union (ERC, HI-Audio, 101052978). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was performed using HPC resources from GENCI–IDRIS (Grant 2024-AD011014072R1)

one behind new humorous traditions which greatly contributed to the ambiance at ADASP.

Une thèse c'est non seulement un travail mais aussi un état (on dit bien *être en thèse*). J'aimerais donc profiter de ces pages pour remercier les ami·e·s qui ont rendu cette aventure vivante et collective.

À la team rando et bivouacs estivaux, aux trouvères et trouveresses de la "Communauté du bâton", aux "Artefacts, Valha & Co", au "Zoo" à Douai, Paris ou Oléron, à ceux du lycée dont le nom de groupe change à chaque fois qu'on se voit, Merci. Merci pour ces semaines de vacances musicales ou sportives, ces week-ends pour fêter d'heureux évènements, ces soirées calmes autour de jeux de société et discussions profondes sous les étoiles, ou moins calmes dans des bars bondés, en concerts et soirées. Merci d'avoir gardé le contact, et toujours rendu les retrouvailles aussi riches en émotions. Tous·tes vous nommer individuellement prendrait trop de place pour ces deux maigres pages (mais je suis sûr que vous vous reconnaitrez), c'est pourquoi j'aimerais simplement mentionner les quelques un·es que j'ai vu·e·s le plus souvent et qui m'ont aidé durant la rédaction : Hugo, Michel, Lucille, Solal, Élisa.

Merci à ma famille, qui a toujours pris de mes nouvelles, et apporté leur soutien malgré la distance.

À mes parents enfin, pour m'avoir tout donné. Vous m'avez constamment encouragé dans mes études, et énormément sacrifié pour moi. Je n'aurais jamais été aussi loin sans votre soutien inconditionnel, et vous serai éternellement reconnaissant pour tout ce que vous m'avez apporté.

Résumé substantiel

Les enregistrements vocaux sont sujets à des altérations liées aux effets de réverbération, induits par la propagation du son dans l'environnement. Ces phénomènes dégradent significativement l'intelligibilité et la qualité des signaux audio, posant ainsi un défi majeur pour les applications de communication et d'interaction homme-machine. La déréverbération, processus visant à restaurer un signal source sec à partir d'un enregistrement réverbéré, constitue donc un enjeu central dans le domaine du traitement du signal audio. Cette problématique s'inscrit dans le cadre plus large des problèmes inverses, où la déréverbération est intrinsèquement liée à l'identification des systèmes acoustiques : tandis que la première tâche cherche à estimer la source sonore, la seconde vise à caractériser les propriétés de l'espace de propagation. Deux paradigmes principaux émergent pour aborder ces défis. Le premier, dit approche guidée par un modèle, repose sur la modélisation explicite du signal à restaurer ou de la dégradation en intégrant des contraintes physiques ou statistiques. Le second, plus récent et qualifié d'approche guidée par les données ou déréverbération neuronale, exploite des réseaux de neurones profonds entraînés sur de vastes corpus de données pour résoudre ce problème inverse. Bien que les méthodes guidées par les données constituent actuellement l'état de l'art en matière de déréverbération, elles présentent des limitations notables. Leur performance dépend étroitement de la disponibilité de grandes quantités de données annotées, qui, dans le cas de la déréverbération, correspondent à des signaux anéchoïques, difficilement mesurables. De plus, de par leur fonctionnement qui reste largement opaque, ces approches peinent à intégrer des contraintes physiques ou des modèles statistiques classiques de production ou de dégradation des signaux sonores, ce qui soulève des questions quant à leur capacité à modéliser le lien entre déréverbération et acoustique des salles.

Les stratégies hybrides émergent comme une voie prometteuse pour surmonter les limitations de chacun des paradigmes de résolution des problèmes inverses. Cette thèse propose donc de développer des modèles permettant de combiner la performance des approches guidées par des données à l'interprétabilité et la faible dépendance aux données offerte par les approches guidées par un modèle, pour la tâche de déréverbération. Plus précisément, ce travail s'articule autour de deux points : d'une part, nous cherchons à évaluer la compatibilité des architectures neuronales pour la déréverbéra-

tion avec des modèles physiques d'acoustique des salles. D'autre part, nos recherches visent à exploiter ces modèles d'acoustique des salles pour améliorer les performances et la compréhension des déréverbérateurs neuronaux.

Pour y parvenir, notre démarche s'organise en deux parties. Dans une première partie, nous évaluons la capacité des déréverbérateurs neuronaux à contribuer à la tâche d'identification de système acoustique. Pour ce faire, nous utilisons un réseau neuronal afin d'estimer le signal source associé à un signal mesuré, car, étant donné ce signal source, le problème d'identification de système acoustique devient un problème de déconvolution, a priori plus simple à résoudre. Toutefois, cette approche se heurte à une difficulté majeure : alors que la plupart des méthodes de déconvolution existantes supposent un faible bruit sur le signal mesuré, aucune étude n'a, à notre connaissance, évalué leur robustesse face à des artéfacts induits par des erreurs d'estimation sur la source. Nous analysons donc les performances des méthodes de déconvolution dans ce contexte particulier. Pour pallier les limitations causées par une source présentant des artéfacts, nous proposons une méthode de déconvolution régularisée par un modèle de réverbération. Ensuite, nous concevons une fonction de perte motivée par la physique, destinée à entraîner un déréverbérateur neuronal pour qu'il puisse effectuer une tâche d'identification de système acoustique. Cette fonction de perte impose des contraintes physiques sur le système acoustique prédit en déconvoluant l'entrée du déréverbérateur neuronal par sa sortie. Théoriquement, une telle fonction de perte devrait aussi encourager le déréverbérateur neuronal à estimer un signal sec proche de la vérité-terrain. Cependant, nos expériences révèlent que l'optimisation de cette fonction de perte seule dégrade souvent les performances de déréverbération du réseau. Cette observation suggère que les déréverbérateurs neuronaux ne peuvent pas systématiquement identifier à la fois un système acoustique et un signal sec. Ainsi, il apparaît que peu d'approches neuronales actuelles sont compatibles avec ce cadre physique. En conséquence, nous proposons une stratégie de régularisation de l'entraînement d'un déréverbérateur neuronal visant à lui permettre d'accomplir conjointement les deux tâches de déréverbération et d'identification de système acoustique, sans compromettre ses performances.

Dans une seconde partie, nous abordons la question cruciale de la dépendance aux données des déréverbérateurs neuronaux. En effet, la plupart des architectures existantes ont besoin de paires d'enregistrements appariés, composés d'un signal sec et de sa version réverbérée. Même les approches génératives, qui modélisent la distribution a priori des signaux secs, restent contraintes par la disponibilité de tels enregistrements. Or, l'acquisition de signaux secs exige des conditions d'enregistrement anéchoïques, contrairement aux signaux réverbérés, abondants dans les environnements naturels. Pour surmonter cette limitation, nous introduisons un nouveau cadre de déréverbération non-supervisée, ne requérant que des données réverbérantes pour l'entraînement. À cette fin, nous proposons une formulation en maximum de vraisemblance de la dé-

réverbération guidée par un modèle stochastique et paramétrique de réverbération. Nous implémentons cette formulation au sein d'un protocole d'entraînement flexible, adapté à divers niveaux de disponibilité de données, allant d'une supervision dite forte exploitant la réponse impulsionnelle exacte, à une supervision faible ne nécessitant que quelques paramètres acoustiques, et enfin, une configuration totalement non-supervisée.

Ce travail montre que l'alliance des modèles physiques d'acoustique des salles avec les architectures neuronales permet à la fois d'évaluer leur compatibilité et d'améliorer les performances des déréverbérateurs, tout en réduisant leur dépendance aux données. Nos résultats ouvrent ainsi la voie à des approches hybrides plus efficientes en données et interprétables.

Contents

List of Figures

List of Tables

Acronyms

ACE Acoustic Characterization of Environments

ADMM Alternating Direction Method of Multipliers

AEC Acoustic Echo Cancellation

AR Auto-Regressive

ASR Automatic Speech Recognition

DDSP Differentiable Digital Signal Processing

DFT Discrete Fourier Transform

DNN Deep Neural Network

DRR Direct to Reverberant Ratio

EARS Expressive Anechoic Recordings of Speech

EDC Energy Decay Curve

EDR Energy Decay Relief

FAD Fréchet Audio Distance

FCN Fully Convolutional Network

FDN Feedback Delay Network

FSN FullSubNet

FVN Filtered Velvet Noise

IDFT Inverse Discrete Fourier Transform

IID Independant and Identically Distributed

ISM Image Source Method

ISTFT Inverse Short-Time Fourier Transform

LASSO Least Absolute Shrinkage and Selection Operator

LSTM Long Short Term Memory

BiLSTM Bidirectional Long Short Term Memory

LTI Linear Time-Invariant

MAE Mean Absolute Error

ML Maximum Likelihood

MSE Mean Squared Error

PESQ Perceptual Evaluation of speech Quality

WB-PESQ Wide-Band Perceptual Evaluation of Speech Quality

PM Parameter Matching

RIR Room Impulse Response

RM Reverberation Matching

$\mathbf{RT_{60}}$ Reverberation Time

SDR Source-to-Distortion Ratio

SI-SDR Scale-Invariant Source-to-Distortion Ratio

SGD Stochastic Gradient Descent

SNR Signal-to-Noise Ratio

SRMR Speech-to-Reverberant Modulation Ratio

STFT Short-Time Fourier Transform

STOI Short-Term Objective intelligibility

ESTOI Extended Short-Term Objective intelligibility

SWFT Statistical Wave Field Theory

TFL TF-Locoformer

WPE Weighted Prediction Error

Symbols

$\mathbb{R}, \mathbb{R}_+, \mathbb{C}, \mathbb{Z}, \mathbb{N}$	Sets of real, non-negative real, complex, integer, non-negative integer numbers
$x \in \mathbb{R}^N, X \in \mathbb{C}^N, \boldsymbol{X} \in \mathbb{C}^{F \times T}$	time-, frequency- and time-frequency domain signals
$n, \quad f, \quad t$	time-domain, frequency-domain, time-frequency frame index
$j, \quad \lvert\cdot\rvert, \quad \angle, \quad \cdot^*$	Complex number s.t. $j^2 = -1$, Magnitude, angle on the complex plane, conjugation
$g_a, \quad g_s$	STFT analysis window, synthesis window
$\star, \quad \circledast, \quad \mathcal{T}$	Convolution, Cross-correlation, Toeplitz operator
$f_s, \quad \delta, \quad \mathscr{H}$	Sampling frequency, Dirac impulse, Hilbert transform
$\odot, \quad \oslash$	Pointwise (Hadamard) vector multiplication and division
$.^\intercal, \quad .^\dagger$	Matrix transposition, Moore-Penrose pseudo-inverse
$\mathcal{L}, \quad \nabla, \quad \eta, \quad \lambda$	Loss function, gradient, step size, regularization parameter
$\lVert\cdot\rVert_2, \quad \lVert\cdot\rVert_F$	Euclidean, Frobenius norms

$\triangleq$	Equal by definition

Chapter 1

Introduction

1.1 Reverberation and dereverberation

Reverberation enables us to understand sonic spaces, at the cost of understanding sonic sources. Indeed, acoustic waves propagating in enclosed environments are significantly influenced by reflection and diffraction effects from surrounding surfaces and objects. These interactions alter the original waveform in a singular manner and result in *reverberation*, a unique characteristic of the space it traverses [1]. Reverberation, in turn, has had an immense influence on art and culture throughout history. Starting at late prehistoric times, rock art was located in places where intense echoes could be heard [2], suggesting the probable ritualistic meaning of reverberation across cultures [3, 4]. Western music performance [5] and reception [6] are largely influenced by the acoustics of performance spaces, and in turn composers are influenced by the space in which their music is supposed to be heard [7, chap. 21]. Acoustics is tightly bound to architectural design of vocal and musical performance spaces, such as theaters [8], classrooms [9, 10] or concert halls [11, 12]. As they are part of our cultural heritage, it is of crucial importance to preserve acoustics of historical spaces. Preservation of this intangible heritage can be done thanks to auralization [13, 14], the audio equivalent of visualization, which aims at reproducing an acoustic environment. Auralization techniques often involve measurement and characterization of an acoustic space, a task that we call *acoustic system identification* in the remaining of this manuscript.

In the context of voice processing, which will be the focus of this dissertation, reverberation has long been recognized as a critical factor affecting speech intelligibility [15, 16], and its detrimental effect on audio clarity has motivated decades of research. The task of reverberation suppression, commonly referred to as *dereverberation*, has received renewed attention in recent years due to its relevance in a wide range of audio processing applications. Dereverberation can be framed as an *inverse problem*, where the goal is to restore the original, reverberation-free signal (often called the

dry source) from a reverberant recording. An inverse problem is one where the goal is to recover underlying parameters or signals from observed data, typically subject to noise and uncertainty. Effective dereverberation is essential in enhancing the performance of hearing aids [17, 18], improving communication quality in hands-free [19] and hand-held [20] telephony, and enabling robust Automatic Speech Recognition (ASR) in human-machine interaction scenarios [21]. It also serves as a key preprocessing step in general-purpose speech enhancement frameworks [22].

Both tasks of dereverberation and acoustic system identification are intertwined, as they can both be seen as a system identification task: the first aims to estimate a sound-source system whereas the latter estimates an acoustic system. Related tasks involving reverberation modeling are Acoustic Echo Cancellation (AEC), Acoustic Feedback Control, and Acoustic Matching. AEC seeks to prevent a known source signal that has been played through a loudspeaker and recorded by a microphone from being captured again [23]. Acoustic feedback control extends this idea by cancelling the entire feedback loop, preventing signals emitted by loudspeakers from being re-recorded and re-amplified [24]. The recently-introduced task of acoustic matching [25] aims to disentangle both acoustic and source systems from a recording. It consists in transferring the acoustic content of one recording onto another, while keeping unaltered all features of the target recording, making it sound like it was recorded in the environment of the first recording. While both tasks of AEC and Acoustic Feedback Control assume a known source, both dereverberation and acoustic matching know neither of the source nor the acoustic system. In the remaining of the manuscript, such settings will be called *blind*.

1.2 Dereverberation: an ill-posed problem

Both tasks of blind dereverberation and blind acoustic system identification are considered to be ill-posed as per Hadamard [26, 27]. A problem is said to be ill-posed if one of the following properties does not hold:

1. Existence: the problem has a solution;
2. Uniqueness: the solution is unique;
3. Stability: the solution's behavior changes continuously with the initial conditions.

In the case of dereverberation of an unknown acoustic system, the uniqueness and stability conditions might not hold. First, the solution of the blind dereverberation problem is not uniquely identifiable [28, Section 8.2.4]. This is due to the fact that applying the acoustic system to the source signal can be expressed as sequentially applying several valid acoustic filters, a notion coined as reducibility [29]. Indeed, time-invariant reverberation filters can be factorized as a minimum-phase filter defin-

ing the coloration of the room, and an all-pass component, accounting for the delay of the echoes. Therefore, the dry source to which room coloration has been applied represents a plausible solution of the blind dereverberation problem, although this solution is undesired as we wish to remove both echoes and room coloration effects. This ambiguity makes the dereverberation solution not unique.

Even when the uniqueness condition holds, Hadamard's stability condition might not be fulfilled for dereverberation. The uniqueness condition of dereverberation can be fulfilled when the acoustic system is perfectly known. In this case, coined as *non-blind*, dereverberation becomes a *deconvolution* problem. It was shown [30], that the acoustic system defined for a room and a pair of source and microphone positions changes drastically when the source or microphone's position varies slightly. This makes the dereverberation problem unstable.

1.3 Data-driven and model-driven solutions

Definition of the model-based and data-driven paradigms

Approaches to solve ill-posed problems nowadays fall into two main paradigms [1]: *Model-driven* and *Data-driven*.

In the model-driven paradigm, ill-posed problems can be made well-posed by adding prior domain-specific knowledge of the phenomenon at the core of the problem. A model can be defined as [32, translated]: *"a representative, idealized, and open framework, acknowledged as approximate but considered fruitful with respect to a given goal (to predict, to act upon nature, to understand it better, etc.)"*. For instance, in the context of dereverberation, it is widely accepted that sound propagation is a causal phenomenon. Other models can be the result of a deliberate choice and simplification. In the context of speech dereverberation, model-driven approaches will typically introduce a model of the source, or of the reverberation phenomenon. These models will be detailed in Chap. 3. Modeling each point of the source's time-frequency representation to be Independant and Identically Distributed (IID) circular [33] or considering the autoregressive nature of speech signals [34] are two examples of approximate yet fruitful frameworks to solve the dereverberation task.

On the other hand, data-driven methods have taken an orthogonal approach to problem solving. They consist in leveraging *data* instead of domain-specific knowledge. Indeed, introducing data provides a ground-truth valid solution to a given problem, which can be generalized to some extent. Data-driven techniques for audio inverse problems typically consist in machine-learning, and more recently *deep-learning* [35] approaches. Deep learning methods involve a parametric mapping organized to mimic the structure of the human brain, and consisting in a stack of layers of artificial neurons

1. We use Kuhn's definition of paradigms [31]: "universally recognized scientific achievements that for a time provide models, problems and solutions to a community of practitioners".

called *Deep Neural Network (DNN)*. In the remaining of this manuscript we will call purely data-driven dereverberation methods *neural dereverberators*. Before being used to predict the solution of an ill-posed problem, DNN's optimal parameters for a given task have to be found from data, in a process called *training*.

It could be argued that data-driven methods can still be considered as models.

Differences between model-driven and data-driven paradigms

A question arises: can DNNs be considered as models? While answering this question would require a proper epistemological study, we would like to provide our insight on the extent to which model-based approaches and data-driven approaches differ. At first thought, throughout the history of science, it is not the first time that the parameters of a model have to be tuned using data. For example, the constant defining Newton's gravitational law (published in 1687) could only be determined more than 100 years later by Cavendish (in 1797). Moreover, deep learning approaches represent state-of-the art solvers for most audio inverse problems, including dereverberation [36], illustrating their high potential for prediction. It could also be argued that DNNs can be seen as a model for the human brain.

These arguments fall short as some strong differences between model-based and data-driven approaches can be found. These differences can be summarized by two main arguments: DNNs are overparametrized and their behaviour is opaque.

First, DNNs are highly overparametrized, meaning that they have more parameters than are strictly necessary to solve the task. From an epistemological point of view, this contradicts the widely accepted principle of Occam's razor [37, p33], stating that entities should not be multiplied [2]. From a practical point of view, the reliance on data makes data-driven methods very task-specific, and prone to a phenomenon called *overfitting*. Overfitting can apply to a data distribution, or even a task. The phenomenon of overfitting to a data distribution can happen, for instance, when neural dereverberators trained on synthetic reverberation are tested on real-world reverberation, and exhibit a loss in performance that can only be compensated by training on a dataset of realistic reverberation. Overfitting can also happen with respect to a task, when a DNN which is trained to optimize a metric performs poorly when evaluated on another metric. An illustrative example is found in the case of speech restoration [38]: a data-driven speech enhancement method that was trained to optimize the Perceptual Evaluation of speech Quality (PESQ) metric (that will be described in 3.3.5). Initially, it appeared to perform well based on PESQ scores. Yet, listening tests and evaluation using other metrics were deceiving. Indeed, the network inadvertently learned a flaw in the PESQ metric itself, which was designed to evaluate speech signals starting with low amplitudes. By generating a brief Dirac

2. Occam's statement is often rephrased to better be applied to epistemology. Funnily, its literal sense is enough to discredit DNNs.

pulse at the start of the output (intended to be a clean signal), the DNNs could make virtually any signal appear optimal according to this metric. This example show that data-driven methods can suffer from a form of *misalignment*. Misalignment occurs when the objective a model is trained to optimize does not correspond to the broader, desired goals, leading to suboptimal or even misleading performance.

A second difference between classical models and data-driven techniques is that data-driven techniques behave in an intransparent manner, making them unable to be used as instruments of knowledge [39]. It was stated that classical models can be used to investigate both the world and a theory during both their conception and their usage [40, Section 2.4]. At conception, models result from deliberate simplifications of the world, resulting from fitting a theory to observations. At usage, we learn about which situations can or cannot be framed by a model. These argument fall short in the case of DNNs. Indeed, DNNs are *universal approximators* [41], meaning that they can approximate any continuous function to any desired degree of accuracy, provided they have sufficient parameters. From the point of view of continuous optimization, data-driven techniques are unable to be constrained by a theory, or to integrate a simplification of the world. Their predictions can only be steered by techniques such as *regularization*, which can be considered as a weak way to incorporate constraints into a training objective as it allows for some flexibility. A contradiction arises: because of their expressive power, DNNs can only fail to succeed on a task if they are wrongly trained, or their number of parameters is insufficient for the universal approximation theorem to apply. It would then be incorrect to assume that failure of a data-driven method could be caused by wrong simplifications or assumptions at the core of its design. This argument on falsifiability limits the role of data-driven methods as mediators of knowledge unlike classical models [39, 40].

The black-box nature and overparametrization of data-driven approaches have not only epistemological but also societal consequences. O'Neil showed that data-driven methods can fuel vicious circles of bias, learnt from the worldview imposed by unquestioned and irresponsible data extraction and usage [42]. Similarly to the model that deceived its evaluators by learning a bias of the metric, data-driven methods appear efficient for sovereign and commercial use when they instead reinforce inequalities. Overconfidence in the results of these intransparent methods causes them to be overly trusted and less and less regulated. Crawford [43] moreover warns about inconsiderate usage of resources by data-driven methods, whether they take the form of rare earth, data, or energy. Audio-related tasks such as voice generation are now being also evaluated through the lens of their energy consumption [44], which enables to partially quantify their impact on climate change [45]. While these arguments are not directly related to our contributions, we believe that the societal and environmental implications of data-driven methods should not be neglected by the academic community.

1.4 Hybrid deep learning

The model-driven and data-driven paradigms are however not mutually exclusive. In the recent years, model-based and data-driven approaches have been combined in *hybrid* approaches. Hybrid models combine the advantages of model-based and data-driven approaches. Indeed, they are able to be interpreted like classical models, while leveraging the performance of DNNs. More precisely, unlike purely data-driven which required post-hoc explainability techniques to be understood, hybrid approaches are interpretable by design [46]. Balancing model-driven and data-driven techniques for signal processing defines a whole continuous spectrum of hybridization techniques, that are neither purely model-based nor data-driven [47].

In the domain of audio restoration, such approaches fall into three main categories: *Plug-and-Play* methods aim to leverage deep-learning-based priors in order to regularize a classical optimization problem. Typically, the Alternating Direction Method of Multipliers (ADMM) offers an iterative algorithm that enables to solve an inverse problem via a classical iterative algorithm that leverages a prediction by a DNN at each iteration [48]. The second category regroups *unrolling* [49] techniques, defined as a variant of Plug-and-Play methods where a complete iteration of a classical algorithm is modeled by one or several layers of a neural architecture. Finally, data-driven techniques can be leveraged to estimate the parameters of a classical model, that has been made differentiable. For instance, the popular Differentiable Digital Signal Processing (DDSP) framework allows a DNN to be trained using an optimization objective involving differentiable implementations of signal operators. Other approaches in this category, such as physics-informed neural networks, have been less used in single-channel audio restoration, due to the fact that they require several measurements of a single system to be trained.

1.5 Research questions

So far, we have seen that dereverberation and reverberation modeling are two sides of the same coin of system identification, characterizing respectively *source* and *space* from a reverberant recording. On the other hand, while purely data-driven techniques could not model this duality, hybrid deep learning seems to provide the necessary framework to jointly model room acoustics and dereverberation. Two research questions emerge from this statement.

We first want to evaluate existing neural dereverberators on their ability to model room acoustics. This objective is summarized in the following research question:

Research Question 1

To what extent are existing neural dereverberators consistent with room acoustics?

We also wish to alleviate the drawbacks of both classical and purely data-driven approaches, focusing on performance, data consumption, and ability to model the source-space duality. Hence, we aim to also answer the following question:

Research Question 2

How to leverage room acoustics models in data-driven methods for dereverberation?

1.6 Contributions

To answer these questions, our approach is organized into two parts. In the first part, we evaluate the ability of neural dereverberators to contribute to the task of acoustic system identification. To do so, we use a neural network to estimate the source signal associated with the measured signal, because, given this source signal, the problem of acoustic system identification becomes a deconvolution problem, which is a priori simpler to solve. We evaluate existing deconvolution methods in the specific context where the artefacts are present on an estimated source signal and not on a measured signal. To address the limitations caused by a source signal with artifacts, we propose a deconvolution method regularized by a reverberation model. We then design a physics-driven optimization objective to train a neural dereverberator to perform an acoustic system identification task. This loss function computes physical properties from the acoustic system induced by the dereverberation process. Experiments show that considering this optimization objective alone often degrades the dereverberation performance of the network, suggesting that neural dereverberators cannot systematically identify both an acoustic system and a clean signal simultaneously. Consequently, we propose a training regularization strategy for neural dereverberators aimed at enabling them to jointly perform both dereverberation and acoustic system identification without compromising their performance. These findings are summarized in Chapter 4, and the training regularization strategy for neural dereverberators has been published in:

[50]: Louis Bahrman, Mathieu Fontaine, Jonathan Le Roux, and Gaël Richard. « Speech Dereverberation Constrained on Room Impulse Response Characteristics ». In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 622–626

In the second part, corresponding to Chapter 5, we address the crucial issue of

data dependence in neural dereverberators. Indeed, most existing architectures require paired recordings, consisting of a dry signal and its reverberated version. Even approaches which estimate the prior distribution of dry signals, remain constrained by the availability of such recordings. However, acquiring dry signals requires anechoic recording conditions, in contrast to reverberated signals, which are abundant in natural environments. To overcome this limitation, we introduce a novel unsupervised dereverberation framework that requires only reverberant data for training. To achieve this, we propose a maximum-likelihood formulation for dereverberation, guided by a parametric reverberation model. We implement this formulation within a flexible training protocol, suitable for various levels of data availability, ranging from strong supervision using the exact characterization of the acoustic system, to weak supervision requiring only a few acoustic parameters, and finally, a fully unsupervised configuration. These results have been summarized in the following articles:

[51]: Louis Bahrman, Mathieu Fontaine, and Gaël Richard. « A Hybrid Model for Weakly-Supervised Speech Dereverberation ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5

[52]: Louis Bahrman, Mathieu Fontaine, and Gaël Richard. « Déréverbération Non-Supervisée de La Parole Par Modèle Hybride ». In: *XXXe Colloque Francophone de Traitement Du Signal et Des Images*. Strasbourg, France: GRETSI, Aug. 2025

[53]: Louis Bahrman, Marius Rodrigues, Mathieu Fontaine, and Gaël Richard. « U-DREAM: Unsupervised Dereverberation Guided by a Reverberation Model ». In: *IEEE Transactions on Audio, Speech and Language Processing* 34 (2026), pp. 1552–1563

1.7 Manuscript structure

After this introduction, the remaining of this manuscript is organized as follows: in Chapter 2, we introduce the theoretical framework and tools used throughout this manuscript, including audio signal representations and operations, and selected approaches for optimization. Then, in Chapter 3, we present current research directions in reverberation modeling, acoustic system identification, and dereverberation. We detail reverberation models that can be leveraged in hybrid models for acoustic system identification and dereverberation. We also present a typology of dereverberation methods based on the assumptions they make about the reverberation process. A

second part of this manuscript gathers our contributions. In Chapter 4, we deal with acoustic system identification by neural dereverberators, and aims to answer the first research question on compatibility of data-driven approaches for dereverberation with room acoustics. Then, in Chapter 5, we propose to leverage a room acoustics model in the training objective of a neural dereverberator, in order to reduce its reliance on large quantities of dry source datasets. Finally, we conclude this manuscript, answer the research questions and present future research directions in Chapter 6.

Part I

Theoretical framework and literature review

Chapter 2

Theoretical framework

Chapter abstract

In this chapter, we introduce commonly used tools for audio signal processing, starting with signal representations in the time, frequency and time-frequency domains. Next, we define filtering technique for time-domain signals and various adaptations in the time-frequency domain. Finally, we present optimization techniques that will be relevant for the throughout this manuscript, with examples for the task of deconvolution as it plays a central role in our contributions.

2.1 Signal representations

2.1.1 Time domain

Throughout this manuscript, audio signals are considered to be obtained by sampling a function of the local air pressure at discrete time steps. The *sampling frequency* of such signals, denoted f_s, limits their *bandwidth*, defined as the maximal frequency under which the original function can be deterministically reconstructed. This frequency is coined as Nyquist frequency, of value $\frac{f_s}{2}$. Typical sampling frequencies for audio signal processing include 16 kHz (narrow-band), 44.1 kHz, and 48 kHz (full-band). 16 kHz is sufficient to record intelligible speech, while 44.1 kHz and 48 kHz are used to encompass the full range of the human auditory system, which extends up to 20 kHz.

When they are of finite length, they will be considered as vectors and indexed by the time-domain index n; their n^{th} element is written $x[n]$.

2.1.2 Frequency domain

Discrete time-domain audio signals can also be studied in the dual frequency domain, through a change of basis performed using the Discrete Fourier Transform

(DFT) operator. This operator is defined for a time-domain signal $x \in \mathbb{R}^N$ as:

$$\mathrm{DFT}(x)[f] \triangleq X[f] = \sum_{n=0}^{N-1} x[n] e^{-2j\pi nf/N}, \qquad f \in 0, 1, \ldots, N-1. \tag{2.1}$$

In this expression, f represents the frequency and j is the complex number such that $j^2 = -1$. Throughout this document, $\triangleq$ denotes equality by definition. As a complex-valued signal, $X[f]$ can be expressed as the combination of its magnitude $|X[f]|$ and phase $\angle X[f]$ components.

This transform is a bijection and the Inverse Discrete Fourier Transform (IDFT) is defined for a frequency-domain signal $X \in \mathbb{C}^N$:

$$\mathrm{IDFT}(X)[n] \triangleq x[n] = \frac{1}{N} \sum_{f=0}^{N-1} X[f] e^{+2j\pi nf/N}, \qquad n \in 0, 1, \ldots, N-1. \tag{2.2}$$

2.1.3 Time-frequency domain

Given a discrete signal x, a positive analysis window g_a of finite size N, and a hop-size L, the N-band Short-Time Fourier Transform (STFT) of x is the matrix $\boldsymbol{X} \in \mathbb{C}^{F \times T}$, which elements are defined as:

$$\mathrm{STFT}(x)_{f,t} \triangleq \boldsymbol{X}_{f,t} = \sum_{n=0}^{N-1} x[n+tL] g_a[n] e^{-j2\pi fn/N}, \text{ for } f \in [\![0, F-1]\!] \text{ and } tL \leq N-L. \tag{2.3}$$

Similarly to the Fourier transform, the STFT is complex-valued. The magnitude of the STFT representation of a signal is called *spectrogram*, and often expressed in decibels (dB) for interpretability and visualization. Phase processing and retrieval in the STFT domain has been a critical aspect of model-driven [54, 27] and data-driven [55] signal processing.

The concept of time-frequency representations can be extended to a larger class of continuous distributions, called Cohen's class [56]. Notable distributions include the Wigner-Ville, or the Page [57] distributions [58]. The Wigner-Ville distribution provides a high-resolution representation of a signal's energy jointly in time and frequency, without the trade-off imposed by the STFT's fixed window.

The N-band STFT can be inverted, and the Inverse Short-Time Fourier Transform (ISTFT) of a time-frequency representation $\boldsymbol{X} \in \mathbb{C}^{F \times T}$ is defined as:

$$\mathrm{ISTFT}(\boldsymbol{X})[n] \triangleq x[n] = \frac{1}{N} \sum_{\substack{t \\ 0 \leq n - Lt < N}} g_s[n - Lt] \sum_{f=0}^{N-1} \boldsymbol{X}_{f,t} e^{j2\pi f(n-tL)/N}. \tag{2.4}$$

The synthesis window g_s has to be carefully designed for perfect reconstruction, using

the normalized overlap-add condition:

$$\forall n \in \mathbb{N}, \quad \sum_{t=0}^{T} g_a[n - tL] g_s[n - tL] = 1. \tag{2.5}$$

2.2 Operations on signals

In this section, we define some operators on signals that will be used in our works.

All Linear Time-Invariant (LTI) systems, or *filters*, are characterized by their *impulse response*. The impulse response h of a filter is defined as the resulting signal when the filter is applied to a Dirac delta signal $\delta : \delta[0] = 1$ and $\forall n \neq 0, \delta[n] = 0$. In the general case, the output of a filter is the *convolution* of the filter's impulse response and the input.

2.2.1 Convolution

In time domain

The convolution between two sequences h and s, is defined as:

$$(s \star h)[n] \triangleq \sum_{k \in \mathbb{Z}} s[k] h[n - k]. \tag{2.6}$$

The convolution operation is linear, associative, and commutative. The DFT can be used to compute convolution in an efficient manner, if the support of both sequences are finite. Let N_s and N_h be the length of the support of s and h respectively. By zero-padding s and h such that both their lengths are equal to $N_s + N_h - 1$, it can be shown [59, Theorem 3.10] that:

$$(s \star h) = \mathrm{IDFT}\left(\mathrm{DFT}(s) \odot \mathrm{DFT}(h)\right), \tag{2.7}$$

where $\odot$ denotes pointwise product. This condition is also necessary to express the convolution operation using matrix product. Indeed:

$$(s \star h) = \mathcal{T}(s) h, \tag{2.8}$$

where $\mathcal{T}$ is a Toeplitz operator defined given a vector $x \in \mathbb{C}^N$ as:

$$\mathcal{T}(x) = \begin{bmatrix} x[0] & x[-1] & x[-2] & \cdots & x[-N+1] \\ x[1] & x[0] & x[-1] & \cdots & x[-N+2] \\ x[2] & x[1] & x[0] & \cdots & x[-N+3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[N-1] & x[N-2] & x[N-3] & \cdots & x[0] \end{bmatrix}. \tag{2.9}$$

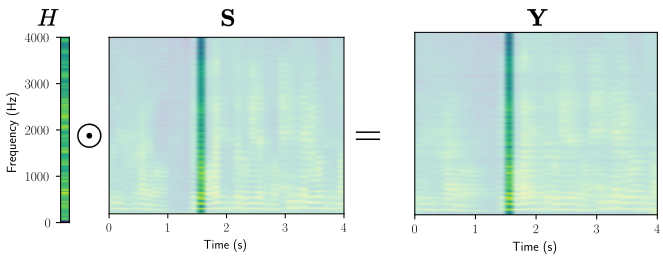


Figure 2.1 – Illustration of narrow-band filtering.

In time-frequency domain

Several formulations of convolution in the STFT domain are widely used in the literature, and detailed hereunder.

Narrow-band filtering Assuming the filter's impulse is short with respect to the STFT analysis window, it can be considered that the filter is applied to each STFT frame independently. Hence, for each frame, the frequency response of the filter's output is the frequency response of the filter pointwise multiplied by the frequency response of the source. Denoting $H \in \mathbb{C}^F$ the F-point DFT of the filter's impulse response, and $\boldsymbol{S}, \boldsymbol{Y} \in \mathbb{C}^{F \times T}$ the STFT of the source and output respectively, the *narrow-band* approximation states that:

$$\forall f \in [\![0, F[\![, t \in [\![0, T, [\![, \boldsymbol{Y}_{f,t} = H[f]\boldsymbol{S}_{f,t}. \tag{2.10}$$

Narrow-band filtering is illustrated in Figure 2.1. This approximation is in practice used in instantaneous mixing applications such as source separation [22].

Sub-band filtering In the case when the filter's impulse response is larger than the window size, using the Fourier Transform of the filter causes cyclic convolution artifacts. These can be reduced by considering *sub-band* filtering. In this case, the convolution operation is applied on each sub-band independently. Each frequency band of the STFT of the filtered signal is assumed to be computed from the convolution of the corresponding band of the STFT of the input signal $\boldsymbol{S} \in \mathbb{C}^{F \times T_s}$ and a line of the matrix $\boldsymbol{H} \in \mathbb{C}^{F \times T_f}$ representing the sub-band convolution kernel, as:

$$\boldsymbol{Y}_{f,t} = \sum_{\substack{\tau=0 \\ t-\tau \geq 0}}^{T_f} \boldsymbol{H}_{f,\tau} \boldsymbol{S}_{f,t-\tau}. \tag{2.11}$$

Sub-band filtering is illustrated in Figure 2.2. This approximation is however inexact. Indeed, the STFT windowing doesn't allow for perfect low-pass filtering, which causes spectral leakage between the frequency bands. Hence, filtering each sub-band

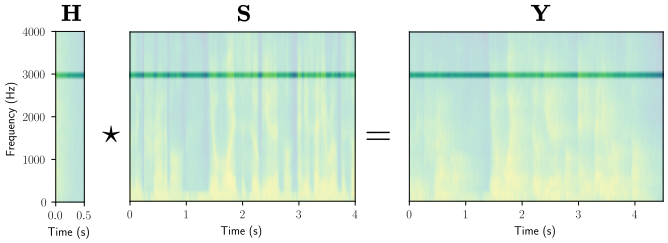


Figure 2.2 – Illustration of sub-band filtering.

independently ignores the fact that STFT bands are correlated, Originally designed for AEC [60], the sub-band filtering method is now widely used in autoregressive reverberation modeling for dereverberation, as we will see in Section 3.3.2.

Sub-band filtering approximations can be corrected by considering cross-band terms [61]. This enables to compute the exact convolution in the STFT domain [62].

Cross-band filtering In the noiseless case, the linear time-invariant filtering in time-domain in (2.6) can be equivalently expressed in the short-time Fourier transform (STFT) domain using an inter-frame and inter-band convolution operator denoted $\mathcal{C}$, defined in [62] as:

$$\boldsymbol{Y}_{f,t} = \mathcal{C}(\boldsymbol{S}, h)_{f,t} \triangleq \sum_{f'=0}^{F-1} \sum_{t'=0}^{\min(t;T_h)} \mathcal{H}_{f,f',t'} S_{f',t-t'}, \tag{2.12}$$

where $\boldsymbol{Y} \in \mathbb{C}^{F \times T_y}$ denotes the STFT of the filtered signal, $\boldsymbol{S} \in \mathbb{C}^{F \times T_s}$ is the corresponding STFT of the dry signal, and $\mathcal{H} = \{\mathcal{H}_{f,f',t}\}_{f,f',t=0}^{F-1,F-1,T_h-1} \in \mathbb{C}^{F \times F \times T_h}$ represents the time-frequency convolution kernel induced by the RIR, capturing both spectral and temporal spread. The convolution kernel $\mathcal{H}$ is derived from the time-domain RIR $h \in \mathbb{R}^{N_h}$ by [62]:

$$\mathcal{H}_{f,f',t'} = \sum_{m=-N+1}^{N-1} h[t'L - m] W_{f,f'}[m], . \tag{2.13}$$

where N denotes the STFT window length, L the hop size, and

$$W_{f,f'}[m] = \frac{1}{N} \sum_{n=0}^{N-1} g_s[n+m] g_a[n] e^{\frac{j2\pi(f'(n+m)-fn)}{F}} \tag{2.14}$$

Here, g_s and g_a denote the synthesis and analysis window functions, respectively. Cross-band filtering is illustrated in Figure 2.3. The formulation of Eq. (2.12) enables the integration of time-domain models of reverberation into time-frequency processing frameworks, which is particularly advantageous for both model-driven and data-driven

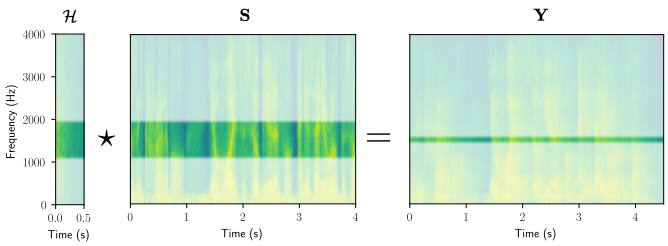


Figure 2.3 – Illustration of cross-band filtering.

approaches.

This model can be equivalently formulated as a matrix product, by introducing the concatenation of the $2F'$ neighbouring frequency bands of $\mathcal{H}_{f,f,t'}$ and Toeplitz matrices of $\boldsymbol{S}_{f,t}$. Each frequency band f of $\boldsymbol{Y}$ can then be computed as:

$$\boldsymbol{Y}_f = \bar{\boldsymbol{S}}_f C_f. \tag{2.15}$$

where $\bar{\boldsymbol{S}}_f$ is the column-wise concatenation of the Toeplitz matrices $\mathcal{T}\big(S_{f'}\big)$ constructed from frequency bands $f' = f - F', \ldots, f + F'$ of the dry STFT coefficients:

$$C_f \triangleq \left[\mathcal{H}_{f,f'}^{\mathsf{T}}\right]_{f'=f-F'}^{f+F'} \in \mathbb{C}^{(2F'+1)T_h}, \tag{2.16}$$

$$\bar{\boldsymbol{S}}_f \triangleq \left[\mathcal{T}\big(S_{f'}\big)\right]_{f'=f-F'}^{f+F'} \in \mathbb{C}^{T_y \times (2F'+1)T_h}. \tag{2.17}$$

Note that sub-band filtering is a special case of cross-band filtering with $F' = 0$ cross-bands.

2.2.2 Cross-correlation

Another important operator for signal processing is the *correlation* operator. The cross-correlation operation of two sequences s and h, denoted $\circledast$, is defined as:

$$(s \circledast h)[n] \triangleq \sum_{k \in \mathbb{Z}} s[k]^* h[k+n], \tag{2.18}$$

where $\cdot^*$ denotes complex conjugation. The cross-correlation operator is linear but neither commutative nor associative. As the convolution between time-reversed and conjugated s with h, the following property holds:

$$(s \circledast h) = \mathrm{IDFT}\left(\mathrm{DFT}(s)^* \odot \mathrm{DFT}(h)\right) = \mathcal{T}(s^*)^{\mathsf{T}} h, \tag{2.19}$$

where $\cdot^{\mathsf{T}}$ denotes matrix transposition.

2.2.3 Properties of filters

In order to be consistent with physics, realistic filters have to respect two conditions: *causality* and *stability*. A causal filter is characterized by its impulse response having only zero-valued negative times. Moreover, the convolution of a causal signal with a causal filter is always causal. From a physical point of view, a system being causal means that it cannot forsee the future. A stable filter (in the Bounded-Input, Bounded Output sense) is a filter that, for any bounded input signal, outputs a bounded signal. A filter is said to be *Minimal-phase* if it is causal and stable and its inverse is causal and stable. One characteristic of minimum-phase filters is that they exhibit the shortest possible group delay and concentrate their energy as early as possible in time. Mathematically, a minimum phase filter's zeroes are all inside the unit circle.

A key property of minimum-phase filters is that their phase response is uniquely determined by the magnitude response. Given a filter's magnitude response $|H[f]|$, and if $\forall f, H[f] \neq 0$ (the filter has no zeroes on the unit circle), a minimal-phase filter can be computed using the Hilbert transform, by setting the angles $\angle H$ to:

$$\angle H = -\mathscr{H}\left(\log|H|\right), \tag{2.20}$$

where $\mathscr{H}$ is the Hilbert transform, that can be computed for a signal x using [63, Section 7.3, p357]:

$$\mathrm{DFT}(\mathscr{H}(x))[f] = -j\,\mathrm{sign}(f)\,\mathrm{DFT}(x)[f], \tag{2.21}$$

$$\text{with } \mathrm{sign}(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x = 0 \\ -1 \text{ if } x < 0 \end{cases}.$$

2.2.4 Analytic signal and envelope

The Hilbert transform of x produces a 90° shift of the negative frequencies of the DFT of x and a −90° shift of its positive frequencies. This can be used to compute the *analytic signal* associated with x, defined as the complex-valued signal that contains no negative frequencies of x. The analytic signal of x is:

$$x + j\mathscr{H}(x). \tag{2.22}$$

The *envelope* of a signal x is the magnitude of the analytic signal associated to x.

2.3 Optimization for audio inverse problems

In this Section, we introduce some mathematical tools for solving audio inverse problems. A first class of convex problems is defined, then non-convex problems are introduced along with regularization and data-driven techniques to solve them. Examples are provided for the non-blind deconvolution problem.

In this manuscript, we focus on continuous optimization problems on a finite dimension $\mathbb{K} = \mathbb{R}^D$ or $\mathbb{C}^D$. These problems can be written as finding $x \in \mathbb{K}$ such that it is in the set of the minimizers of a *loss* function $\mathcal{L}$, under some constraints $(c_i)_{1 \leq i \leq m}$:

$$x \in \underset{x \in \mathbb{K}}{\operatorname{argmin}} \mathcal{L}(x) \tag{2.23a}$$

$$\text{s.t. } c_i(x) \leq 0 \text{ for } 1 \leq i \leq m. \tag{2.23b}$$

2.3.1 Convex optimization

When the functions $\mathcal{L}$ and $(c_1, \dots c_m)$ are convex, the problem is said to be a convex optimization problem. A thorough introduction on convex optimization can be found in [64]. An important property of convex optimization problems is that their set of solutions is convex. If, moreover, the loss is strictly convex, then this set of minimizers is a singleton. For instance, two widely-used convex loss functions for inverse problems in signal processing are the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) losses. The main difference between these two opitmization objectives lies in the way they penalize errors: the MSE enphasizes large errors, while the MAE treats error linearly. Moreover, under assumptions of Gaussian additive noise, the MSE solution corresponds to the Maximum Likelihood (ML) estimator.

Common optimization techniques to solve convex problems where the loss function is differentiable are based on the gradient method, which consists in iteratively defining the sequence $\left(x^{(i)}\right)_{i \in \mathbb{N}}$ using the gradient of the loss function denoted $\nabla_{\mathcal{L}}$ as:

$$x^{(i+1)} = x^{(i)} - \eta^{(i)} \nabla_{\mathcal{L}}(x^{(i)}), \tag{2.24}$$

where $x^{(i)}$ designates the variable x at iteration i and $x^{(0)}$ designates the initialization of the sequence. The *step size* $\eta \in \mathbb{R}_+$ has to be defined, and needs to be precisely tuned for performant gradient descent [65, 66].

The optimal step size can be computed as:

$$\eta^{(i)} = \underset{\eta \in \mathbb{R}_+}{\operatorname{argmin}} \mathcal{L}\left(x^{(i)} - \eta \nabla_{\mathcal{L}}\left(x^{(i)}\right)\right). \tag{2.25}$$

One of the most encountered optimization problems is the linear least-squares

approximation [64]. It is an unconstrained problem in the form:

$$\hat{x} = \operatorname{argmin} \|\boldsymbol{A}x - b\|_2^2 , \tag{2.26}$$

where $\boldsymbol{A}$ is a matrix and b a vector.

If $\boldsymbol{A}$ has a full column rank, a closed-form of the solution can be computed as:

$$\hat{x} = \boldsymbol{A}^{\dagger} b, \tag{2.27}$$

where $\boldsymbol{A}^{\dagger} = \left(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathsf{T}}$ is the Moore-Penrose pseudo-inverse of the matrix $\boldsymbol{A}$.

For convex optimization problems of the least-square form, it can be interesting to measure to what extent the solution $\hat{x}$ is robust to perturbated measurements b. The condition number provides such a quantification, and, for the least-squares approximation problem, is equal to [67, p341]:

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\|_F \left\|\boldsymbol{A}^{\dagger}\right\|_F \geq 1, \tag{2.28}$$

where $\|\boldsymbol{A}\|_F$ denotes the Frobenius norm of matrix $\boldsymbol{A}$. The condition number is real-valued and always greater or equal than 1. If $\kappa(A) \gg 1$, the problem is said to be ill-conditioned.

Application to deconvolution

Given a source $s \in \mathbb{R}^N$, and a measurement $y \in \mathbb{R}^N$, the non-blind deconvolution problem aims to retrieve the impulse response $h \in \mathbb{R}^N$, such that $y = s \star h$. It can be formulated with the least-squares norm as:

$$\operatorname*{argmin}_h \|\mathcal{T}(s)h - y\|_2^2 . \tag{2.29}$$

This problem is an instance of the least-squares approximation problem defined in (2.26) where a particular structure is imposed on the measurement matrix. It has been proven that, in the case of a Toeplitz structure, the condition number of the deconvolution problem is related to the eigenvalues of $\mathcal{T}(s)\mathcal{T}(s)^{\mathsf{T}}$, which in turn are related to the Fourier transform of s denoted S [68]. More precisely, the condition number of the problem described in Eq. (2.29) is asymptotically (when the length of y, s tend to infinity) equal to [69]:

$$\kappa(s) = \frac{\max_f |S(f)|^2}{\min_f |S(f)|^2}. \tag{2.30}$$

This condition number explains why deconvolution for acoustic system identification is often considered after a pre-whitening step [70].

These results can be generalized to the full- and sub-band models, in which case the following least-squares problem has to be solved for each STFT band [62]:

$$\underset{C_f}{\operatorname{argmin}} \left\| \bar{\boldsymbol{S}}_f C_f - \boldsymbol{Y}_f \right\|_2^2, \tag{2.31}$$

where $\bar{\boldsymbol{S}}_f$ and C_f are defined in Eq. (2.17). This technique will be referred to as *Full-band* or *sub-band* deconvolution in the remaining of this manuscript, depending on the number of cross-band filters F' considered.

In order to solve ill-posed but convex problems, various regularization techniques can be leveraged.

2.3.2 Regularization

In some cases it is useful to combine two optimization objectives in a single optimization problem. The problem is then written as:

$$\underset{x \in \mathbb{K}}{\operatorname{argmin}} \mathcal{L}_1(x) + \lambda \mathcal{L}_2(x), \tag{2.32}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are two loss functions, and $\lambda > 0$ is a regularization parameter, used to balance both objective functions. Lambda can be interpreted as the Lagrange multiplier associated with the constraints $c_1(x) = \mathcal{L}_2(x) \leq 0$. Tikhonov's regularization considers the case where $\mathcal{L}_2(x) = \|\boldsymbol{B}x\|_2^2$, where $\boldsymbol{B}$ is a matrix. A popular instance of Tikhonov's regularization considers $\boldsymbol{B}$ to be the identity matrix, and is coined as ridge regression. Another regularization technique, called LASSO, enforces sparser solutions using the ℓ_1 norm.

In the case when the regularization function is not differentiable, proximal gradient descent can be employed [71]. It consists in alternating minimization of the proximal operator

$$\operatorname{prox}_{\lambda \mathcal{L}_2}(x) \triangleq \underset{y}{\operatorname{argmin}} \lambda \mathcal{L}_2(y) + \frac{1}{2} \|y - x\|_2^2, \tag{2.33}$$

and the gradient step for $\mathcal{L}_1$. The Fast Iterative Shrinkage-Thresholding Algorithm [72] has been developed to efficiently solve ridge-regularized problem.

Another way to solve regularized problems, is to introduce variable splitting, that is, reformulating Eq. (2.32) as:

$$\underset{x,z \in \mathbb{K}^2}{\operatorname{argmin}} \mathcal{L}_1(x) + \mathcal{L}_2(z) \text{ s.t. } x = z. \tag{2.34}$$

This problem can be solved using the ADMM algorithm [73], consisting in iteratively optimizing x by minimizing $\mathcal{L}_1$, z by minimizing $\mathcal{L}_2$, and the distance between x and z. This method enables more complex loss functions to be employed [74], including

DNN-based regularizers [48], under the name of Plug-and-Play methods.

Application to deconvolution

In the case of deconvolution, a popular variant of Tikhonov's method can be defined in the form:

$$\underset{h}{\operatorname{argmin}} \left\| (\mathcal{T}(s) + \lambda I)\, h - y \right\|_2^2 , \tag{2.35}$$

where λ can be interpreted as the regularization constant, to balance the data-fidelity and the robustness of the solution. Solving this equation can be performed in an efficient manner using the DFT and pointwise division denoted $\oslash$, as [75]:

$$\hat{h} = \mathrm{IDFT}\left(\mathrm{DFT}(y) \oslash (\mathrm{DFT}(s) + \lambda)\ \right). \tag{2.36}$$

In the remaining of this manuscript, this method will be referred to as *Fourier deconvolution*. This form is close to the one obtained by Wiener deconvolution, which minimizes the quadratic error in the Fourier domain assuming a known spectral density of additive noise $N = \mathbb{E}(|\mathrm{DFT}(\varepsilon)|^2)$, and a known source spectral density $S = \mathbb{E}(|\mathrm{DFT}(s)|)$.

$$\hat{h} = \mathrm{IDFT}\left(\mathrm{DFT}(y) \odot S^* \oslash \left(S^2 + \frac{1}{\lambda^2}\right)\right), \tag{2.37}$$

where λ is the signal to-noise ratio $\|S\|_2 \,/\, \|N\|_2$. This illustrates the relationship between regularization and robustness, as noisier scenarios require stronger regularization to be solved accurately.
Finally, some approaches for non-blind deconvolution can be considered under the plug-and-play framework [76, 77], and will be explicited in Section 3.2.2.

2.3.3 Non-convex problems

When one of the loss function or the feasible set of an optimization problem is not convex, the solution of an optimisation problem cannot be determined using gradient descent alone, as they would converge to local minima which might not be global. Various methods can be leveraged for solving such problems. We mention only two of them as they are used in audio signal processing: genetic algorithms and deep learning.

Genetic Algorithms consist in mimicking natural selection [78], to make sets of solutions evolve over the course of the optimization using crossover, mutation and selection operations. These algorithms have been used in order to tune filter parameters to match a measured acoustic system [79], a task that will be detailed in Section 3.2.1. Nowadays, most non-convex optimization problems call for solutions involving deep learning methods, especially in the context of audio signal processing.

Deep learning extends classical optimization-based modeling by parametrizing complex nonlinear mappings through deep neural architectures. In the supervised learning setting, the goal is to learn a function $f_w : \mathcal{X} \to \mathcal{Y}$, parameterized by weights (and biases) w, that maps an input $x \in \mathcal{X}$ to an output $y \in \mathcal{Y}$. Given a training dataset $D = (x_i, y_i)_{i=1}^N$, the weights are estimated by minimizing an empirical loss function that measures the discrepancy between predicted and target outputs:

$$\mathcal{L}(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_w(x_i), y_i), \tag{2.38}$$

where ℓ denotes a sample-wise loss function, such as MSE for regression. Optimization proceeds by gradient-based updates, iteratively adjusting w to minimize $\mathcal{L}(w)$. Given the typically large size of datasets, Stochastic Gradient Descent (SGD) allows for training using small batches of data, with *backpropagation* enabling efficient computation of gradients, even in deep networks.

While this formulation defines the essence of supervised learning, many deep learning approaches differ in how the supervision is defined or in the modeling philosophy underlying f_w. In contrast to supervised training, no explicit targets are needed in unsupervised training. Instead, the network is trained to capture latent structures or statistical regularities within the data distribution itself. A specific case of unsupervised learning strategies is when they are combined with so-called generative approaches. Generative approaches aim to model the joint distribution of their input and output, unlike discriminative approaches which aim to only learn the output distribution conditioned to their input. Most neural generative approaches for audio inverse problems learn the distribution of dry speech signals. Popular frameworks for generative unsupervised learning involve variational autoencoders [80], or diffusion models [81]. Such models can synthesize new data samples or latent representations consistent with the clean audio statistics.

Note that in this manuscript, we also consider a supervised setting where the dimension of the supervision labels is much lower than the typical dimension of an audio signal. This setting is called *weak supervision*.

The degree of supervision and the modeling philosophy (generative or discriminative) define a landscape of deep learning methods. Supervised discriminative training remain the most direct and widely used framework for audio dereverberation when paired data are available. These methods will be detailed in Section 3.3.

2.4 Conclusion

In this chapter, we have introduced the time, frequency and time-frequency representations of signals that are commonly used in both model-driven and data-driven dereverberation and room acoustics modeling. Filtering has been presented, and cor-

responds to the operation of applying reverberation models that will be introduced in Section 3.1. We have detailed how this process can be implemented through several approximations of convolution. The same approximations are at the core of acoustic system identification and deconvolution methods that will be summarized in Section 3.2 and 3.3.2 respectively. Then, the data-driven and model-driven paradigms have been analyzed under the light of optimization techniques. These techniques lay the foundation for understanding our contributions. Indeed, convex optimization and regularization techniques will be directly leveraged in our contributions of Chapter 4, as they will enable to define physically consistent loss functions.Finally, non-convex optimization and especially deep learning training approaches set the stage for the unsupervised approaches we develop in Chapter 5.

Chapter 3

Literature review

Chapter abstract

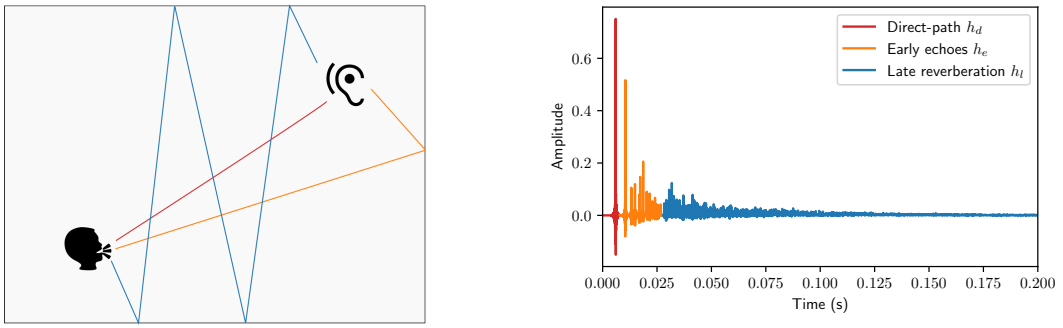
In this Chapter, we present an overview of methods used for the three joint fundamental tasks linked to this thesis: representing the acoustic effects of a room using reverberation, identifying reverberation parameters using acoustic system identification, and mitigating acoustic effects using dereverberation.

In this Chapter, we present a literature review of methods for reverberation modeling, acoustic system identification and dereverberation. The structure of this study is meant to mirror the main computational blocks shared across all of our contributions. Namely, we first present reverberation models, then acoustic system identification, and finally dereverberation. In Section 3.1, we explicit some reverberation models that will be used in our contributions for dataset creation, reverberation modeling and synthesis, and analysis of acoustic characteristics. Then, in Section 3.2, we present methods to determine characteristics of the reverberation process when different types of information are available. We define the analysis-synthesis framework for acoustic system identification in both blind and non-blind settings. Finally, in Section 3.3, we introduce a typology of dereverberation techniques based on the assumption they make about a reverberation model.

As our contributions are only in the context of single-source recordings using a single channel, we restrict our analysis to monophonic signals.

3.1 Reverberation models

In this section, a short glimpse of reverberation models is provided. A more in-depth review of classical reverberation models has been made by Välimäki [82]. After presenting the room impulse response and specific characteristics of the reverberation process, we define three classes of reverberation models. Physical reverberation models naturally stem from studying sound wave propagation. Then, statistical reverberation



(a) Schematic view of the reverberation process.

(b) Room Impulse Response example. (Adapted from [22])

Figure 3.1 – Schematic view of the reverberation process and its associated Room Impulse Response. The direct path is represented in red, early echoes in orange, and late reverberation in blue.

models to derive asymptotic properties of physical reverberation processes Finally, artificial models provide filter-based implementations of reverberators.

3.1.1 Room Impulse Response

In the case where the acoustic conditions are linear and time-invariant, for instance when the sound source and microphone are static during recording, reverberation can be modeled as a filter, characterized by its Room Impulse Response (RIR).

While non-linearities can exist in the context of wave propagation due to the absorption of the walls being nonlinear [83], this manuscript considers a linear behaviour of acoustic processes. This assumption holds as disturbances caused by sound propagation are of small amplitudes [84, Chap. 11], and is often made in practice [85, 86].

The RIR uniquely characterizes the acoustic reverberation process between a source and a microphone within a room. For the sake of clarity, we remove the dependency of the RIR to source and microphone positions, and denote its discretization $h[n]$. Convolution by an RIR can be interpreted as an Auto-Regressive (AR) process, and justifies AR modeling for dereverberation that will be explicited in Section 3.3.2.

Figure 3.1b illustrates an example of such an RIR. It can be decomposed into three components: after an initial delay, corresponding to the time that the acoustic wave takes to travel from the source to the microphone, a first impulse corresponds to the *direct-path* response h_d. It is followed by sparse echoes corresponding to *early echoes* h_e, representing the first-order reflections of the sound against the walls. These echoes then become more and more dense, as first reflections themselves encounter again new walls and acoustic objects, to form *late reverberation* h_l.

In this manuscript, we formulate reverberation as the sum of a direct path, early

reflections and late reverberation, as:

$$h = h_d + h_e + h_l \tag{3.1}$$

where h_d, h_e and h_l are the direct path, early echoes and late reverberation respectively. It is often considered that the supports of h_d, h_e and h_l are disjoint and that the transition from early to late reverberation corresponds to a shift from deterministic to stochastic reverberation modeling. Criteria to define the validity domain of stochastic models include echoes density [87], mean free path of the room [88], or higher order statistics of the reverberation such as kurtosis [89]. This formulation also corresponds to hybrid approaches for reverberation modeling which use two distinct models for early and late reverberation [90, 91, 92]. In practice, for single-channel reverberation, the initial delay of the RIR is often ignored for both tasks of reverberation modeling [93, 94, 95] and dereverberation [36, 96] as it only causes a temporal misalignment which is unperceptible in subjective evaluation but causes a notable degradation of most objective metrics.

Complementary to the time-domain representation, the RIR can be better analyzed using key characteristics, presented below.

3.1.2 Reverberation characteristics

Reverberation characteristics enable to derive meaningful acoustic properties from reverberation, accounting for the specific structure of the RIR. In this section, we present a few of them. Note that the mean-squared error between those characteristics is also often used as a metric to evaluate the performance of reverberation models.

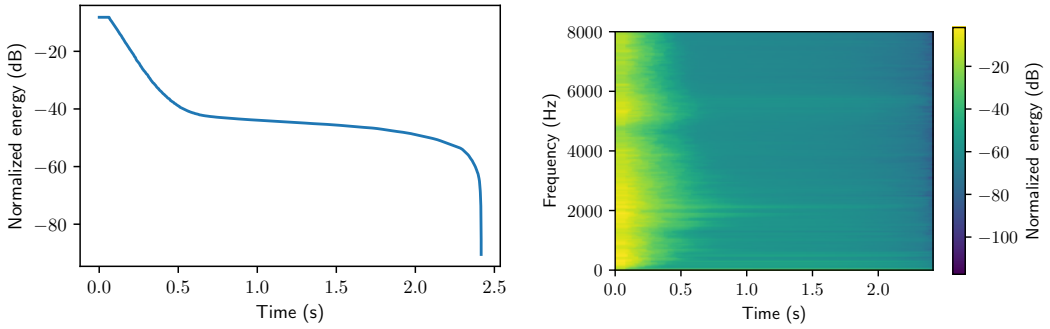
Energy ratios

A key descriptor of room acoustics is the Direct to Reverberant Ratio (DRR), which quantifies the energy balance between the direct path and the reverberant tail. As defined in [28, Sec. 2.4.2], the DRR of the RIR h is given by:

$$\mathrm{DRR}(h) = 10 \log_{10} \left(\frac{\sum_{n=0}^{N_d} h^2[n]}{\sum_{n=N_d+1}^{\infty} h^2[n]} \right) \mathrm{dB}, \tag{3.2}$$

where N_d represents the end of the direct path, commonly fixed at approximately 2.5 ms after the initial RIR peak, corresponding to 40 samples at a sampling rate of 16 kHz [97].

The DRR depends on both the distance between source and microphone and the reflective properties of the environment, and is widely used as a quantitative measure for evaluating and modeling reverberant conditions. Other characteristics of the RIR such as the clarity index can be expressed in a similar manner as the DRR, by changing



(a) EDC of an RIR

(b) EDR of an RIR

Figure 3.2 – Energy Decay Curve and Energy Decay Relief of a measured Room Impulse Response.

the value of N_d. For instance, the clarity index considers N_d corresponding to 50 ms after the direct-path [1].

Energy decay

As seen in Figure 3.1b, measured RIRs exhibit a key property: their late reverberation is exponentially decreasing in amplitude. Introduced by Schroeder [98], the Energy Decay Curve (EDC) captures the cumulative energy remaining in the RIR beyond index n and is defined as:

$$\mathrm{EDC}(h)[n] = \sum_{k=n}^{+\infty} h[k]. \tag{3.3}$$

Dividing the EDC by the total energy of the RIR and plotting it in the dB scale enables to observe how the exponential decrease of the RIR energy over time becomes more interpretable and linear. An example is provided in Figure 3.2a. The EDC starts by a flat line at 0 dB corresponding to the initial delay of the RIR, followed by a linear energy decrease. Finally the noise floor appears as an almost flat line [99, 100].

A counterpart of the EDC for time-frequency distributions can be defined using the Page distribution [57] and is coined Energy Decay Relief (EDR). For faster computation, Jot [101] proposes to estimate the EDR from the spectrogram time-frequency representation as it is easier to manipulate despite some approximations [58, chap5]. Another approximation of the EDR as a band-passed EDC can be formulated [102]. Figure 3.2 presents the EDC and EDR of the RIR presented in Figure 3.1b side by side.

Reverberation time

Another key characteristic can be considered: the Reverberation Time (RT_{60}), is defined as the time it the sound pressure level takes to decrease by a factor of 1000, corresponding to an energy reduction of 60 dB. While first attempts by Sabine [103, 1] to compute the reverberation time were empirical, the relationship between RT_{60} and the absorption of the walls and room volume has been demonstrated in several reverberation frameworks [104, 105].

The RT_{60} can be measured from the EDC, by computing the slope of the energy decay region of the EDC using linear regression on the domain where the EDC decreases [100] or non-linear regression [106, 93] accounting for RIR measurement noise. It has been shown that the reverberation time has an impact on speech intelligibility [107].

After detailing some general characteristics of reverberation, we briefly present some reverberation models that will be used in our contributions.

3.1.3 Physical models of reverberation

Several reverberation models are based on physical features of rooms and consist in solving the wave equation or simulating the propagation of sound rays. An overview of such models can be found in [108].

As room acoustics results from solving the wave equation within enclosed spaces, immediate approaches for reverberation modeling took the form of numerical solvers for the wave equation. These techniques are coined *Wave-based* reverberation models. Examples include Finite Difference Time-Domain methods for discretizing time and space [109, 110], Finite-Element methods for space and frequency discretization [111], and Boundary-Element methods, which focus on discretized boundary conditions on the room walls [112]. In wave-based methods that discretize space, the bandwidth of simulated RIRs is bounded by mesh resolution. Geometry-based methods alleviate this issue by modeling acoustic rays [1].

The Image Source Method (ISM) and its variants provide better results than wave-based methods at higher frequencies. First introduced for rectangular rooms [113], then extended for arbitrary polyhedra [114], it is based on the modeling of direct sound rays from the microphone to virtual sources, illustrated in Figure 3.3. For each wall reflection of a sound ray that leads to the microphone, a virtual source is constructed by mirroring the actual source against the reflection wall. Virtual sources are then recursively created by mirroring lower-order sources. Then, the contribution of each source is delayed according to the length of the direct path from source to microphone, and filtered to model the frequency-dependent absorption of each virtual wall on its path. Finally, the contribution of all sources are summed to produce an RIR. The ISM method has been also popularized thanks to a powerful implementation

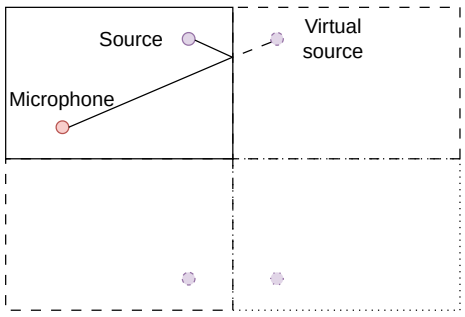


Figure 3.3 – Image source method on a rectangular room. The original room and source are represented by plain lines, first-order virtual sources by dashed lines, and second-order virtual sources (constructed by symmetry of a first-order virtual source) by dotted lines.

named `Pyroomacoustics` [115], enabling the wide-spread usage of this method to create reverberation datasets [116].

Considering the delay of each source's contribution to be slightly varied (using a Gaussian virtual impulse instead of a Dirac), allows this process to be modeled in a differentiable manner [117, 118]. This model has been also extended by considering sources randomly scattered across the room [119]. Furthermore, stochastic variants of the ISM are used in computationally-efficient implementations [116].

The accuracy of ISM methods is however limited by the number of virtual image sources which corresponds to the order of reflections along the wave path. Ray-tracing [120] allows for a higher number of reflections, limited only by computational time. Extensions of this method consider the reflection to be diffuse, in a deterministic, or stochastic manner. On one hand, deterministic diffuse reflections have been implemented in the form of volumetric cones of reflection [121] in the so-called beam-tracing method [122] or diffuse reflection patches in the radiance-transfer method [123]. On the other hand, stochastic ray-tracing consists in randomly sampling reflected rays. Stochastic ray-tracing allows to model diffuse reflections [124], or diffraction of objects within a room which are necessary to render a perceptually realistic reverberation [90].

Some physical approaches to simulate reverberation combine several modeling techniques, to balance each method's strengths and defaults: either by compensating the high-frequency inaccuracy of wave-based methods using ray-tracing [125, 126], or combining image-source and ray-tracing [127] methods.

While wave-based, image-source and ray-tracing based methods for reverberation modeling consider a deterministic source position, some statistical models consider the source to be positioned at random and derive properties of late reverberation. They are detailed hereunder.

3.1.4 Statistical models

Stochastic reverberation models were first motivated by the study of ergodic properties of rooms [128, 88], or a random distribution of scatterers [129]. Most of these results have been recently unified by Badeau under the Statistical Wave Field Theory (SWFT). The SWFT assumes that the source position is located at random and focuses on the modeling of late reverberation at relatively high frequency. Several approaches to the SWFT have been considered, first in the case of ergodic rooms [105], then non-ergodic and highly symmetric rooms [130] allowing to draw a connection to the ISM method [119].

Furthermore, the SWFT unifies and generalizes one of the broadly used reverberation models: Polack's model [87]. This model states that the reverberant tail of an RIR can be modeled as an exponentially decaying stochastic process, confirming previous empirical observations [90]. Specifically, the late reverberation component h_l is defined at index n as:

$$h_l[n] = b[n]e^{-n/\tau}, \tag{3.4}$$

where $b[n] \sim \mathcal{N}(0, \sigma^2)$ denotes a zero-mean white Gaussian noise process, and the decay constant τ is related to the reverberation time RT_{60} and the sampling frequency f_s by:

$$\tau = \frac{\mathrm{RT}_{60} f_s}{3\ln(10)}. \tag{3.5}$$

This statistical model provides a way to simulate reverberant environments using only a small set of physically meaningful acoustic parameters albeit not considering any frequency-dependant reverberation parameter. Frequency-dependant reverberation parameters can still be obtained in the discrete implementation of the SWFT [131].

Furthermore, combining stochastic models of late reverberation with deterministic models of early echoes such as the ISM allows to model the full RIR [91].

3.1.5 Artificial models

Artificial reverberation models encompass a broad range of techniques designed to synthesize reverberation, from traditional filter-based approaches to modern neural, data-driven methods. They are regrouped under the same class as they are meant to provide efficient and real-time compatible implementations of reverberators.

The stochastic position of the source for modeling late reverberation has inspired models considering reverberation as a filtered random sequence. For instance, in the context of the SWFT, the RIR can be represented as a filtered Poisson noise [131]. Other representations model RIRs as a recursively delayed and attenuated finite short random impulse train [132, 133]. Studies of the distribution of the repeated random sequence outlined the interest of velvet noise [134] which density can vary throughout the response [82] in so-called Filtered Velvet Noise (FVN) reverberators.

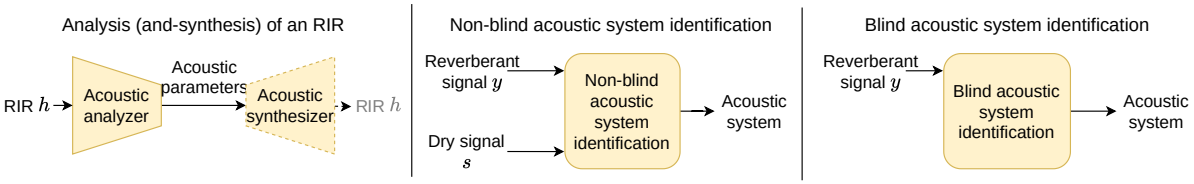


Figure 3.4 – Various acoustic system identification scenarios.

Using a filter to model reverberation is however not new. Schroeder introduced in [135] the use of delay and gain (or comb) filters as a building block for reverberators. He built the first model of reverberation by chaining interlaced comb filters in series to make the first all-pass reverberant filter. Such reverberators have been further expanded into Feedback Delay Networks (FDNs), where parallel comb filters are mixed by a gain matrix. Furthermore, FDNs can be connected to stochastic models [136], physical models under the digital waveguide mesh framework [137] or ray-tracing techniques [138].

Finally, approaches considered deep-learning-based synthesizers of RIRs, where the stochastic nature of reflections was the result of the generative nature of the RIR sampler. They can be conditioned on high-level acoustic properties (RT_{60}, DRR) [139], spatial information [139, 140, 141, 142] or an RIR from the same room [143].

In the next section, we will see that determining an acoustic model from a reverberant signal can be interpreted as a form of acoustic system identification.

3.2 Acoustic System identification

After detailing the tasks of analyzing and reconstructing an RIR using an artificial filter model, we now shift our attention to acoustic system identification, which aims to determine the characteristics of an acoustic system using a reverberant recording. Several scenarios can be identified, and are detailed in the next sections. While the RIR uniquely characterizes the reverberation process, determining geometric properties or filter parameters from a measured room impulse response remains a challenge. In Section 3.2.1, we present the analysis-synthesis framework that can be considered to determine acoustic properties, and that will also be leveraged in non-blind and blind acoustic analysis. Then, in Section 3.2.2, we detail some adaptation of deconvolution techniques to the specific case of acoustic system identification. As this setting requires both the reverberant signal and the dry source, we coin it as *non-blind* acoustic system identification. Finally, in Section 3.2.3, we focus on how acoustic systems can be determined in a *blind* setting, that is when only a reverberant signal is provided as input to the acoustic identification system. A diagram of the three acoustic system identification scenarios is provided in Figure 3.4.

Note that we purposefully omit numerous methods for acoustic system identifi-

cation using image or mesh information, as these multimodal approaches are out of the scope of our contributions. We first focus on acoustic system identification, as the task of source system identification, commonly coined as dereverberation, will be presented in the next Section 3.3.

3.2.1 Analysis (and synthesis) of RIRs

We have seen in Section 3.1.1, that the RIR uniquely characterizes an acoustic system. However, determining geometric properties from RIRs is a notoriously hard task, especially in a mono-channel setting [144]. Hence, focus has shifted to the analysis and synthesis of a room within an RIR model. For instance, the ISM provides a convenient framework to determine geometric properties of rectangular rooms, since it requires only 18 parameters to be jointly estimated [145, 146, 147]. Some methods focus on estimating average parameters of ray-based methods, such as the total volume, which can be computed using diffuse field acoustics [148], the mean absorption using data-driven approaches [149], Other methods focus on room dimensions only [150].

Popular methods can be considered under the analysis-synthesis framework [101], where RIR parameters are analyzed, and then used to resynthesize an RIR following a given model. This approach is particularly popular to fit artificial reverberators to match RIRs measured from existing rooms. For instance, FVNs filters have been optimized using linear predictive coding [151], enabling interpretable reverberation shortening [152]. On the other hand, FDN optimization represents more of a challenge. Their complex parameter space typically requires usage of non-convex optimization techniques such as genetic algorithms [79, 153, 154]. The recently-introduced framework of DDSP [155] is perfectly tailored for this task. It consists in leveraging a differentiable implementation of a synthesis model, whose parameters are determined by an analysis DNN. Differentiable FDN designs have been introduced, where the mixing matrices are learnt [94], or both the mixing matrices and attenuation gains are trained, in order to produce colorless artificial reverberation [156]. These designs also enable to distinguish between position-dependent and position-independent parameters, allowing to perform RIR interpolation [157]. Other notable examples of DDSP for RIR matching include FVNs [94, 158], wave-based methods [159], or ISM [117].

The procedure of synthesizing an RIR matching a given model has led to the introduction of new RIR metrics such as the EDR [101],

3.2.2 Non-blind acoustic system identification

The RIR itself can be obtained from pairs of reverberant and dry signals. In fact, estimating the RIR from a pair of reverberant and dry signals has become the new standard [160], superseding previous methods based on impulsive excitations. Accurate estimation of RIR now consists in deconvolving a measured swept-sine [161]

or computing the cross-correlation of maximum-length sequences [162]. These techniques however require specific dry signals to be used for measurement. Indeed, both swept-sine and maximum-length sequences exhibit a flat frequency spectrum, thus making deconvolution robust to measurement noise, as its condition number defined in Eq. (2.30) remains low. However, when the source signal does not exhibit such a property (for instance in case of speech), or the noise level is too high, regularization techniques can be employed to obtain a more robust estimate. Popular regularization criteria involve sparsity enforced using the nuclear [163], ℓ_1 [164, 165] or ℓ_2 [166] norms. The Analysis-Synthesis framework detailed above in Section 3.2.1 can also be adapted to RIR estimation, for instance by using variational expectation-maximization to estimate physical model inspired by the SWFT [167], or DDSP-like techniques to identify various audio effects [168]. A specific case can also be considered, when RIRs have to be extrapolated to unseen source and microphone locations. Popular techniques then involve regularizing deconvolution using the Wasserstein distance to an RIR measured from a neighboring location [169], RIR dimensionality reduction [170], or deep learning [171]. Finally, when the measurement noise level doesn't allow for precise estimation of the RIR tail, late reverberation can be generated from early echoes using white noise [172] or more recently differentiable FVNs [173, 174].

3.2.3 Blind Acoustic system identification

In the case when the source signal is not known, blind acoustic system identification aims to fulfill two distinct objectives: first approaches aimed only at retrieving RIR characteristics or geometric properties of the room in which a signal was recorded. Retrieving the full RIR has only been tackled later, as it represented a harder task.

Among RIR characteristics, the RT_{60} has received a lot of attention, as it is directly related to sound decay rate. First attempts aimed to compute the minimal decay time across sliding windows of reverberant signals [175, 176]. A popular approach was later designed by Prego [177], that focused on specific regions of the spectrogram to accurately compute the RT_{60}. In this method, free-decay regions of the reverberant spectrogram are first identified based on the fact that they exhibit a strictly-decreasing energy. Then, for each decay region, the sub-band reverberation time is computed via linear fitting of the energy decay curve in this region. Finally, a polynomial mapping is used to match the median sub-band reverberation time to its corresponding time-domain value. This approach has won the Acoustic Characterization of Environments (ACE) challenge in 2015. Later methods for RT_{60} estimation mostly involved deep learning, at the cost of interpretability [178, 179, 180, 181]. Another characteristic that has received attention in the context of the ACE challenge is the DRR. The DRR can be estimated from the ratio of dry signal energy to reverberant signal energy (also coined as SRR), provided the source is spectrally white. This property has been leveraged in data-driven masking methods for DRR estimation [182]. Similarly

to the DRR, clarity can also be estimated from reverberant signal features in an interpretable manner. Its high correlation to ASR performance has been leveraged in a blind estimator [183]. This correlation has inspired a method to disentangle clarity and speech content in reverberant signals [184].

Efforts have also been made to blindly estimate geometric characteristics of the room in which a reverberant signal has been recorded. Notable examples include estimation of the distance between source and microphone [185], or of the room volume [186]. As these estimators are mostly data-driven, it was shown that they require substantial data augmentation [187] to compensate for lack of realism [188]. Blind acoustic parameter estimation performance is also limited by the fact that some acoustic parameters are better measured from an estimated RIR than directly from the reverberant signal [189], justifying the need for blind RIR estimation.

Only very recent work aimed to retrieve the full RIR, thanks to data-driven methods. Such methods fall in two categories: hybrid and purely data-driven. Similarly to non-blind and analysis-and synthesis RIR estimation methods, hybrid blind RIR estimation methods leverage an RIR model in an analysis-synthesis framework. They consist in a classical reverberation model whose parameters are estimated from reverberant signals by a DNN encoder. Popular reverberation models that were trained under the DDSP framework include FDNs [94] or FVNs [158, 94]. When the reverberation effect is not differentiable, as for commercial reverberation effects, data-driven methods can be used to blindly predict a reverberation plugin preset in a supervised manner [190]. On the other hand, purely data-driven methods for blind RIR estimation take the form of generative models. Several architectures have been proposed. Modeling long sequences, which are convolved to form reverberant signals naturally calls for architectures that are able to model an AR process, such as autoregressive decoders [191] or a novel segmental generation procedure [192]. Other enable to distinguish between position-specific and room-specific informations [193]. Some approaches ignore any RIR model [139]. It could be argued that even purely data-driven approaches could be considered in the analysis-synthesis framework, as they compute an RIR representation in a compressed latent space.

3.3 Dereverberation

The task of retrieving the dry source instead of the acoustic system from a reverberant signal is coined as dereverberation.

Assuming fixed source and microphone positions, the monaural reverberant observation y can be modeled as the convolution of the anechoic source signal s with the RIR h, contaminated by additive noise ε. The resulting signal is expressed as:

$$y[n] = (s \star h)[n] + \varepsilon[n], \tag{3.6}$$

where $\star$ denotes the linear convolution operator and n is the discrete time index.

Similarly to the acoustic system identification task, dereverberation can be performed in a non-blind or blind setting, depending on whether the RIR is known. Nevertheless, unlike acoustic system identification models, some dereverberation methods have been designed to use partial information about a reverberation model. Indeed, some methods are based upon an AR model of reverberation, and also perform dereverberation in an autoregressive manner. We denote them as autoregressive dereverberators and detail them in Section 3.3.2. The AR model of reverberation can be enhanced by considering its connection to reverberation characteristics such as the RT_{60}, yielding a new class of acoustic-parameter-driven dereverberation methods presented in Section 3.3.3. Finally, most neural dereverberators use almost no knowledge of any reverberation model whatsoever. In Section 3.3.4, we coin them as reverberation-agnostic models and we try to exhibit assumptions they make about very implicit reverberation models. We conclude by presenting some metrics for speech dereverberation in Section 3.3.5.

3.3.1 Non-blind dereverberation

As acoustic system identification, deconvolution can be performed in a non-blind setting, assuming knowledge of the full RIR. As with acoustic system identification, deconvolving a reverberant speech by an RIR presents some difficulties. Apart from degradations caused by noisy measurements [194], deconvolution by an RIR presents the challenge that if the source which is aimed to be deconvolved moves slightly, the RIR will change drastically [30]. RIRs are also not minimum-phase, so usual inversion of a filter-based model of RIR would not be stable [195, 196]. Popular methods split the RIR into a minimum-phase and an all-pass component to alleviate these problems. For instance, minimum-phase filtering has been combined with neural dereverberators to perform correction of the remaining all-pass filtered signal obtained after minimum-phase filtering [197]. Another application of data-driven methods for non-blind deconvolution consists in plug-and-play algorithms presented in 2.3.2, combined with neural speech priors. These neural speech priors can be diffusion-based [198], or discriminative neural denoisers [77].

3.3.2 Autoregressive dereverberation

A second family of joint source and reverberation models leverage the AR nature of the reverberation and dereverberation process. Early dereverberation models tailored for speech signals considered both speech and reverberation to be AR. Indeed, speech can be modeled according to the source-filter model which inspired linear predictive coding techniques that are robust to reverberation [34]. Contrary to these approaches, the Weighted Prediction Error (WPE) [33] method doesn't assume such autoregressive

nature of speech and instead leverages a complex circular time-varying Gaussian prior for the speech spectrogram. In this model, the reverberant spectrogram is assumed to be constructed by delaying and scaling the dry signal, which characterizes an AR process. The WPE method focuses on estimating the inverse filter of the reverberation process. It formulates the dereverberation problem as a Maximum Likelihood (ML) objective, where the likelihood to observe reverberant speech with respect to the reverberation inverse filter has to be maximized. Unlike previous approaches, the ML objective is weighted by the dry speech variance. The resulting optimization problem is solved by iteratively estimating the reverberation inverse filter, the dry speech signal, and the dry speech variance. The initial dry speech power spectral density, which is required to start the iterative process, was originally approximated as the reverberant speech power spectrum. Hence, neural improvements of the WPE method first aimed to estimate the dry speech power spectral density [199], allowing to make this estimation more accurate for real-time enhancement [200], and in the presence of noise [201]. These methods allow to perform less iterations of the WPE methods, at no cost on performance. While these neural improvements required to be trained using pairs of reverberant and dry data, an estimator of the dry speech power spectral density for WPE could equivalently be trained by backpropagating through the classical WPE iteration, alleviating the need for dry data [202]. Later neural improvements of the WPE method enhanced its results at each iteration under the Plug-And-Play framework. The iterative nature of the WPE algorithm inspired an approach based on deep unfolding [203]. Deep Unfolding models are constructed by replacing iterations of a classical algorithm by DNN layers, in order to combine the performance of data-driven approaches with the interpretability of iterative models [204]. On the other hand, the Plug-and-Play framework offers another interpretable way to combine WPE with learnt priors, focussing only on speech priors [205]. While the plug-and-play method refined the dry speech estimate only, denoising diffusion restoration models can be leveraged to refine the inverse reverberation filter used in WPE, and are trained using dry data only [206]. Meanwhile, the performance of WPE can be increased by creating virtual channels computed from the recorded signal using a DNN [207]. Other improvements of the WPE algorithm enhanced the AR model. They allow for instance to use a dictionary of inverse filters [208] to allow for a more robust dereverberation, or to model the reverberation process in the cross-band model instead of the sub-band approximation (detailed in Section 2.2.1) [209].

While WPE focused on estimating the inverse of the reverberation filter in a process coined as backward linear prediction, recent approaches such as forward convolutive prediction extend this paradigm by modeling reverberation as a forward convolutional process. In the STFT domain, this corresponds to modeling the reverberation as sub-band filtering, and explicitly computing the filter. First approaches used this filtering model in conjunction with classical signal processing methods, such as the

non-negative matrix factorization of speech's time-frequency representation [210]. Another advantage of the forward convolutive prediction is that it enables more accurate and trainable representations aligned with neural architectures. Hence, the forward convolutive prediction has been used in discriminative neural dereverberation trained under the supervision of pairs of dry and wet signals [211] and an unsupervised manner using wet signals only in USDNet [212]. USDNet however shows subpar performance when used in a single-channel setting. The need for supervision using wet signals has been alleviated by leveraging deep generative models of dry signals. Common approaches for dereverberation then combine a pre-trained neural source prior with an AR reverberation model at inference. These priors can take the form of Variational AutoEncodeurs (VAEs) [213] or diffusion models [81]. For VAEs, several methods to sample the dry speech according to a reverberation observation model are based on the sub-band decomposition. They use expectation maximization algorithms to iteratively estimate the clean spectrogram and acoustic parameters. The optimization can be done in the latent space of the VAE using Monte-Carlo Markov-Chain sampling [214], or in the signal space [215].

3.3.3 Acoustic-parameter-driven dereverberation

Autoregressive RIR models used for dereverberation can be vastly improved by introducing some RIR parameters. By leveraging reverberation parameters (blindly estimated using techniques mentioned in Section 3.2.3), dereverberation models have been improved.

The most commonly used reverberation parameter for dereverberation is the reverberation time RT_{60}, as it is particularly meaningful to define the exponential decay of the RIR predicted by Polack's model (defined in Section 3.1.4), and has a strong connection to speech intelligibility [107]. One one hand, several dereverberation methods explicitly define the convolutive model of reverberation using the RT_{60}. Given the RT_{60}, the variance of the late reverberant spectrum can be computed and reused in spectral subtraction methods for dereverberation [216]. This method has been further enhanced by also leveraging the DRR parameter and considering both parameters to be frequency-dependent [217]. Generative speech models have also been combined with autoregressive reverberation models defined using the RT_{60}. They were first based on a Student-T distribution of speech [218], then on diffusion-based pre-trained neural priors for dereverberation [219, 220]. On the other hand, some neural dereverberators considered the RT_{60} without making the autoregressive model of reverberation explicit. They are mostly based on the fact that the RT_{60} characterizes the length of the RIR, and thereby the length of the filter that would represent dereverberation. In the case of DNN, which are non-linear and often not time-invariant, this concept translates to the notion of receptive field, defined as the region of the input space that influences a single unit in the latent space. The receptive field of

a dereverberation neural network can be adjusted using the RT_{60}, by steering the feature extractor at inference [221], or an internal attention layer at training [222]. A less interpretable manner to use the RT_{60} at inference, is to directly connect the latent space of a neural RT_{60} estimator to a data-driven dereverberation model [223]. Another advantage of leveraging reverberation parameters at training is that they can be modified at inference, enabling user-controllable dereverberation. For instance, the balance between dereverberation-induced distortion and incomplete removal of reverberation can be steered by adjusting the DRR induced by a reverberation model [224]. This is especially useful in the context of acoustic matching [225], or even modifications of attributes such as clarity [184].

3.3.4 Reverberation-agnostic dereverberation

The last category of dereverberation uses almost no knowledge of a reverberation model whatsoever, and regroups most neural dereverberators. On one hand, generative models for speech only aim to model the dry speech distribution and thereby cannot make any assumption on a reverberation model. For instance, in the context of diffusion generative models, the noisy signal [226] or the output of a pre-trained neural dereverberation model [227, 228] can be used to start a reverse diffusion process, without taking into account any reverberation model to steer the reverse diffusion. On the other hand, deterministic approaches for data-driven dereverberation are meant to directly represent a mapping from a reverberant to a dry signal. In this case, the choice of the architecture for such DNNs is crucial as it implicitly sets assumptions on reverberant and dry signals. Two factors can be decisive: the DNN architecture and its ability to process phase.

Among the first neural dereverberation architectures, convolutional neural networks represented a popular choice, whether they were used in time-domain [229] or frequency domain [230]. As convolutional layers are translation-invariant, they can be interpreted as a form of non-linear filtering. This remains true in the STFT domain, where a connection to the cross-band filtering model detailed in Section 2.2.1 can be drawn by considering a time- and frequency-invariant filter. Moreover, it was shown that convolutional neural networks are able to implicitly model properties of reverberation. For instance, analysis of the receptive field of convolutional networks working in time-domain showed that it should be increased with respect to the RT_{60} for optimal speech dereverberation [231]. This motivated the design of deformable [232] or multi-resolution [233] architectures to improve robustness to implicit variations in reverberation characteristics. Another popular choice consists in recurrent neural networks such as Long Short Term Memory (LSTM)s, that were primarily considered due to their ability to model temporal dynamics of speech [234]. A more powerful model was later proposed under the name FullSubNet (FSN), that consisted in processing both the sub-bands and the cross-bands of a reverberant STFT recursively [235]. This

principle was then reproduced in TF-Gridnet [36], where intra-band (temporal) and intra-frame (frequency) LSTM processing was followed by a full-band self-attention module. An improvement of TF-Gridnet, named TF-Locoformer (TFL), proposed to leverage attention heads located in both the temporal and frequency modules [236]. This latter architecture showed state-of-the art performance in reverberant speech enhancement, but can hardly be interpreted with respect to a room acoustics model. Such an implicit modeling of reverberation has also been used for data-driven acoustic matching [25, 237, 238] or dereverberation in the latent space of deep learning-based architectures [239].

As most neural dereverberators operate in the complex STFT domain but their weights are real-valued, obvious choices for phase modeling are often not interpretable [55]. Current approaches for phase processing can be divided into phase-sensitive, complex masking, and phase agnostic techniques. Phase-agnostic methods only process the STFT magnitude. Their loss function can either be phase-agnostic as well, when it only considers the STFT magnitudes [230], or phase-aware when in order to take into account the interaction between phase error and magnitudes [240]. The potential improvement that could be obtained by jointly estimating clean magnitude and phase, motivated the use of complex ratio masking, when the mapping between reverberant and dry signals is computed as a complex mask [241]. This approach was implemented for dereverberation in FSN [235], that estimated the complex mask only from the reverberant STFT magnitudes. A similar approach can be found in two-stage processing methods which first enhanced STFT magnitudes, then phase, using classical methods based on STFT consistency [54] or data-driven methods [242, 243]. Finally, phase-sensitive models operate using real-valued projections of the complex STFT on both input and output. A popular choice is to consider real part and the imaginary part as two channels, and has been implemented in state-of-the-art dereverberation models that perform end-to-end phase processing, such as TF-GridNet [36].

3.3.5 Dereverberation metrics

Finally, performance of speech dereverberation methods has been evaluated using the Scale-Invariant Source-to-Distortion Ratio (SI-SDR) [244, 245], Short-Term Objective intelligibility (STOI), Wide-Band Perceptual Evaluation of Speech Quality (WB-PESQ), or Speech-to-Reverberant Modulation Ratio (SRMR) metrics. All metrics except the SRMR are intrusive, meaning that they require the ground-truth (or clean) signal to be computed.

The SI-SDR is a signal enhancement metric based on the time-domain difference between the clean and processed signals. First introduced for the source separation task [244], it can be adapted to dereverberation by considering its single source variant. In the single source-variant, the estimated source $\hat{s}$ is decomposed as the sum of the ground-truth source s, a distortion or artifact term ε_a, and a remaining noise/re-

verberation term ε_n, as:

$$\hat{s} = s + \varepsilon_a + \varepsilon_n \tag{3.7}$$

Several metrics are then introduced: the Source-to-Distortion Ratio (SDR) is defined as:

$$\text{SDR} \triangleq \frac{\|s\|_2^2}{\|\varepsilon_a + \varepsilon_n\|_2^2}. \tag{3.8}$$

The Signal-to-Noise Ratio (SNR) is defined as:

$$\text{SNR} \triangleq \frac{\|s\|_2^2}{\|\varepsilon_n\|_2^2}. \tag{3.9}$$

Note that on unprocessed signals, the distortion term disappears and SDR is equal to the SNR. The SDR metric can be improved, by scaling the estimate $\hat{s}$ such that the residual $\hat{s} - s$ is orthogonal to s. The SI-SDR is then defined as

$$\text{SISDR} \triangleq \frac{\|\alpha s\|_2^2}{\|\alpha s - \hat{s}\|}, \text{ for } \alpha = \underset{\alpha}{\operatorname{argmin}} \|\alpha s - \hat{s}\|_2^2. \tag{3.10}$$

The optimal scaling factor is then $\alpha = \frac{\hat{s}^\mathsf{T} s}{\|s\|_2^2}$. The SI-SDR is expressed in dB.

The STOI [246] is a perceptual metric designed to compare two speech sequences in terms of intelligibility. Computing the STOI involves applying identical processing steps to both the target and the estimated speech signals. First, silent regions of the target signal are identified and removed from both signals to ensure a consistent comparison. Next, each signal is decomposed into one-third octave bands, and the resulting band-limited representations are segmented into frames of 384 ms duration. After applying an appropriate scaling to the processed (noisy or dereverberated) signal, the short-term correlations between the temporal envelopes of the two signals are computed within each 384 ms window. Finally, these correlation coefficients are averaged across all time windows and frequency bands to yield the overall STOI score. The STOI metric has been extended [247], by improving the normalization procedure. The STOI takes values between 0 and 1 , whereas the Extended Short-Term Objective intelligibility (ESTOI) is valued between 0 and 4. Higher scores indicate better performance.

The PESQ [248] is a widely used objective metric for assessing the quality of speech signals, particularly in telecommunications. It compares the perceptual quality of the processed speech signal against a reference signal. It is computed by measuring the distance between auditory transforms of both clean and processed speech. While its relevance and alignment to subjective tests are more and more debated [249], PESQ is still widely used to detect processing artifacts. Originally designed for speech signals measured at 8 kHz, it has been extended for speech sampled at higher frequencies [250]. It is valued between -0.5 and 4.5.

Finally, SRMR [251] is a metric designed to measure the specific degradation caused by reverberation on speech. It is built upon the observation that reverberant speech tends to exhibit more Gaussian white-noise like properties and hence its high-frequency temporal envelope should vary at a higher rate than dry speech. Hence, the SRMR is computed as the ratio between low-frequency and high-frequency speech modulation. As this metric is un-intrusive and requires no clean speech reference, it can be used as an unsupervised target to train data-driven dereverberation models. MetricGAN-U [252] uses a discriminator neural network designed to mimic the SRMR metric in order to train a dereverberation model. While the dereverberation model trained in this setting seems to excel when it is evaluated using the SRMR metric, it shows subpar dereverberation performance when evaluated using other metrics.

3.4 Conclusion

In this chapter, we reviewed existing literature by drawing connections between reverberation modeling, acoustic system analysis, and dereverberation. Several acoustics models have been detailed, such as the ISM used for fast and interpretable physical reverberation modeling and dataset creation, the SWFT and Polack's model as statistical reverberation models for efficient and physically-consistent reverberation modeling, and artificial methods for practical filter-based implementation of reverberators. These artificial models enabled us to draw a connection to acoustic system identification in the widely-used analysis-synthesis framework that has been adapted to various scenarios, such as non-blind and blind acoustic system identification. Finally, we introduced a typology of dereverberation methods, outlining the various ways they leverage reverberation models.

This analysis also provides an insight about how our contributions can benefit to both tasks of acoustic analysis and reverberation. On one hand, while reverberation models have been used for dereverberation, to the best of our knowledge, no dereverberation model has been used to estimate the RIR in a blind acoustic system identification setting. Moreover, only very weak assumptions about the reverberation process can be derived from neural dereverberators. In the general case, such assumptions highly depend on the DNN architecture, and adaptations to derive individual acoustic characteristics such as the RT_{60} currently require additional DNN parameters. In Chapter 4, we propose two contributions on these statements. We show that DNNs designed to only dereverberate speech are also able to implicitly model reverberation without increasing the number of parameters, and we explicitly synthesize an RIR from a neural dereverberator.

On the other hand, we remark that very few data-driven dereverberation systems are trained using only reverberant data [212, 252], and only exhibit subpar performance compared to classical algorithms such as the WPE method. In Chapter 5, we

build upon the analysis-synthesis technique for acoustic system identification to develop a framework for joint acoustic parameter estimation and dereverberation. Our model is especially suited for low-data regimes, when only reverberant signals are used for training. More precisely, our most data-efficient variant requires only 100 reverberation-parameter-labelled reverberant samples at training to outperform an unsupervised baseline.

Part II

Evaluation and improvement of neural dereverberators using room acoustics

Chapter 4

Dereverberation for acoustic system identification

Chapter abstract

In this Chapter, we investigate the use of neural dereverberators for predicting RIRs, a task linked to acoustical system identification. The goal is to determine if a data-driven dereverberator can learn the physical properties of a room by analyzing the impulse response induced by the dereverberation process. This approach serves as a proxy for evaluating how well DNNs align with room acoustics. Then, a novel physical coherence loss is introduced to train neural dereverberators in a way that preserves acoustic meaning. While optimizing this loss alone can degrade dereverberation performance, a regularization strategy is proposed, enabling data-driven methods to jointly perform both dereverberation and acoustic system identification without compromising performance.

4.1 Introduction

In this chapter, the ability of neural dereverberators to perform RIR prediction is evaluated. This task is closely related to system identification since a dereverberation model can be considered as a black-box system that has to be identified. A key difference is that, in order to perform dereverberation, a DNN supposedly learns some representation of the physical properties in which its reverberant input has been recorded. Hence, these physical properties could be retrieved by computing properties of the RIR induced by the dry signal estimated from a performing dereverberation. This task therefore serves as a proxy to determine to what extent neural dereverberators are compatible with room acoustics.

A summary of the task is presented in Figure 4.1. From the dereverberated output $\hat{s}$ obtained by a dereverberation model from a reverberant signal y, a system

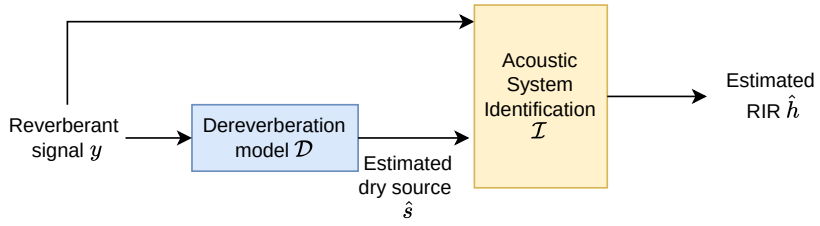


Figure 4.1 – Overview of the system identification by deconvolution task.

identification method $\mathcal{I}$ is used to compute an Room Impulse Response h.

In this chapter, we first review and evaluate deconvolution methods that can be leveraged to solve the acoustic system identification task in this novel setting. Then, a novel physical coherence loss is introduced, in order for a data-driven dereverberator to be trained using acoustic constraints. The results of this chapter are presented incrementally and structured as follows: in section 4.2, existing deconvolution approaches are evaluated on the task of RIR prediction from a dereverberation model. In section 4.3, a regularized deconvolution method is proposed in order to correct some of the defaults of existing dereverberation approaches in a physically meaningful manner. In Section 4.4, the ability of neural dereverberators to be trained to yield a sound source for the task of acoustic system identification is evaluated. Finally, a procedure to train a neural dereverberator to jointly perform both tasks of dereverberation and acoustic system identification is proposed in Section 4.5. This work has been published in

> [50] Louis Bahrman, Mathieu Fontaine, Jonathan Le Roux, and Gaël Richard. « Speech Dereverberation Constrained on Room Impulse Response Characteristics ». In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 622–626

4.2 Evaluation of acoustic system identification methods on a dereverberation model

4.2.1 Introduction

In this Section, we aim to leverage existing convolutive models for room impulse response estimation from a deep-learning model.

Given a reverberant signal y and a dereverberation model $\mathcal{D}$, the problem of room impulse response estimation from a dereverberation model can be formulated as:

$$\hat{h} = \underset{h}{\operatorname{argmin}} \mathcal{L}\left(\mathcal{C}\left(\mathcal{D}\left(y\right), h\right), y\right) \tag{4.1}$$

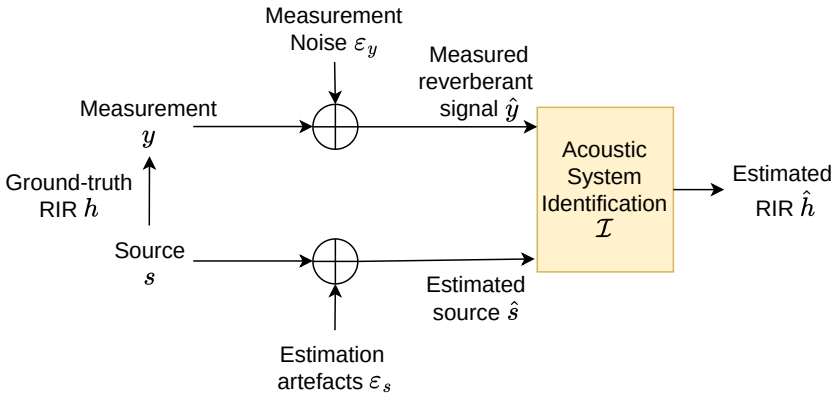


Figure 4.2 – Overview of the convolutive model and sources of noise.

Where $\mathcal{C}$ describes the convolutive model that is inverted, and $\mathcal{L}$ represents a loss function measuring the distance between the estimated and ground-truth reverberant signals. As described in Section 2.3.1, convolutive models include time-domain and STFT domain deconvolution, in which case the distance d can be a distance between the magnitudes or log-magnitude spectrograms of estimated and ground-truth reverberant signals. While some convex losses offer better robustness to outliers in the reverberant signal, such as the Huber [253] loss, we focus here on the MSE estimator, because its optimization yields a closed-form solution for most convolutive models $\mathcal{C}$, and does not require any parameter tuning.

The subtlety of this task is that the estimated dereverberated signal is not exact, and deconvolution methods should be evaluated on their robustness to erroneous measurements y and kernels $\mathcal{D}(y)$. While the robustness of deconvolution methods has been studied from a theoretical point of view, using simple noise models, this is the first time that the robustness of deconvolution methods is evaluated on artifacts induced by neural dereverberators.

4.2.2 Method

In order to represent both deconvolution errors and additive measurement noise, we consider the model illustrated in Figure 4.2. It can be described as follows: a reverberant signal y is the convolution of a dry source s and an RIR h. However, the measured source $\hat{s}$ and the measured signal $\hat{y}$ are respectively perturbed by additive noise ε_s and measurement noise ε_y. In the case presented above in Eq. (4.1), the perturbation on the source ε_s originates from dereverberation artifacts:

$$\varepsilon_s = \mathcal{D}(y) - s. \tag{4.2}$$

Several deconvolution methods are evaluated on their performance:

— Fourier deconvolution described in Eq. (2.36). The regularization parameter λ is only meant to avoid division by zero errors and hence set to 1×10^{-8}.
— Wiener deconvolution described in Eq. (2.37). In this case the regularization parameter is set to the supposedly known SDR.
— Sub-band deconvolution described in Eq.(2.31), with a number of cross-bands F' set to 0 to match the sub-band convolution assumption defined in Eq. (2.11).
— Cross-band deconvolution described in Eq. (2.31). We consider a single cross-band $F' = 1$.

The time-frequency domain system identification method proposed in [62] can be used to compute the cross-band convolution kernel associated with the RIR, but requires an additional processing step in order to go back to the STFT domain.

We prove that the STFT $\boldsymbol{H}$ of the impulse response h can be computed from the convolutive cross-band filter $\mathcal{H}_{f,f',t}$ defined in (2.13):

$$H_{f,t} = \sum_{f'=0}^{F-1} (-1)^{f'} \mathcal{H}_{f,f',t} \tag{4.3}$$

where the multiplication by $(-1)^{f'}$ stems from the centering of the first STFT window. The full proof can be found in Annex A.1.

The deconvolution performance is tested on several metrics, using several noise sources and deconvolution models to evaluate their performance on the task of RIR estimation from a neural dereverberator.

4.2.3 Experimental setting

The experimental setting includes a dataset, a neural dereverberator, and several sources of artifacts.

Noise sources

Several noise sources are used to synthesize ε_s and ε_y:
— Dereverberation artifacts by the pre-trained DNN computed using Eq. (4.2)
— A white noise sampled at various SNRs. Its variance is $\sigma^2 = \frac{\|s\|_2^2}{\mathrm{SNR}}$. Both the noiseless case ($\mathrm{SNR} = +\infty$ dB) and a noisy case ($\mathrm{SNR} = 20$ dB) are considered. Additionally, a measured dry source is synthesized by adding white noise to the ground-truth source s at a SNR corresponding to the measured SNR of the dry source estimated by the DNN: $\mathrm{SNR}(\mathcal{D}(y), s)$. This enables us to compare the degradation on the dry source caused by dereverberation artifacts, with the degradation caused by white noise, and how it influences deconvolution performance.

Dataset

Similarly to [36], we simulated a training dataset by dynamically convolving dry speech signals with simulated RIRs. The dry speech signals are randomly sampled from the close-talking microphone recordings in the WSJ0 dataset [254]. The training set is composed of a total of 61 hours of recordings split into 31,350 audio excerpts. The simulated RIR dataset consists of 32,000 RIRs simulated using the `pyroomacoustics` library [115] with 2000 rooms whose dimensions and RT_{60} are uniformly sampled in the respective ranges of $[5, 10] \times [5, 10] \times [2.5, 4]$ m^3, and $[0.2, 1.0]$ s. In each room, a source is randomly positioned and 16 microphones are sampled such that the source-microphone distance D is uniformly distributed in $[0.75, 2.5]$ m and both source and microphone are at least 50 cm from the walls. In order to align the dry signal target and the direct-path, the samples before the direct path are discarded and the RIR is normalised so that its direct-path impulse is of amplitude 1. This does not change the RIR distribution and compensates for the delay induced by the direct-path, both on the STFT $\boldsymbol{H}$ and on measured acoustic properties.

Neural dereverberator

We consider the bidirectional variant of the FSN model [235], as a baseline dereverberation model. The ability of FSN to process spectrograms both in the full-band and sub-band directions in a recursive manner is required to estimate a cross-band convolutive model. This variant has also been successfully used to solve the physically meaningful task of reverberation-time shortening [255]. As in the original FSN, 49151 sample excerpts (around 3 s at 16 kHz) reverberant audios are processed in the STFT domain using a 512-sample Hann window with an overlap of 50 %. While the dereverberation model is trained using signals of length 3 s, we evaluate deconvolution on 10 s signals. In this case, the measured signal y is of length 11 s The network is trained for 330 000 steps using the Adam optimizer with an initial learning rate of 1×10^{-4} and a One-cycle learning rate scheduler with a maximum at 1×10^{-3}.

4.2.4 Results

The system identification performance is evaluated using RIR characteristics defined in Section 3.1.2. Three metrics are considered, and defined from the DRR, EDR and EDC characteristics of the RIR, as the MAE between these characteristics measured on the ground-truth and estimated RIRs. The deconvolution methods are evaluated on 1000 pairs of dry speech signals and RIRs.

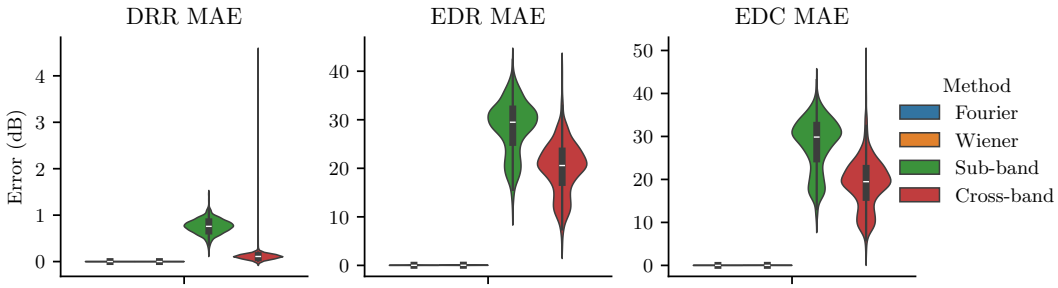


Figure 4.3 – Comparison of deconvolution methods in the noiseless case. Lower bars indicate better performance. Black lines represent 95 % confidence intervals.

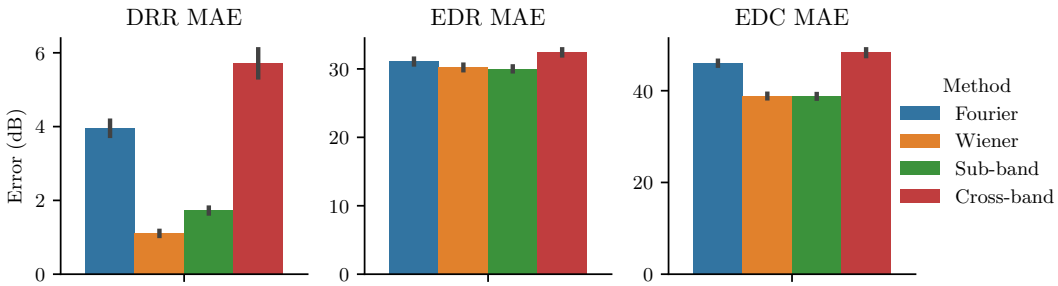


Figure 4.4 – Comparison of the deconvolution methods when a 20 dB SNR is applied to the reverberant signal. Lower bars indicate better performance. Black lines represent 95 % confidence intervals.

Comparison of deconvolution techniques

Noiseless case Figure 4.3 presents the performance of the proposed deconvolution techniques in the noiseless case, i.e., when $\varepsilon_y = \varepsilon_s = 0$. In such cases, the Wiener deconvolution is equivalent to the Fourier deconvolution. We observe that for all metrics, Fourier and Wiener deconvolution yield an almost perfect performance, whereas sub-band and cross-band models exhibit errors. This is due to the inaccuracy of such models, which are approximation of the inter-band and inter-frame convolution presented in section 2.2.1. Unsurprisingly, the more cross-band are considered, the better the deconvolution performance.

Noise on the measured reverberant signal Figure 4.4 presents the results of the deconvolution methods with an additive white noise at 20 dB SNR. As expected, all methods exhibit a substantial degradation in performance compared to the noiseless case. However, their robustness to additive noise varies notably. Indeed, the cross-band deconvolution method, which previously achieved the better results than the sub-band method in the absence of noise, now performs the worst. Similarly, the Fourier and Wiener deconvolution methods are no longer equivalent, since the Wiener formulation explicitly accounts for the noise power spectrum, whereas the

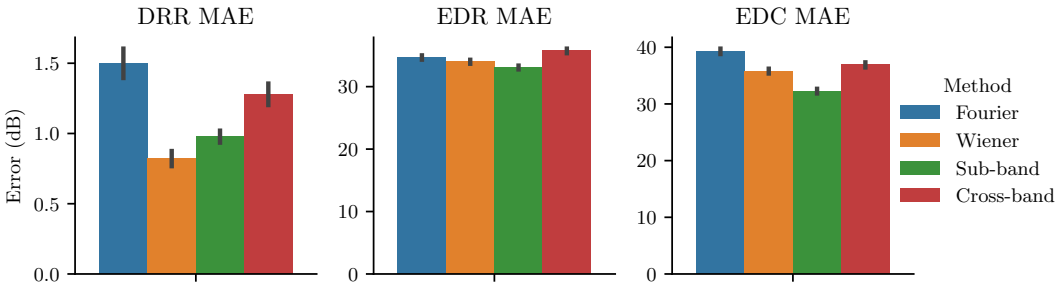


Figure 4.5 – Comparison of the deconvolution methods when a 20 dB SNR is applied to the dry signal. Lower bars indicate better performance. Black lines represent 95 % confidence intervals.

Fourier method does not. A Wilcoxon paired signed-rank test ($p < 1 \times 10^{-4}$) confirms significant performance differences on each metric independently. The Fourier deconvolution method performs significantly better than the cross-band approach on all metrics. The Wiener and sub-band deconvolution method further outperform both Fourier and cross-band techniques. Specifically, the Wiener method yields RIRs which are more accurate in terms of DRR, while the sub-band method performs significantly better on both EDC and EDR metrics. Overall, the sub-band deconvolution remains the preferred approach in noisy conditions, as it achieves competitive performance without requiring an explicit estimate of the noise variance.

Noise on the estimated source Figure 4.5 presents the results when a 20 dB SNR noise is applied on the dry source signal. Compared to the situation when the same noise distribution is applied to the dry source signal, all deconvolution methods are performing better on all metrics except for the EDR MAE metric. This demonstrates that the task of acoustic signal estimation by deconvolution might be easier to solve in the case where the noise is on the kernel rather than on the measurement.

As with the noisy measurement, both the Wiener and sub-band deconvolution methods are performing better than the Fourier and cross-band methods on all metrics, with the sub-band method significantly ($p < 1 \times 10^{-4}$) outperforming the Wiener deconvolution technique on all metrics except the DRR MAE. Surprisingly, Wiener deconvolution still performs better than Fourier deconvolution, albeit not being suited for this noise setting. This could be explained by the fact that adding white noise to the dry signal virtually increases the condition number of the deconvolution described in Section (2.30).

In the remaining of this section, we will consider the sub-band deconvolution variant as a baseline, because it is the least sensitive to noise and does not require a SNR estimation which would not be available at inference.

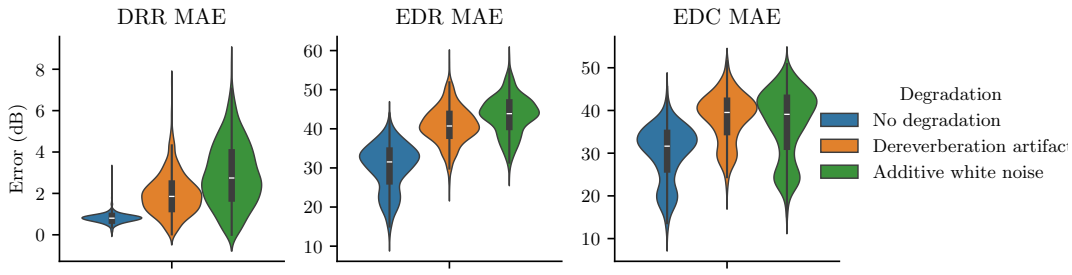


Figure 4.6 – Influence on the nature of noise on the estimated source. Lower bars indicate better performance.

Comparison between the noise models

This experiment evaluates if deconvolution artifacts can be modeled as white noise for the task of RIR estimation by deconvolution. We compare the performance of the proposed dereverberation estimator when the noise on the source ε_s is synthesized using either dereverberation artifacts as in Eq. (4.2), or a white noise with the same SNR. We consider the sub-band deconvolution method, as it offers the best performance estimation and did not rely on an estimation of the SNR that would not be available at inference.

Figure 4.6 shows the performance of the sub-band deconvolution method for RIR estimation with various distributions of noise. We observe that the distribution of the results greatly varies with respect to the degradation method. A statistical Wilcoxon paired test confirms that the differences are significant with p-value $<1\times10^{-4}$. Hence, we can state that for the problem of RIR estimation by deconvolution, it is false to assume that DNNs yield artifacts which can be modeled as stationary white noise for deconvolution. This could be explained by the behaviour of neural dereverberators not being consistent with room acoustics in several ways:

— They approximate a linear mapping (the inversion of an RIR which is a linear operation), by a nonlinear one.

— They create artifacts which are time-sensitive, unlike the reverberation process which is assumed to be time-invariant.

A more thorough evaluation of several DNN architectures for the task of physical properties estimation will be provided in Section 4.4.

An example of deconvolution artifacts

An example of an estimated RIR obtained by deconvolving the reverberant input of a neural dereverberator by its dereverberated output, is provided in Figure 4.7. Two methods are shown alongside the exact RIR: deconvolution in the Fourier domain and in the STFT domain under the sub-band approximation. We observe that, unlike physically consistent RIRs which have an exponentially decreasing tail (as seen

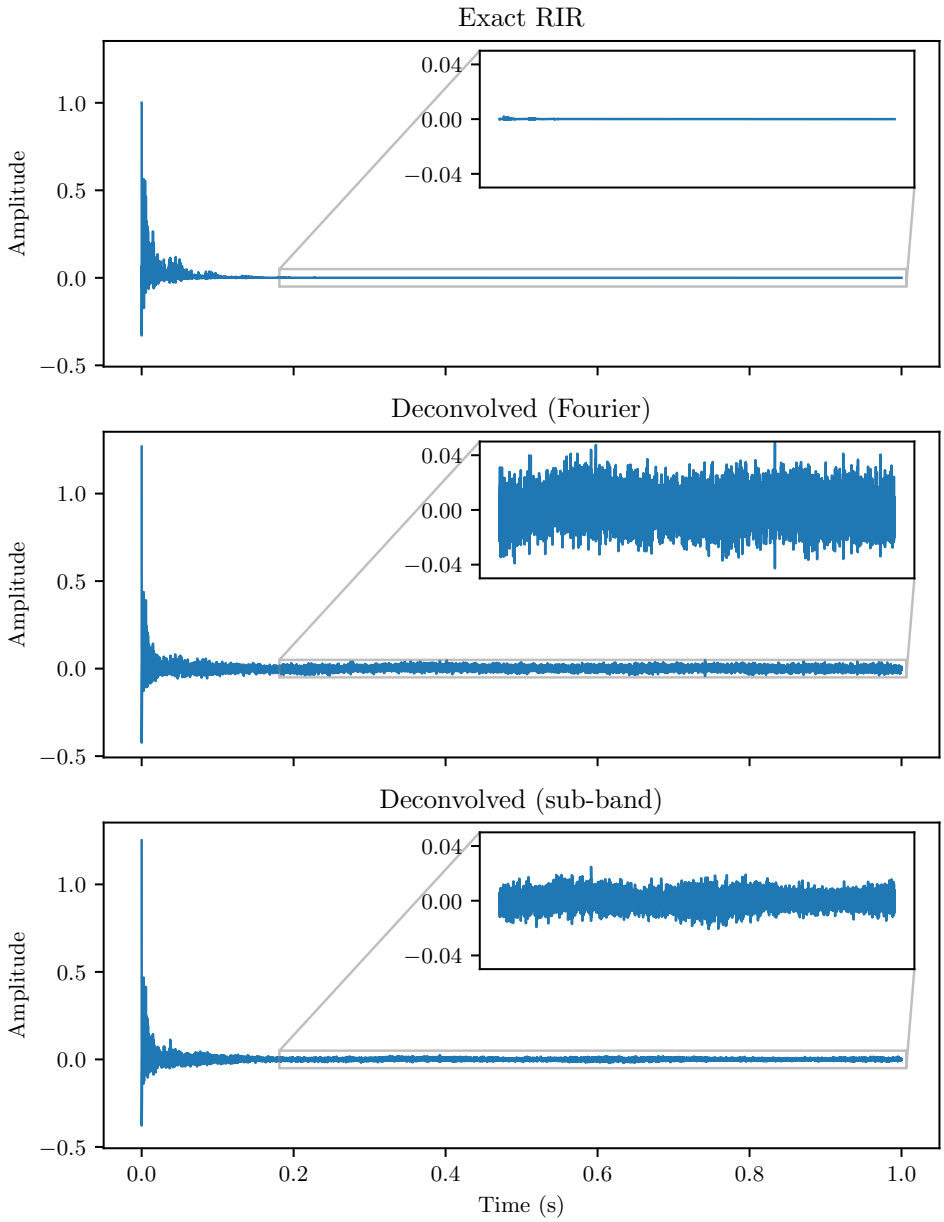


Figure 4.7 – Example of the RIR obtained by deconvolving a reverberant speech by a dereverberated speech (Equivalent SNR=4.00 dB).

in Section 3.1.2), our estimated RIRs exhibit a very noisy tail. The noise seems to be more present when the RIR has been estimated in the Fourier domain than by deconvolution in each STFT sub-band. In the next section, we introduce a deconvolution regularization in order to mitigate this effect.

4.2.5 Conclusion

In this section, the task of acoustic system identification using the output signal of a dereverberation model has been introduced. Existing deconvolution methods have been compared on this novel task and several observations have been made: first, the performance of deconvolution methods seems to be less affected by presence of estimation errors on the dry signal than additive noise on the measured signal. However, as these estimation errors consist of artifacts caused by a neural dereverberator, they cannot be simply modeled as white noise. Furthermore, while the sub-band approximation seem to yield the best performance among existing deconvolution methods, it is still unable to produce a realistic RIR which tail is supposed to be exponentially decreasing. This calls for domain-specific knowledge to enhance the deconvolution process. In the next section, we will introduce a novel regularization method to improve the physical consistency of deconvolved RIRs.

4.3 RIR characteristics regularization for acoustic system identification

In this section, a novel regularization function is introduced, that encourages a deconvolution method to produce physically-consistent RIRs even in the presence of noise. A proximal gradient descent algorithm is derived and evaluated on the task of RIR estimation from a neural dereverberator.

4.3.1 Method

As presented in Section 3.1.1, a key characteristic of RIRs is their decreasing amplitude. However, previous experiments related in Section 4.2 showed that such a property is lost during deconvolution by a dereverberated source. In this Chapter, we present a regularization function that encourages the RIR to have an amplitude below a predefined threshold.

More precisely, given a reverberant signal y and a (possibly erroneous) dry estimate s of the source, we propose to formulate the non-blind RIR estimation problem as:

$$\underset{h}{\operatorname{argmin}} \|h \star s - y\|_2^2 + \lambda g(h), \tag{4.4}$$

where $g(h)$ is our proposed regularization function, and $\lambda \in \mathbb{R}_+$ is the regularization

constant. Given Polack's reverberation parameters σ and τ defined in Eq. (3.4), and n_d the time index of the start of late reverberation, the proposed energy decay regularization function g can be defined as a function of the RIR h:

$$g(h) \triangleq \sum_{n>n_d} \max\left(0, h[n]^2 - \sigma^2 e^{-2n/\tau}\right). \tag{4.5}$$

It encourages the squared amplitudes of the RIR to be below an exponentially decreasing threshold parametrized by its decay rate and initial amplitude. Compared to other regularization techniques found in the literature [165], our proposed regularization does not enforce sparsity and allows for some peaks in the RIR to be present, provided their amplitude is below the exponentially decreasing threshold.

The decay term g, is convex in h, as the sum and composition of convex functions. As g is not differentiable everywhere, a proximal gradient descent algorithm [64] can be used in order to solve Eq. (4.4). The proximal operator of λg is defined from Eq (2.33) as:

$$\text{prox}_{\lambda g}(x) \triangleq \underset{h}{\text{argmin}}\, \lambda g(h) + \frac{1}{2}\|h - x\|_2^2 \tag{4.6}$$

This proximal operator is separable, so it can be computed element-wise as:

$$\text{prox}_{\lambda g}(x)[n] = \underset{h[n]}{\text{argmin}}\, \frac{1}{2}\left(h[n] - x\right)^2 + \lambda \max\left(0, h[n]^2 - \sigma^2 e^{-2n/\tau}\right) \tag{4.7}$$

$$= \begin{cases} x & \text{if } |x| \leq \sigma^2 e^{-2n/\tau} \\ \frac{x}{2\lambda+1} & \text{if } \left|\frac{x}{2\lambda+1}\right| \geq \sigma^2 e^{-2n/\tau} \\ \text{sign}(x)\sigma e^{-n/\tau} & \text{otherwise} \end{cases} \tag{4.8}$$

Details of the computation of the prox are given in Annex A.2.

On the other hand, the gradient of the deconvolution loss $\mathcal{L}_d \triangleq \|h \star s - y\|$ is:

$$\nabla_{\mathcal{L}_d}(h) = 2s \circledast (y - s \star h), \tag{4.9}$$

where $\circledast$ denotes the cross-correlation operator defined in Section 2.2.2.

A line search method is applied in order to compute the optimal gradient step η as:

$$\eta = \underset{\eta \geq 0}{\text{argmin}}\, \mathcal{L}_d\left(h - \eta \nabla_{\mathcal{L}_d}(h)\right) \tag{4.10}$$

$$= \frac{a^\intercal b}{\|a\|_2^2}, \tag{4.11}$$

where $b = s \star h - y$ and $a = 2s \star (s \circledast b)$. Note though, that this line-search method is

not guaranteed to yield the optimal gradient step for the proximal gradient descent. A method to compute the optimal line-search for proximal gradient descent in an efficient manner was recently proposed [256], and could represent an interesting alternative for our problem.

Our proposed method requires the reverberation parameters σ and τ to be estimated, which can be done either from the RIR using the method described in Section 3.2.1 or in a blind setting using methods described in Section 3.2.3. The problem of jointly estimating reverberation parameters and the RIR h is however ill-posed. The complete algorithm to estimate a physically plausible RIR from a reverberant and dry signal is detailed in Algorithm 1.

Algorithm 1 Proximal gradient descent for RIR estimation

Require:
s dry (or dereverberated) signal
y reverberant signal
I number of iterations
λ regularization parameter
σ, τ exponential decay parameters
Ensure: h estimated RIR
initialize $h^{(0)}$
for $0 \leq i < I$ **do**
$b \leftarrow s \star h - y$
$a \leftarrow 2s \star (s \circledast b)$
$\eta \leftarrow \frac{a^\top b}{\|a\|_2^2}$ ▷ Line search
$h^{(i+1)} \leftarrow h^{(i)} - 2\eta s \circledast (y - s \star h)$ ▷ Gradient step
for $0 \leq n < N_h$ **do**
if $\left|h^{(i+1)}[n]\right| > \sigma \exp^{-n/\tau}$ **then** ▷ Proximal operator
if $\left|\frac{h^{(i+1)}[n]}{1+\lambda}\right| > \sigma \exp^{-n/\tau}$ **then**
$h^{(i+1)}[n] \leftarrow \frac{h^{(i+1)}[n]}{1+\lambda}$
else
$h^{(i+1)}[n] \leftarrow \text{sign}\left(h^{(i+1)}[n]\right) \sigma e^{-n/\tau}$
end if
end if
end for
end for

4.3.2 Experimental setting

We test our method on the same setting as described in Section 4.2.3. Two variants of the proposed method are tested:

— Proximal-gradient: solves the problem described in Eq. (4.4) using the proposed Algorithm 1.

— Adam: solves the same problem using the Adam optimizer [257]. Although

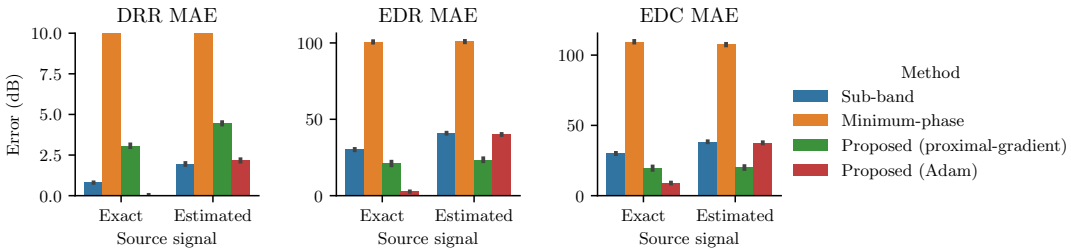


Figure 4.8 – Results of our proposed RIR estimator. Black lines represent 95 % confidence intervals.

the loss function of the term g is not supposed differentiable everywhere, we use the implementation of the max function provided by the `Pytorch` [258] library, and allows to backpropagate our proposed loss on its whole domain of definition.

Based on preliminary experiments, the parameter λ is set to 1×10^{-2}, and $I = 1 \times 10^4$ iterations are performed. The maximum learning rate of Adam is set to 1×10^{-4}. The exponential decay parameters are estimated directly from the ground-truth RIR. However, the parameter σ used for thresholding is set to ensure that the reverberant energy could be twice as high as the ground-truth energy predicted by Polack's model. More precisely, we assume that the ground-truth RIR has a reverberant energy of $\sigma^2 \frac{\tau}{2} e^{-2n_d/\tau}$, with σ defined in Eq. (3.4) and τ defined in Eq. (3.5). We simply multiply σ by $\sqrt{2}$ at input of our algorithm in order to multiply the energy threshold by 2. In addition to the sub-band deconvolution baselines presented above, we also compare our proposed approach to a method encouraging a short RIR decay, in the form of minimum-phase filter estimation. Indeed, minimum-phase filtering is guaranteed to yield the RIR having the sharpest energy decay, causing the noise in the tail of an estimated RIR to move towards the start of the impulse response. Such an approach has for instance been used in [219]. To compute the minimum-phase equivalent filter of the acoustic system, we first estimate the magnitudes of the RIR by deconvolution using the Fourier method. Then, we derive the minimum-phase filter using the procedure described in Section 2.2.3.

4.3.3 Results

Figure 4.8 evaluates the performance of our proposed method against sub-band and minimum-phase deconvolution techniques. Two settings are evaluated: when the ground-truth source is provided (denoted *Exact*) and when it is estimated by a neural dereverberator (denoted *Estimated*). A first observation is the inefficiency of the minimum-phase deconvolution technique. The RIR decay obtained using minimum-phase deconvolution is too sharp to accurately model late reverberation, and yields huge errors on all RIR metrics. Moreover, one of our proposed methods (Proximal-

gradient or Adam optimizer) outperforms sub-band deconvolution on all metrics with the exception of the DRR MAE metric when the source has been estimated from a reverberant signal. Comparing the proposed proximal gradient descent algorithm with direct optimization via Adam, we remark that they converge to distinct optima. When the ground-truth source is provided, our proposed proximal-gradient-based method is less accurate than Adam optimization on all metrics. This tendency however reverts when dry source signals are estimated using a neural dereverberator. In such cases, the proposed proximal-gradient-based method yields an RIR that more accurately represents the EDR and EDC of the ground-truth RIR.

Observing one RIR example, as shown in Figure 4.9 provides an explanation to such differences. Indeed, the proposed proximal-gradient optimizer imposes a stronger constraint on the tail of the RIR, even though the regularization parameter λ is set to the same value as for the Adam optimizer. On one hand, this constraint forces the tail of the RIR to be exponentially decreasing which is consistent with physics. On the other hand, some peaks in the early echoes are absent from the RIR obtained using the proposed proximal gradient descent, while they are correctly estimated by the Adam optimizer. A solution to alleviate such issues would be to apply the proximal gradient not directly after the direct path, but to leave some time to allow some sparse early echoes above the threshold.

4.3.4 Conclusion

A physically meaningful regularization of the deconvolution for non-blind acoustic system identification has been proposed. Experiments show that it improves existing deconvolution techniques, by forcing an exponential decay of the estimated RIR. Yet, a question seems to remain unanswered: are such artifacts solely caused by the dereverberation method employed, or could the dereverberation model be trained specifically for the purpose of RIR estimation via deconvolution ? In other words, could a dereverberation model be trained to optimize such a criterion?

4.4 Dereverberation model training using a physics-driven loss

We now consider a model trained solely with a loss function inspired from room acoustics.

We propose to introduce a novel loss term that imposes physical constraints on the RIR characteristics measured from the acoustic system predicted by deconvolving a deep neural network's input by its output.

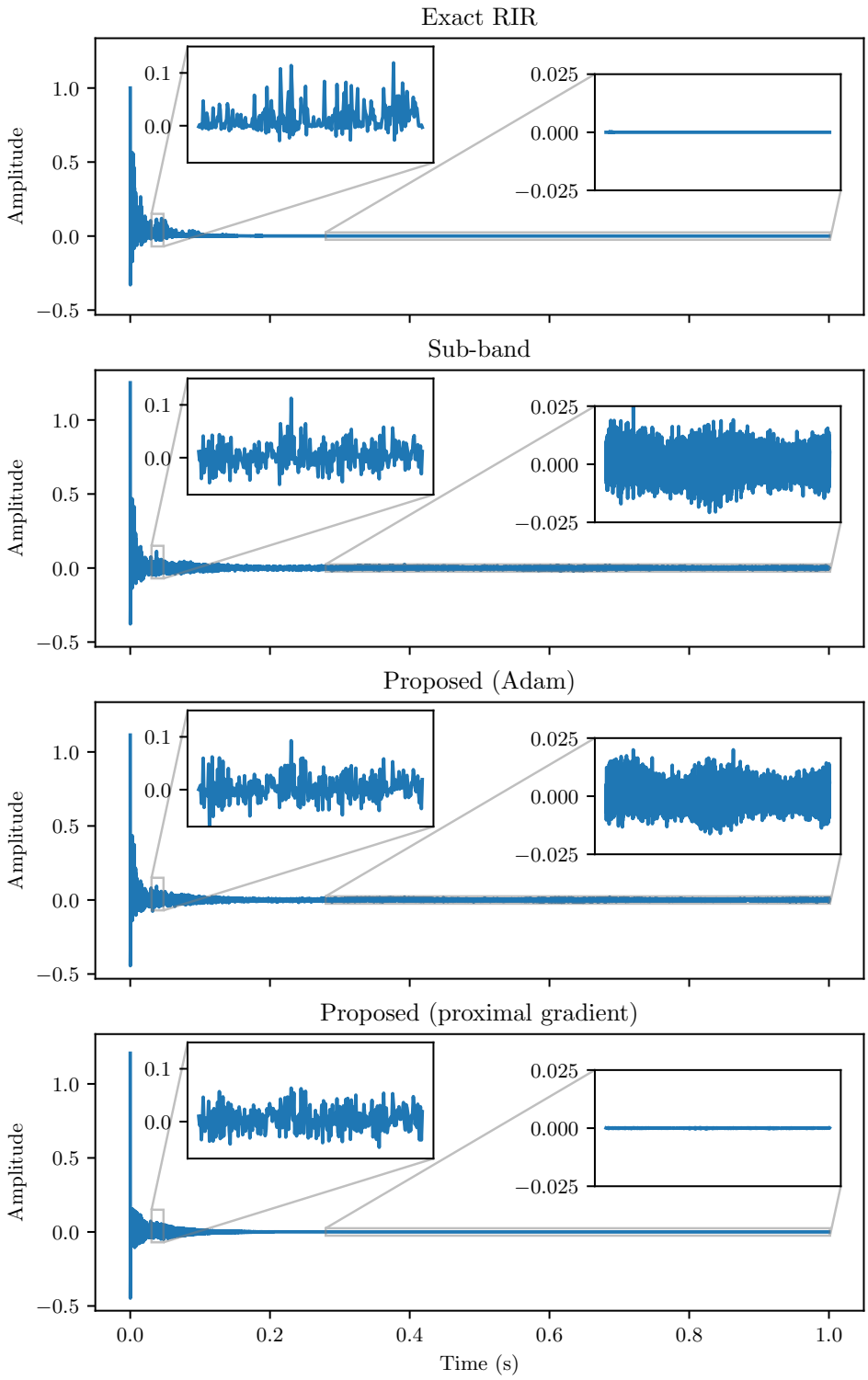


Figure 4.9 – Example of an RIR obtained from a neural dereverberator using our proposed regularization.

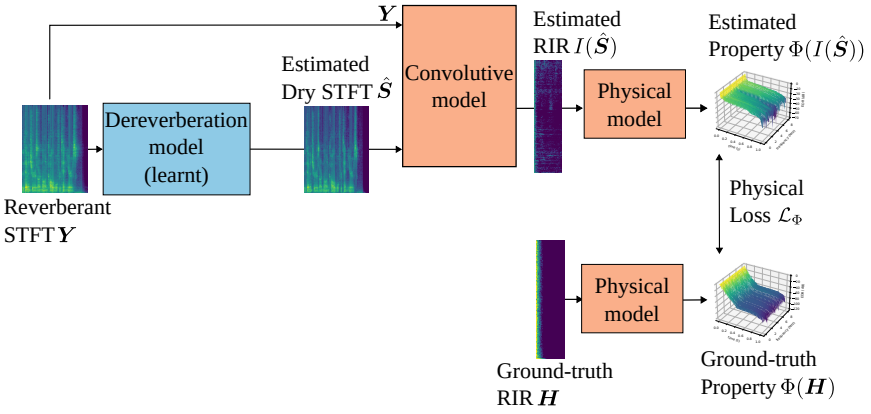


Figure 4.10 – Dereverberation model training using a physics-driven loss.

4.4.1 Method

The compatibility between several deep-learning models and room-acoustics-based loss functions is studied by training a DNN solely on a room-acoustics-inspired loss function, and measuring its performance on the dereverberation task. The room-acoustics inspired loss function is computed on the RIR induced by the dereverberation model output. The general procedure to define our physical loss term is as follows. From the dereverberated output $\hat{\boldsymbol{S}}$ obtained by the DNN $\mathcal{D}_w$ from a reverberant signal $\boldsymbol{Y}$, an acoustic system identification module $\mathcal{I}$ computes the estimated RIR $\hat{h}$ mapping the output of the DNN to its reverberant input. A physical model is then used to compute an estimated physical property $\Phi(\hat{h})$ from the estimated RIR, and similarly a target physical property $\Phi(h)$ from the oracle RIR obtained with the ground-truth anechoic signal. Their distance is finally used to define our physical loss function $\mathcal{L}_\phi$. The training procedure is detailed in the diagram 4.10.

To study the influence of both the DNN architecture and RIR estimation loss, we consider an overfitting setup where the DNN is trained on a single batch. No pre-training is performed. Hence, misalignment between the reverberation-inspired training objective can only be attributed to the incompatibility of the neural dereverberator, the deconvolution method and the RIR characteristics studied.

4.4.2 Experimental setting

The training dataset consists of a single batch of size 2 randomly extracted from the dataset described in Section 4.2.3.

Considered variants

Several combinations of deep dereverberation models, deconvolution methods and RIR characteristics are considered.

Neural dereverberators include:

— FCN-PI: a Fully Convolutional Network (FCN) introduced in [259]. *PI* stands for *phase-invariant* and denotes the fact that this neural network processes magnitude spectrograms, and completely discards STFT phase information. The clean phase is approximated as being equal to the reverberant phase.
— FCN-RI2RI: a variant of FCN that enables phase processing. This variant, described in [243], concatenates the real and imaginary parts of the spectrogram as distinct channels of the input and output.
— FSN-R2RI: FSN presented in [235], that estimates the complex ideal mask mapping the reverberant STFT to the dry STFT.
— FSN-PI: a phase-invariant variant of FSN, where the synthesized mask is real-valued, forcing the phase of the dereverberated spectrogram to be the same as the phase of the reverberant.
— ConvTasNet: a temporal convolutional neural network, proposed in [231].

Let the ground-truth RIR be denoted by h in the time domain and $\boldsymbol{H}$ in the time-frequency domain, with corresponding estimates $\hat{h}$ and $\hat{\boldsymbol{H}}$, respectively. Based on these representations, the different physical loss functions considered include:

— Time-domain MSE loss: $\left\| \hat{h} - h \right\|_2^2$
— Spectral-magnitude MSE: $\left\| \left| \hat{\boldsymbol{H}} \right| - |\boldsymbol{H}| \right\|_F^2$
— Log-spectral-magnitude MSE: $\left\| \log \left| \hat{\boldsymbol{H}} \right| - \log |\boldsymbol{H}| \right\|_F^2$
— EDR MSE: $\left\| \mathrm{EDR}_{\mathrm{dB}}(\hat{\boldsymbol{H}}) - \mathrm{EDR}_{\mathrm{dB}}(\boldsymbol{H}) \right\|_F^2$, where the EDR is expressed in dB.
— $\mathrm{EDR}_{>-20}$ MSE: $\sum_{f,t} \left| \mathrm{EDR}_{\mathrm{dB}}(\hat{\boldsymbol{H}})_{f,t} - \mathrm{EDR}_{\mathrm{dB}}(\boldsymbol{H})_{f,t} \right|^2 \mathbb{1}_{\{\mathrm{EDR}_{\mathrm{dB}}(\boldsymbol{H})_{f,t} > -20\}}$, where $\mathbb{1}_{\mathcal{X}}$ denotes the characteristic function of the set $\mathcal{X}$. This loss restricts the evaluation of the dB-scaled EDR metric to the indices where the normalized EDR has an energy greater than $-20\,\mathrm{dB}$.
— EDC MSE: $\left\| \mathrm{EDC}_{\mathrm{dB}}(\hat{h}) - \mathrm{EDC}_{\mathrm{dB}}(h) \right\|_2^2$
— $\mathrm{EDC}_{>-20}$ MSE: $\sum_n \left| \mathrm{EDC}_{\mathrm{dB}}(\hat{h})(n) - \mathrm{EDC}_{\mathrm{dB}}(h)(n) \right|^2 \mathbb{1}_{\{\mathrm{EDC}_{\mathrm{dB}}(h)(n) > -20\}}$
— RT_{60} MSE: $\left| \mathrm{RT}_{60}(\hat{h}) - \mathrm{RT}_{60}(h) \right|^2$
— Envelope MSE: $\left\| \left| g \star \hat{h} \right| - |g \star h| \right\|_2^2$, where g corresponds to the impulse response of a Hilbert filter of length 60 samples, described in Section 2.2.3.

The initial decay variants for the EDC and EDR are motivated by the observation that deep-learning RIR analysis methods focussed on the first 120 ms of reverberation [260].

Two deconvolution methods are considered among those presented in Section 4.2.2: the Fourier and sub-band deconvolution methods.

4.4.3 Results

A first experiment examines whether the dereverberation model can be trained using the physical loss. All DNN variants are trained to optimize each loss for 10

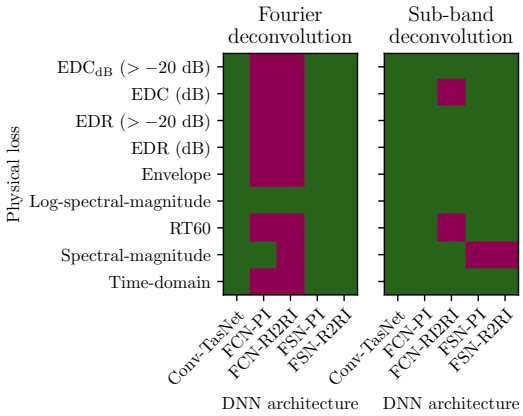


Figure 4.11 – Convergence when overfitting to the physical loss. Green cells indicate the DNN converges on the proposed loss and red cells indicate absence of convergence.

epochs. Figure 4.11 presents the results of training a dereverberation model solely on a physical loss, for all neural dereverberators and physical losses. Each cell is colored according to the convergence of the tuple $(\mathcal{D}, \mathcal{L}_\Phi)$. More precisely, we consider that the deep neural network converges if its training loss decreases over the course of its training on a single batch. Several reasons can explain why a DNN does not converge when trained in this setting. Convergence failure can mean that the loss landscape is highly non-smooth or non-convex, meaning that the architecture or number of parameters of the network is not suited to the optimization objective. We observe that all variants of the STFT-domain convolutional network fail to converge on most physical losses computed using Fourier deconvolution. This means that this architecture is unsuited for the task, which might be due to the too small latent space of the network, or its low performance when processing phase without any other adaptation, detailed in [243]. On the other hand, sub-band deconvolution seems to be more robust, as most of the physical losses converge for this variant.

A second experiment aims to determine whether training with the physical objective can also enable the DNN learn dereverberation. As such, for all models excluding FCN, the improvement of SI-SDR is measured between the first and the 100th training epoch. Results are presented in Figure 4.12. We observe that most models struggle to produce a more meaningful estimated dry signal when trained to solely optimize a physical loss. Fourier deconvolution seems to yield subpar performance, due to its poor robustness to dry signal estimation artifacts. Only a few losses seem to be able to make the FSN model converge on both the RIR estimation and dereverberation tasks.

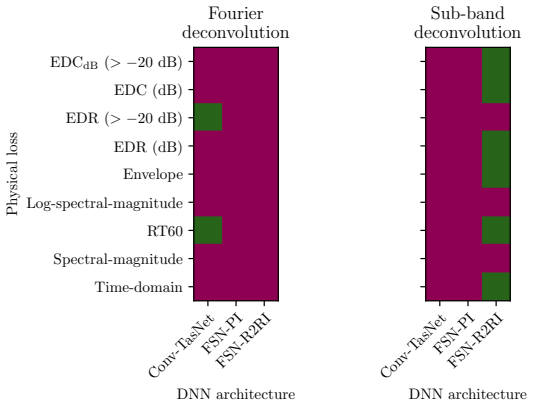


Figure 4.12 – Influence of overfitting a physical loss onto dereverberation performance. Cells are colored in green if and only if the SI-SDR metric improves.

4.4.4 Conclusion

In this section, we proposed to train a neural dereverberator solely for the task of RIR estimation. Results indicate that not all DNN architectures are suited for this task, and confirm the poor performance of the Fourier deconvolution. Optimizing the neural dereverberator for RIR estimation seems to produce a limited improvement on its performance for the speech dereverberation task.

In the next section, a procedure to jointly train a dereverberation model for the task of RIR estimation and dereverberation is proposed.

4.5 RIR-based training loss regularization for dereverberation

In this section, we evaluate whether DNNs can be optimized such that for the room impulse response induced by the dereverberation process to remain consistent with room acoustics.

4.5.1 Method

Overview

We propose to introduce a new loss term that imposes physical constraints on the RIR characteristics measured when training a neural dereverberator. The general procedure to define our physical loss regularization is as follows. From the dereverberated output $\hat{\boldsymbol{S}}$ obtained by the neural dereverberator from a reverberant signal $\boldsymbol{Y}$, an acoustic system identification module computes an estimate $I(\hat{\boldsymbol{S}})$ of the STFT of the corresponding RIR. A physical model is then used to compute an estimated physical property $\Phi(I(\hat{\boldsymbol{S}}))$ from the estimated acoustic system, and similarly a target physical

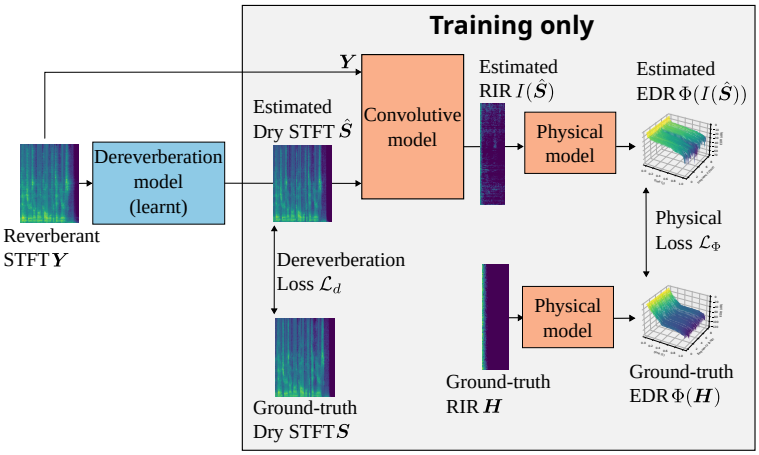


Figure 4.13 – Overview of the physics-driven regularization for dereverberation method.

property $\Phi(I(\boldsymbol{S}))$ from the oracle acoustic system obtained with the ground-truth anechoic signal. Their distance is finally used to define our physical loss function $\mathcal{L}_\phi$. This new loss term $\mathcal{L}_\phi$ can be summed with a classical dereverberation loss $\mathcal{L}_d$ (e.g., assessing the reconstruction quality of the dry or direct-path signal) to form the DNN training loss $\mathcal{L} = \mathcal{L}_\phi + \lambda\mathcal{L}_d$. A diagram of the training procedure is shown in Fig. 4.13.

Because the acoustic system identification and physical models are not parametric, they do not need to be trained. At inference, for the dereverberation task, these blocks are discarded, and only the neural dereverberator is used. Hence, the number of parameters, as well as the computational complexity and memory footprint are the same as for the original DNN.

Corrected Convolutive Model

The number of cross-bands is limited by the dimension of the least-squares system to solve at Eq. (2.31). For the system to have a unique solution, it is required that $\bar{\boldsymbol{S}}_f$ is full-rank, hence the relation $(2F'+1)T_h < T_y$ must hold. Taking into account the length of the dry signals and RIRs in our training data, as well as the computational load, we limit ourselves to considering the sub-band ($F' = 0$) and 3-band ($F' = 1$) approximations for the cross-band convolution. We solve Eq. (2.31) using QR decomposition, and compute $\mathcal{I}(\boldsymbol{S})_{f,t}$ and $\mathcal{I}(\hat{\boldsymbol{S}})_{f,t}$ using Eq. (4.3). Because our model only considers a few cross-bands, this estimate will not yield the exact STFT $\boldsymbol{H}_{f,t}$ of the RIR, but an approximation, even if it is computed from the clean speech $\boldsymbol{S}$. We define the modeling error at each time-frequency bin indexed by f,t as $\boldsymbol{\mathcal{E}}_{f,t} = I(\boldsymbol{S})_{f,t} - \boldsymbol{H}_{f,t}$. To make physical properties less dependent on this approximation, we attempt to compensate for the error via a spectral-subtraction-based correction. The spectral subtraction yields $I(\boldsymbol{S})^c_{f,t}$, an estimator of the RIR STFT. The same error

correction can be applied to the estimate $I(\hat{\boldsymbol{S}})$ of the RIR obtained from the estimate $\hat{\boldsymbol{S}}$ of the dry speech:

$$I(\boldsymbol{S})^c_{f,t} = \left(|I(\boldsymbol{S})_{f,t}|^2 - |\boldsymbol{\mathcal{E}}_{f,t}|^2\right)^{1/2} e^{j\angle I(\boldsymbol{S})_{f,t}}, \tag{4.12}$$

$$I(\hat{\boldsymbol{S}})^c_{f,t} = \left(|I(\hat{\boldsymbol{S}})_{f,t}|^2 - |\boldsymbol{\mathcal{E}}_{f,t}|^2\right)^{1/2} e^{j\angle I(\hat{\boldsymbol{S}})_{f,t}}. \tag{4.13}$$

Note that adjusting both target and estimated convolutive transfer functions by the same quantity will alter the nonlinear behaviour of the physical model employed.

Physical coherence loss

As an example of a physical characteristic of interest to be used as a constraint on the RIR, we consider the dB-scaled EDR [101]. Given an STFT of an RIR or an approximation of it, $\boldsymbol{R}$, the dB-scaled EDR is obtained as:

$$\Phi_{f,t}(\boldsymbol{R}) \triangleq \mathrm{EDR}^s(\boldsymbol{R})_{f,t} = 10\log_{10}\frac{\mathrm{EDR}(\boldsymbol{R})_{f,t}}{\mathrm{EDR}(\boldsymbol{R})_{f,0}}. \tag{4.14}$$

The physical coherence loss $\mathcal{L}_\Phi$ can then be defined as a point-wise mean-squared error between the dB-scaled EDRs obtained from an estimate $\hat{\boldsymbol{R}}$ and a target $\boldsymbol{R}$. Since the tail of the EDR is very sensitive to sub-band approximation errors and has high values on the log scale, both target and estimated EDRs are masked to exclude time-frequency bins where the target EDR is lower than -20 dB:

$$\mathcal{L}_\Phi(\hat{\boldsymbol{R}}, \boldsymbol{R}) = \sum_{f,t} \left|\Phi_{f,t}(\hat{\boldsymbol{R}}) - \Phi_{f,t}(\boldsymbol{R})\right|^2 \mathbb{1}_{\{\Phi_{f,t}(\boldsymbol{R})>-20\}}. \tag{4.15}$$

We consider several variants for the selection of $\hat{\boldsymbol{R}}$ and $\boldsymbol{R}$, such as $I(\hat{\boldsymbol{S}})^c$ and $I(\boldsymbol{S})^c$, as described in Section 4.5.2.

Multi-objective training

To balance both physical coherence and reconstruction losses in a multi-task training setting, we adopt the GradNorm method [261]. GradNorm automatically adjusts the regularization parameter λ such that the Frobenius norm of the gradient of both losses with respect to the model weights w are equal:

$$\text{find } \lambda \text{ s.t. } \left\|\frac{\partial \mathcal{L}_d}{\partial w}\right\| = \left\|\lambda\frac{\partial \mathcal{L}_\Phi}{\partial w}\right\|. \tag{4.16}$$

In our setting, $\mathcal{L}_\Phi$ is highly non-convex with respect to the network parameters, so we prioritize the reconstruction loss $\mathcal{L}_d$ over the physical coherence loss $\mathcal{L}_\Phi$ to stabilize training. After the optimal weight λ has been found using GradNorm, we further multiply the physical coherence loss by a constant weight α. Based on preliminary

experiments, we set $\alpha = 0.1$. The DNN training loss is then:

$$\mathcal{L} = \mathcal{L}_d + \alpha\lambda\mathcal{L}_\Phi \tag{4.17}$$

4.5.2 Experimental setup

Model variants

We assess several variants of our method with FSN [235] as the baseline neural dereverberator (see Fig. 4.13). We select its bidirectional version, and keep the original training loss expressed as the MSE between the estimated and ground-truth complex ratio masks [241] as the dereverberation loss $\mathcal{L}_d$.

The following variants are considered, representing different ways to compute the convolutive model. We define two kinds of approaches depending on whether sub-band or cross-band filters are considered to obtain the estimates of the RIR STFT, and which estimates and targets are compared:

— Sub-band approach (FSN+SB): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{S}}), \boldsymbol{H})$, comparing the estimate from $\hat{\boldsymbol{S}}$ with ground-truth RIR STFT $\boldsymbol{H}$.

— Symmetric Sub-band approach (FSN+SSB): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{S}}), I(\boldsymbol{S}))$, comparing the estimate from $\hat{\boldsymbol{S}}$ with the estimate from $\boldsymbol{S}$.

— Corrected Suband approach (FSN+CSB): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{S}})^c, I(\boldsymbol{S})^c)$, comparing the corrected estimate from $\hat{\boldsymbol{S}}$ with the corrected estimate from $\boldsymbol{S}$.

— 3-band approach (FSN+3B): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{S}}), \boldsymbol{H})$, similar to SB but computed using $F' = 1$ crossbands.

Miscellaneous configurations

We reuse the dataset described in Section 4.2.3. At training time, we use a dynamic mixing procedure consisting in randomly selecting a dry signal and RIR pair. As in the original FSN, 49151 sample excerpts (around 3 s at 16 kHz) reverberant audios are processed in the STFT domain using a 512-sample Hann window with an overlap of 50 %. The network is trained for 330 000 steps using the Adam optimizer with an initial learning rate of 10^{-4} and a One-cycle learning rate scheduler with a maximum at 10^{-3}.

We evaluate the proposed method on two different tasks: speech dereverberation and room impulse response characterization.

Evaluation tasks and metrics

We evaluate the generalization performance of our metrics to both unseen sources and rooms. For dry sources, we consider the test set of WSJ0 [254], and Librispeech clean [262]. Two reverberation datasets are considered: one simulated using unseen rooms matching the same physical parameters as the training dataset described in

Table 4.1 – Dereverberation scores ± standard deviation (std.) for FullSubNet (FSN) and its constraints versions.

	Matched RIRs						Mismatched RIRs					
	WSJ0			LibriSpeech clean			WSJ0			LibriSpeech clean		
	STOI	SISDR	WB-PESQ	STOI	SISDR	WB-PESQ	STOI	SISDR	WB-PESQ	STOI	SISDR	WB-PESQ
FSN	0.93 ± 0.07	$\mathbf{5.1 \pm 4.1}$	2.23 ± 0.60	0.90 ± 0.11	$\mathbf{3.1 \pm 4.3}$	2.06 ± 0.55	0.87 ± 0.06	$\mathbf{0.9 \pm 2.6}$	1.60 ± 0.21	0.84 ± 0.10	$\mathbf{-0.8 \pm 3.4}$	1.53 ± 0.24
+ SB	0.92 ± 0.07	4.3 ± 4.2	2.10 ± 0.56	0.89 ± 0.11	2.5 ± 4.6	1.98 ± 0.51	0.86 ± 0.06	-0.3 ± 2.9	1.46 ± 0.19	0.82 ± 0.10	-1.9 ± 3.5	1.42 ± 0.21
+ CSB	0.92 ± 0.07	4.2 ± 4.6	2.11 ± 0.65	0.89 ± 0.11	2.2 ± 5.1	1.99 ± 0.59	0.86 ± 0.06	-0.7 ± 2.9	1.43 ± 0.18	0.82 ± 0.10	-2.4 ± 3.8	1.41 ± 0.21
+ SSB	0.93 ± 0.07	4.8 ± 4.1	2.19 ± 0.59	0.89 ± 0.11	2.6 ± 4.5	1.99 ± 0.52	0.87 ± 0.06	0.6 ± 2.7	1.57 ± 0.20	0.83 ± 0.10	-1.3 ± 3.8	1.49 ± 0.23
+ 3B	0.93 ± 0.07	4.9 ± 4.1	$\mathbf{2.24 \pm 0.60}$	0.90 ± 0.11	2.9 ± 4.6	$\mathbf{2.07 \pm 0.57}$	0.87 ± 0.06	0.7 ± 2.6	$\mathbf{1.61 \pm 0.21}$	0.84 ± 0.10	-1.0 ± 3.7	$\mathbf{1.54 \pm 0.25}$
input	0.86 ± 0.09	-0.2 ± 4.8	1.76 ± 0.67	0.85 ± 0.12	-1.0 ± 5.5	1.89 ± 0.76	0.75 ± 0.07	-4.5 ± 2.9	1.20 ± 0.11	0.74 ± 0.10	-5.2 ± 3.7	1.24 ± 0.16

Section 4.2.3 ("Matched RIRs"), and the other synthesized in rooms corresponding to more challenging acoustic conditions ("Mismatched RIRs"): $\mathrm{RT}_{60} \in [1.0, 1.5]$ s, room size range in $[10, 15] \times [10, 15] \times [4, 6]$ m^3, $D \in [2.5, 4.0]$ m. The dereverberation performance between the baseline and the proposed approaches is evaluated using the Short-time-objective Intelligibility STOI, the Scale Invariant Signal-to-noise ratio (SISDR), and the wide-band Perceptual Evaluation of Speech Quality WB-PESQ defined in Section 3.3.5.

To demonstrate the acoustic system identification capability acquired by the network trained to match RIR characteristics, we compare the energy distribution predicted at the output of each version of the DNN using three combinations of convolutive models and physical losses defined in Section 4.4.2:

— The EDC MSE computed using the Fourier deconvolution method. Given a ground-truth (respectively estimated) RIR h (respectively $\hat{h}$), this corresponds to $\mathcal{L}_\Phi = \left\| \Phi(\hat{h}) - \Phi(h) \right\|_2^2$, with $\Phi(r) = \mathrm{EDC}(r)$ (defined in Section 3.1.2). This metric has also been used to compute the realism of a physically-interpretable dereverberation training objective in [255].

— The EDR MSE computed in the sub-band approximation. This corresponds to $\Phi(r) = \mathrm{EDR}_{\mathrm{dB}}(r)$.

— The EDR MSE computed in the cross-band approximation.

4.5.3 Results and Discussion

Dereverberation

The results for the dereverberation task are presented in Table 4.1. On matched RIRs conditions, our proposed solution, FSN+3B, has a higher WB-PESQ on all datasets and acoustic conditions than the FSN baseline. All physically constrained variants exhibit similar performance in terms of STOI as the baseline. This means that the physical coherence loss and the dereverberation loss can be jointly optimized and that they both converge to equally performing optima in terms of STOI on the space of the DNN weights. The poorer results of our methods compared to the baseline in terms of SISDR can be explained by the DNN encountering difficulty in optimizing the phase of the complex mask mapping $\boldsymbol{Y}$ to $\boldsymbol{S}$ when it is constrained by a convolutive model.

Table 4.2 – RIR estimation scores ± std. on the WSJ0 test set.

	Matched RIRs			Mismatched RIRs		
	EDC	EDR		EDC	EDR	
	Fourier	Sub-band	Cross-band	Fourier	Sub-band	Cross-band
FSN	66.2 ± 28	39.0 ± 12	99.6 ± 24	86.4 ± 15	37.8 ± 7	116.7 ± 6
+SB	60.5 ± 21	**32.7** ± **7**	100.7 ± 22	66.3 ± 16	27.6 ± 6	114.9 ± 7
+CSB	**52.6** ± **24**	34.1 ± 13	**97.8** ± **24**	**63.1** ± **16**	**25.6** ± **4**	**113.6** ± **7**
+SSB	76.4 ± 23	39.9 ± 10	102.9 ± 23	86.2 ± 14	40.4 ± 8	117.9 ± 6
+3B	67.1 ± 27	38.7 ± 11	100.0 ± 24	86.8 ± 15	37.5 ± 7	117.2 ± 6
dry	0.0 ± 0	36.7 ± 10	75.0 ± 19	0.0 ± 0	38.4 ± 8	84.4 ± 12

Considering this metric, the model trained on SSB performs similarly to the model trained on 3B. These losses are the ones that introduce the least constraints on the training and that are the least well-defined (SSB by introducing sub-band modelling errors, and 3B by being unstable). Because these two losses regularize the training in a physically realistic manner, they enable the model to perform better on unseen cases and to generalize to mismatched RIRs and source signals (in the "LibriSpeech" column of the table). Further experiments show that the dereverberation performance remains consistent when high SNR noise is added to the reverberant input of the model at test time. These results reflect FSN's underlying design assumption that both full- and sub-band recursive modelling are needed for the dereverberation task.

RIR estimation

Table 4.2 compares the performance of all proposed approaches with respect to the energy decay of several convolutive models. The line denoted "dry" shows $\mathcal{L}_\Phi(I(\boldsymbol{S}), \boldsymbol{H})$ for each tuple of convolutive model and energy decay. It represents the best theoretical performance that each convolutive model can offer. A very high error and variance can be observed when the RIR has been estimated in the cross-band model. This confirms the results that we obtained in Section 4.2, where we observed that the cross-band deconvolution had a very low robustness. These results are amplified as the lengths of the signals considered (approximately 3 s) is too short for the cross-band method to perform well. This result confirms Avargel's error analysis of the cross-band filtering [62], where he showed that for a given number of cross-bands, and a fixed SNR on the dry and reverberant signal, there exists only one single dry signal length minimizing the non-blind deconvolution mean-squared error. The results suggest that the RIR estimation task competes with the dereverberation task, as indicated by their differing performance rankings. The FSN+CSB variant is performing the best and is capable of modelling the sub-band model even better than the oracle sub-band model $I(\boldsymbol{S})$. This can be explained by the fact that forcing the model output to respect a sub-band model while maintaining its ability to process cross-bands in

Table 4.3 – Dereverberation scores ± standard deviation on the MedleyDB dataset.

	Matched RIRs		Mismatched RIRs	
	FAD	SISDR	FAD	SISDR
FSN	2.39	0.6 ± 5.1	**3.46**	$\mathbf{-3.9 \pm 5.5}$
+ SB	2.26	0.5 ± 5.6	3.96	-4.4 ± 6.4
+ CSB	**2.08**	0.7 ± 6.3	3.71	-4.7 ± 7.2
+ SSB	2.48	$\mathbf{0.7 \pm 5.4}$	3.48	-4.1 ± 5.9
+ 3B	2.45	0.4 ± 5.7	3.49	-4.5 ± 6.4
input	3.57	0.0 ± 6.5	9.02	-5.5 ± 6.5

its latent representation is very efficient to predict the RIR STFT. This assumption is indeed at the core of FSN's design.

Accordingly, a guideline might be to resort to FSN+CSB for the RIR estimation task, and to FSN+3B for the dereverberation task.

Dereverberation on out-of-domain data

A third experiment is conducted to evaluate the proposed methods in a more challenging out-of-domain scenario. Specifically, we assess the generalization of neural dereverberators trained on speech data on singing voice recordings from the MedleyDB dataset [263], using raw vocal tracks without instrumental bleed. As traditional speech dereverberation metrics are not fully appropriate for singing voice, we instead assess the performance using the Fréchet Audio Distance (FAD) [264] computed with VGGish embeddings [265], alongside with the SI-SDR metric.

Results are presented in Table 4.3. On mismatched source but matched RIRs, several of our proposed variants outperform the baseline. Under RIR domain mismatch, both the baseline and our proposed variants exhibit performance degradation, but the decline is more pronounced for our variants trained with losses that explicitely encode RIR characteristics. Indeed, in this evaluation setting of mismatched RIRs, none of our proposed training strategies outperform the baseline. This behaviour suggests that the proposed physical loss regularization effectively drives the neural dereverberator to capture the acoustic properties of reverberation more faithfully. This encouraging result means that our proposed physical loss regularization enables a neural dereverberator to be more robust to source mismatch but does not generalize to unseen reverberation conditions.

4.6 Conclusion

In this chapter, we investigated the interplay between dereverberation and acoustic system identification, formulating the latter as a non-blind deconvolution problem based on the output of a neural dereverberator. We demonstrated that, while tradi-

tional deconvolution approaches are highly sensitive to noise and estimation artifacts, incorporating physical regularization on the late RIR can substantially improve the physical realism of the identified acoustic systems. Building upon this insight, we proposed a physically motivated training loss for neural dereverberators, designed to enforce consistency with room acoustics. This training framework enables the neural dereverberator to simultaneously perform both tasks, compensating for the lack of robustness of deconvolution techniques. It successfully produces physically plausible RIRs without compromising dereverberation performance, and can be applied to out-of-domain sources such as singing voice. However, our results revealed a fundamental limitation: DNNs struggle to jointly optimize for dereverberation quality and physical accuracy, and their performance improvement over the baseline remains marginal, To address these challenges, we introduce in the next chapter a maximum-likelihood reformulation of dereverberation, in which the reverberation process is modeled statistically. This formulation allows us to embed physical consistency directly into the learning objective through explicit modeling of the acoustic system, rather than by means of soft regularization.

Chapter 5

Towards unsupervised dereverberation

Chapter abstract

In this Chapter, we explore the outcome of training state-of-the-art dereverberation models with supervision settings ranging from weakly-supervised to fully unsupervised, relying solely on reverberant signals and a room acoustics model for training.

5.1 Introduction

While the previous approach required a joint supervision of both the dry signal and the RIR, the framework defined in this Chapter deals with data-scarce scenarios. We present a novel training framework in order to overcome the limitation of most data-driven methods not being suited for such settings. Indeed, despite their empirical success, data-driven methods, particularly those based on DNNs, typically require large volumes of supervised training data in the form of paired dry and reverberant signals. These dry signals must be recorded in anechoic conditions, rendering data collection expensive and impractical. Furthermore, supervised systems often exhibit limited generalization to unseen reverberant conditions, reducing their robustness in real-world scenarios. Even unsupervised methods that learn dry speech priors without requiring paired data remain limited by the availability of dry recordings.

To overcome these limitations, we introduce a monaural dereverberation framework that operates in a fully unsupervised manner, relying solely on reverberant signals for training. Our approach integrates an explicit acoustic model into the dereverberation process and offers two main contributions:

— We derive a novel Maximum Likelihood (ML) formulation for dereverberation guided by a parametric reverberation model. While ML-based methods have

been proposed for multichannel settings [33, 266], this is, to the best of our knowledge, the first application of such a formulation to monaural dereverberation with explicit use of acoustic parameters.

— We develop an unsupervised learning strategy that integrates the physical model of reverberation into a deep neural network. Our method demonstrates strong performance under realistic conditions using a small subset of acoustic parameters (from 100 RIRs) to train the reverberation model. Despite this limited data regime, our method maintains competitive dereverberation performance.

Three articles are summarized here:

— [51] Louis Bahrman, Mathieu Fontaine, and Gaël Richard. « A Hybrid Model for Weakly-Supervised Speech Dereverberation ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5

— [52] Louis Bahrman, Mathieu Fontaine, and Gaël Richard. « Déréverbération Non-Supervisée de La Parole Par Modèle Hybride ». In: *XXXe Colloque Francophone de Traitement Du Signal et Des Images*. Strasbourg, France: GRETSI, Aug. 2025

— [53] Louis Bahrman, Marius Rodrigues, Mathieu Fontaine, and Gaël Richard. « U-DREAM: Unsupervised Dereverberation Guided by a Reverberation Model ». In: *IEEE Transactions on Audio, Speech and Language Processing* 34 (2026), pp. 1552–1563

In this manuscript, we mostly focus on the experimental setting of the third article, which extends the first ones by the contributions presented above and generalizes their results to real RIRs.

5.2 Theoretical formulation of the dereverberation problem

In this Section, we derive the formulation of our proposed Unsupervised Dereverberation system guided by a REverberAtion Model (U-DREAM).

The noisy time-domain formulation of Eq. (3.6) can be expressed in the time-frequency domain using Eq. (2.12) by introducing the STFT of the noise term:

$$\boldsymbol{Y}_{f,t} = \mathcal{C}(\boldsymbol{S}, h)_{f,t} + \mathcal{E}_{f,t}, \tag{5.1}$$

where $\boldsymbol{\mathcal{E}} \triangleq \{\mathcal{E}_{f,t}\}_{f,t=0}^{F-1,T_s-1} \in \mathbb{C}^{F\times T_y}$ is the STFT of the additive noise. In this work, we assume that all $\mathcal{E}_{f,t} \sim \mathcal{N}_{\mathbb{C}}(0, \nu^2)$ are Independant and Identically Distributed (IID)

complex Gaussian variables, and that the dry signal STFT is deterministic. Under such assumption,

$$\boldsymbol{Y}_{f,t} \mid h; \boldsymbol{S}, \Theta \sim \mathcal{N}_{\mathbb{C}}(\mathcal{C}(\boldsymbol{S}, h)_{f,t}, \nu^2). \tag{5.2}$$

Introducing the finite RIR h as a random vector conditioned only on the acoustic parameters Θ, the total probability distribution of $\boldsymbol{Y}$ is:

$$p(\boldsymbol{Y}; \boldsymbol{S}, \Theta) = \int p(\boldsymbol{Y} \mid h; \boldsymbol{S}, \Theta) p(h; \Theta) dh. \tag{5.3}$$

The integral can be expressed as an expectation with respect to the distribution $p(h; \Theta)$, and its negative log-likelihood is:

$$-\log p(\boldsymbol{Y}; \boldsymbol{S}, \Theta) = -\log \mathbb{E}_{p(h|\Theta)} \left[p(\boldsymbol{Y} \mid h; \boldsymbol{S}, \Theta) \right]. \tag{5.4}$$

Jensen's inequality can be applied to obtain an upper bound of the generally intractable negative log-likelihood of the expectation:

$$-\log p(\boldsymbol{Y}; \boldsymbol{S}, \Theta) \leq \mathbb{E}_{p(h|\Theta)} \left[-\log p(\boldsymbol{Y} \mid h; \boldsymbol{S}, \Theta) \right]. \tag{5.5}$$

Finally, replacing the conditional probability distribution of $\boldsymbol{Y} \mid h; \boldsymbol{S}, \Theta$ from Eq.(5.2) yields:

$$-\log p(\boldsymbol{Y}; \boldsymbol{S}, \Theta) \leq \mathbb{E}_{p(h|\Theta)} \left[\| \mathcal{C}(\boldsymbol{S}, h) - \boldsymbol{Y} \|_F^2 \right] + C, \tag{5.6}$$

where $\| \cdot \|_F$ denotes the Frobenius norm, and $C = -FT_y \log(\pi \nu^2)$ is a constant with respect to $\boldsymbol{S}$ and Θ.

The task of dereverberation using an acoustic model can be formulated as a maximum likelihood estimation problem, where the goal is to jointly estimate both the STFT of the dry speech signal $\boldsymbol{S}$ and the acoustic parameters Θ of a parametric reverberation model, given the observed reverberant STFT $\boldsymbol{Y}$. Formally, this is expressed as:

$$\underset{\boldsymbol{S}, \Theta}{\operatorname{argmax}}\, p(\boldsymbol{Y}; \boldsymbol{S}, \Theta). \tag{5.7}$$

In this work, we solve a relaxed version of the problem, where we minimize the upper bound of the negative log-likelihood obtained from Eq. (5.6):

$$\hat{\boldsymbol{S}}, \hat{\Theta} = \underset{\boldsymbol{S}, \Theta}{\operatorname{argmin}}\, \mathbb{E}_{p(h|\Theta)} \left[\| \mathcal{C}(\boldsymbol{S}, h) - \boldsymbol{Y} \|_F^2 \right] \tag{5.8}$$

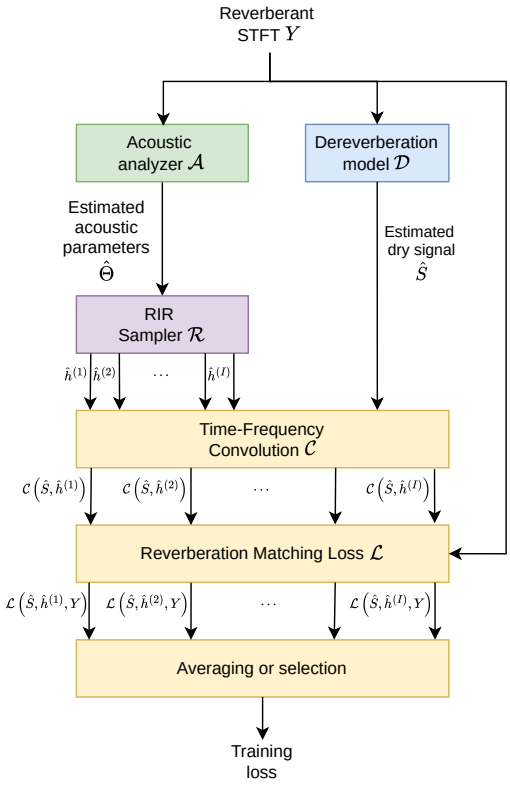


Figure 5.1 – Overview of the U-DREAM method.

5.3 Dereverberation guided by a reverberation model

5.3.1 Overview

To address the ill-posed nature of solving Eq. (5.8) in the monaural case, we introduce a model-based deep learning framework. Specifically, we propose to replace the direct optimization of $\boldsymbol{S}$ and Θ with two trainable, model-based mappings: a dereverberation module $\mathcal{D}_{w_D} : \boldsymbol{Y} \mapsto \hat{\boldsymbol{S}}$ and an acoustic analyzer $\mathcal{A}_{w_A} : \boldsymbol{Y} \mapsto \hat{\Theta}$, where w_D and w_A denote the weights that parametrize each model (sometimes omitted for sake of clarity). Substituting the optimization of $\boldsymbol{S}$ and Θ with the optimization of w_D and w_A, the problem in Eq. (5.8) becomes:

$$w_D, w_A \in \underset{w_D, w_A}{\operatorname{argmin}} \mathbb{E}_{p\left(h|\mathcal{A}_{w_A}(\boldsymbol{Y})\right)} \left[\left\| \mathcal{C}\left(\mathcal{D}_{w_D}(\boldsymbol{Y}), h\right) - \boldsymbol{Y} \right\|_F^2 \right] \tag{5.9}$$

We propose to solve this equation by training both models $\mathcal{A}$ and $\mathcal{D}$ using SGD on a dataset of reverberant signals. The overall forward pass of the proposed framework is illustrated in Figure 5.1 and summarized as follows: given a reverberant signal $\boldsymbol{Y}$, the dereverberation module $\mathcal{D}$ produces an estimated signal $\hat{\boldsymbol{S}}$. Simultaneously, the acoustic analyzer $\mathcal{A}$ (see Section 5.3.4) estimates the corresponding acoustic parameters $\hat{\Theta}$ from $\boldsymbol{Y}$. To approximate the expectation $\mathbb{E}_{p(h|\hat{\Theta})}$ by Monte-Carlo sampling, one

or more RIRs $\hat{h} \in \mathbb{R}^{N_h}$ are drawn from a reverberation sampler $\mathcal{R}$ (See Section 5.3.3). Each RIR $\hat{h}$ is convolved with the estimated dry signal $\hat{\boldsymbol{S}}$ via the operator $\mathcal{C}$, producing one or more estimated reverberant STFTs $\hat{\boldsymbol{Y}}$. The optimization objective is expressed through a reverberation matching loss function $\mathcal{L}$ (described in Section 5.3.5), which quantifies the distance between the estimated reverberant STFT $\hat{\boldsymbol{Y}}$ and the observed reference $\boldsymbol{Y}$.

5.3.2 Adaptations to various supervision scenarios

Unsupervised dereverberation

The training framework presented above is tailored for unsupervised dereverberation, where both the acoustic analyzer $\mathcal{A}$ and the dereverberation module $\mathcal{D}$ are jointly trained. However, this framework potentially has trivial solutions if both $\mathcal{A}_{w_A}$ and $\mathcal{D}_{w_D}$ are trained jointly from scratch. Specifically, the acoustic analyzer $\mathcal{A}$ could converge to predicting acoustic parameters corresponding to an anechoic environment (e.g., low RT_{60} or high DRR), enabling the dereverberation module $\mathcal{D}$ to simply learn an identity mapping, thereby bypassing the intended dereverberation process. To mitigate this issue, we adopt a two-stage training strategy. The acoustic analyzer $\mathcal{A}$ is first pre-trained using available supervised data. Once $\mathcal{A}$ has been pre-trained, we proceed to train $\mathcal{D}$ while keeping $\mathcal{A}$ frozen. This staged approach is motivated by the relative difficulty of the two tasks: predicting Θ is inherently easier than estimating $\boldsymbol{S}$, and the acoustic analyzer can typically be trained with significantly less data than the dereverberation module.

Weak supervision

In case ground-truth acoustic parameters are known and available at training, the acoustic parameters estimated by the acoustic analyzer can be replaced by the oracle acoustic parameters. In such cases, the optimization objective is given by:

$$\underset{w_D}{\operatorname{argmin}}\, \mathbb{E}_{p(\hat{h}|\Theta)} \mathcal{L}(\mathcal{D}_{w_D}(\boldsymbol{Y}), \hat{h}, \boldsymbol{Y}) \tag{5.10}$$

Strong supervision

This framework can still be leveraged when the exact RIR h instead of the dry source is available at training. In this case, the RIR is purely deterministic, and the distribution $p(h \mid \Theta)$ is replaced by a Dirac measure $\delta_\Theta(h)$, where h corresponds exactly to the true RIR used to generate the reverberant signal in the training set. Under this setting, the expectation in Eq. (5.9) simplifies, yielding the following optimization objective:

$$\underset{w_D}{\operatorname{argmin}}\, \mathcal{L}(\mathcal{D}_{w_D}(\boldsymbol{Y}), h, \boldsymbol{Y}). \tag{5.11}$$

Training-less variant

An alternative formulation of Eq. (5.8) is also considered, wherein only the acoustic analyzer $\mathcal{A}_{w_A}$ is employed, without the need for a dereverberation module. This training-less variant assumes that $\mathcal{A}_{w_A}$ has been pre-trained. Given a reverberant signal $\boldsymbol{Y}$ and fixed analyzer parameters w_A, the following optimization problem is solved directly for $\hat{\boldsymbol{S}}$:

$$\hat{\boldsymbol{S}} = \underset{\boldsymbol{S}}{\operatorname{argmin}}\, \mathbb{E}_{p(h|\mathcal{A}_{w_A}(\boldsymbol{Y}))} \left[\|\mathcal{C}(\boldsymbol{S}, h) - \boldsymbol{Y}\|_F^2 \right] \tag{5.12}$$

This variant is named *training-less* because it does not require a dereverberation module $\mathcal{D}_{w_D}$ to be fitted on a large dataset but rather directly optimizes the dry STFT $\boldsymbol{S}$ on a per-sample basis given the output of the acoustic analyzer.

Each part of the overall framework presented above is now detailed.

5.3.3 Reverberation modeling

The RIR sampler is responsible for generating one or more RIRs h sampled from the conditional distribution $p(h \mid \Theta)$, where Θ represents the acoustic parameters. The goal is to synthesize an RIR whose key characteristics, such as RT_{60} and DRR, match those of the original setting in which the reverberant signal was recorded. In this work, the RIR sampler is based on Polack's model, detailed in Eq. (3.4).

In this Section, we detail the stochastic RIR model used in this work.

Late reverberation model formulation

The RIR sampler is based on a statistical model of reverberation, first discovered by Polack [87], and formally generalized and demonstrated by Badeau under the SWFT [105]. Note that Badeau's SWFT explicits the frequency dependency of the Wigner-Ville distribution of the RIR at high frequencies. Experiments conducted in [267] showed that introducing such a frequency dependency is not straightforwardly beneficial for dereverberation performance. Hence, the RIR sampler used is a simplified version of Polack's and Badeau's models, that assumes that late reverberation is an exponentially decaying white noise. To further simplify the model and ensure proper scaling, all direct-path energy is concentrated at the peak of the RIR, with normalization applied such that the peak value is 1. The resulting reverberation sampler $\mathcal{R}$ is defined as:

$$\mathcal{R}(\Theta)[n] = \begin{cases} b[n] e^{-\frac{3\ln(10)}{\mathrm{RT}_{60} f_s} n} & \text{if } n > n_L \\ 1 & \text{if } n = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.13}$$

where $b[n] \sim \mathcal{N}(0, \sigma^2)$.

According to Polack's model, the reverberant energy E_R is given by:

$$E_R = \int_{n_L}^{+\infty} \sigma^2 e^{-2t/\tau} dt \tag{5.14}$$

$$= \sigma^2 \frac{\tau}{2} e^{-2n_L/\tau}, \tag{5.15}$$

where τ is defined in Eq. (3.5)

Assuming the direct-path energy is normalized to 1, σ can be computed from the target DRR as:

$$\sigma = \sqrt{\frac{2e^{2n_D/\tau}}{\tau \, \mathrm{DRR}}} \tag{5.16}$$

Late reverberation model adaptation to synthetic RIRs

A slight difference has to be taken in account when working with synthetic RIRs. In particular, ISM-based RIR generators such as `Pyroomacoustics` [115] often produce RIRs that exhibit nonzero energy at the zero frequency. To better match such RIRs, we modify the noise component of Polack's model. Specifically, when encountering non-centered RIRs, we match the zero-frequency energy of the ground-truth RIR while maintaining a flat spectral response in the other bands by drawing the noise term $b(n)$ from a half-normal distribution $\left|\mathcal{N}(0, \sigma^2)\right|$. We will experimentally verify that this distribution better matches the distribution of synthetic RIRs in Section 5.5.2.

Artifacts caused by early echoes

In the previous section, we detailed the late reverberation model used in our work. However, this statistical model has been proven to be valid only under the conditions defined in [105], namely with a randomly fixed source position, and therefore neglects the contribution of early echoes which are mostly dependant to the source and microphone positions within a room, as seen in part 3.1.1. In this section, we provide an analysis of the interaction of our model with an early echoes model defined hereunder.

Immediate properties of artifacts Our reverberant model assumes a fixed, deterministic source position, that causes a deterministic contribution of the early echoes to the RIR. We further consider that, after a given index noted N_l, the RIR is purely stochastic and strictly follows the model detailed in (5.13), hence we assume that h_e is causal, of finite support. The artifacts caused by early echoes are still linear and time-invariant, and as such, can be modeled by a convolutive model, and there exists an impulse response $h_c \in \mathbb{R}^{2N_h}$ such that the law of the measured RIR h would be equal to the law of:

$$h_c \star (\delta + \hat{h}_l). \tag{5.17}$$

and $\hat{h}_l$ also follows Polack's model, with the same reverberation time RT_{60} and DRR as the observed late reverberation h_l. The distribution of $\hat{h}_l$ remaining the same as the distribution of h_l is justified by the fact that, unlike early echoes which are defined by both the source and microphone positions within a room, late reverberation only depends on a statistical model built upon room geometry and hence its statistical model is unrelated to early echoes. First, h_c is a finite impulse because if it was not, h would not be a finite RIR either.

Moreover, h_c cannot be non-causal, because if h_c was non-causal and of finite support, there would exist an index $m < 0$ such that $h_c[m] \neq 0$ and $\forall n, n < m, h_c[n] = 0$. Then $h_c \star (\delta + \hat{h}_l) = \sum_{k=-\infty}^{+\infty} (\delta + \hat{h}_l)[k] h_c[m-k] = (\delta + \hat{h}_l)[0] h_c[m] = h_c[m] \neq 0$, which would yield a contradiction as h is causal.

Matching covariance matrices Using the constraint that $\hat{h}_l$ and h_l have the same law on late reverberation (i.e. in the domain $[\![N_l, : N[\![$), we derive some properties about h_c. The domain $[\![N_l : N[\![$ is finite, so $\hat{h}_l$ and h_l are two multivariate Gaussian distributions with the same statistics (because their RT_{60} and DRR are equal). As a consequence, their covariance matrices exist and should match on $N_l \leq n < N$. On one hand, let Σ_{h_l} be the covariance matrix of the observed h_l, restricted to the domain $[\![N_l, N [\![$. On the other hand, the convolution of $\hat{h}_l$ by h_c can be expressed as a linear operator in the form of the Toeplitz matrix $H_c \in \mathbb{R}^{(N-N_l)\times(N-N_l)}$. Then, matching the distribution of h with the distribution provided in Eq. (5.17) can be expressed on the second-order statistics of h on $[\![N - N_l, N[\![$ as:

$$\Sigma_{h_l} = H_c \Sigma_{h_l} H_c^{\mathsf{T}} \tag{5.18}$$

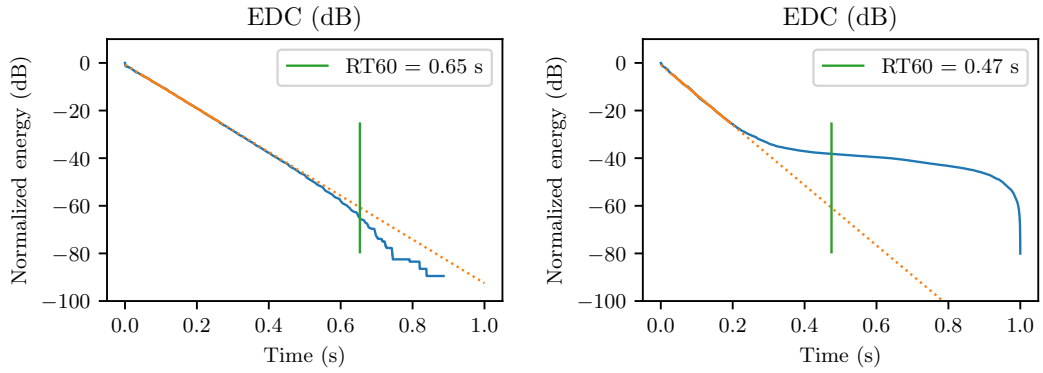
$$= \left(H_c \Sigma_{h_l}^{1/2}\right) \left(H_c \Sigma_{h_l}^{1/2}\right)^{\mathsf{T}} \tag{5.19}$$

$$\text{iff} \quad I_{(N-N_l)} = \left(\Sigma_{h_l}^{-1/2} H_c \Sigma_{h_l}^{1/2}\right) \left(\Sigma_{h_l}^{-1/2} H_c \Sigma_{h_l}^{1/2}\right)^{\mathsf{T}} \tag{5.20}$$

As the diagonal elements of Σ_{h_l} are all positive (they are constructed from the exponential family of the RIR variances according to Polack), they form an orthogonal basis of $\mathbb{R}^{N-N_l}$. Hence, as $\Sigma_{h_l}^{-1/2} H_c \Sigma_{h_l}^{1/2}$ is orthonormal, H_c is also an orthogonal matrix. Then $H_c H_c^{\mathsf{T}}$ is not only symmetric, but diagonal which means that H_c is diagonal. Going back to the time-domain representation of h_c, we have for all $n, 1 \leq n < N - N_l, h_c[n] = 0$

Hence, assuming Polack's model of late reverberation, and in the case where no information about the distribution of early echoes can be derived, an appropriate choice to model early echoes is to consider them silent, as the mismatch in early echoes causes no distortion of the late RIR until its last N_l elements.

Note that this reasoning is only valid for a late reverberation model based on a Gaussian statistical model, and that reasoning fails on half-Gaussian variables used



(a) Synthetic RIR

(b) Measured RIR

Figure 5.2 – Energy decay curves. The expected energy decay is shown in orange.

in synthetic RIRs.

5.3.4 Acoustic Analysis

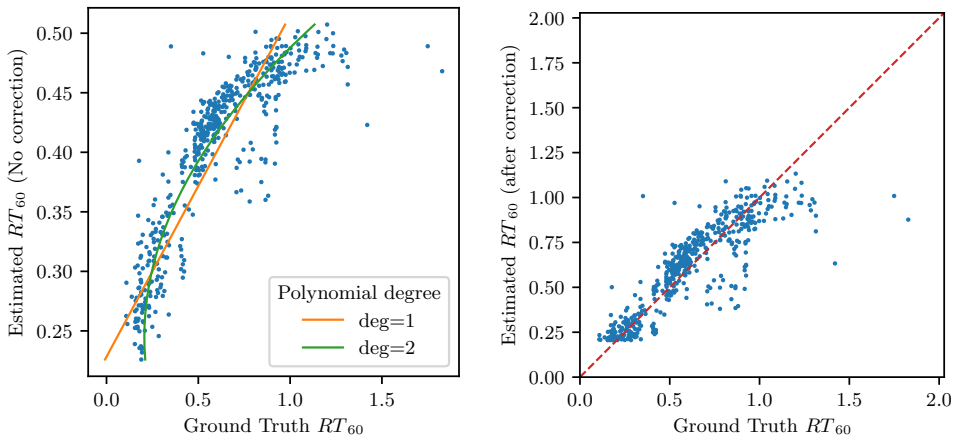
The acoustic analysis module estimates the reverberation parameters Θ, namely RT_{60} and DRR, which are used to guide the reverberation synthesis process described in Section 5.3.3. We distinguish between two modes of analysis: the *Non-blind* case in which the ground-truth RIR h is known, and the *blind* case, where only the reverberant signal $\boldsymbol{Y}$ is available. The blind estimator must be tuned using a small set of ground-truth pairs $(\boldsymbol{Y}, h)$, or by solving (5.9) using an oracle dereverberation module $\mathcal{D}$.

Non-blind analysis

In the non-blind setting, acoustic parameters are derived directly from the RIR. We follow the procedure described in [28, Chapter 2], based on linear regression of the EDC. Figure 5.2 illustrates the energy decay of a synthetic and a measured RIR. The expected energy decay according to Polack's model is shown in orange. We observe that the late reverberation's tail diverges from the linear decay predicted by Polack's model. Measured RIRs present a noisy reverberation tail whereas synthetic RIRs' energy decreases too fast. This would result in a biased RT_{60} (too low on synthetic RIRs and too large on measured ones), and DRR (too high on synthetic RIRs and too low on measured ones), which would be detrimental to the modelling of late reverberation.

To mitigate this issue, we restrict the analysis to the dynamic range in which Polack's model is considered to be valid, namely between $-5\,\mathrm{dB}$ and $-25\,\mathrm{dB}$, where both the measured and theoretical EDC match. The energy on this range, noted E_{25}^{5}, is expected to follow:

$$E_{25}^{5} = \sigma^2 \frac{\tau}{2} \left(e^{-2T_5/\tau} - e^{-2T_{25}/\tau} \right), \tag{5.21}$$



(a) Before polynomial fitting (b) After polynomial (degree 2) fitting

Figure 5.3 – RT_{60} estimation results on a dataset of 500 measured wet speech signals and RIRs.

where T_5 and T_{25} represent the time it takes for the EDC to decrease by 5 dB and 25 dB respectively. The reverberation time RT_{60} is first computed from the slope of the EDC, which is independent of the total energy of the RIR. Then, using the measured value of E_{25}^{5} and Eq. (5.21), the parameter σ is derived, leveraging the relation between τ and RT_{60} given in Eq. (3.5).

Finally, σ can be directly reused in the reverberation sampler $\mathcal{R}$ detailed in Eq. (5.13), or to compute the DRR using Eq. (5.15). Note that this approach excludes the contribution of late reverberation noise to the estimated DRR, thereby improving alignment with the synthesis model.

Blind analysis

In the blind case, both RT_{60} and DRR must be estimated from the reverberant STFT $\boldsymbol{Y}$. For the RT_{60} estimation, we adopt the method proposed in [177]. This approach computes local energy decay in specific regions of $\boldsymbol{Y}$. A polynomial mapping is then used to match the median RT_{60} found in the time-frequency domain to its corresponding time-domain value. Based on preliminary experiments measuring the accuracy of RT_{60} estimation, we employ a second-order polynomial, instead of the first-order polynomial used in the original study. The three polynomial coefficients are tuned on a very small calibration dataset of 100 pairs $(\boldsymbol{Y}, \mathrm{RT}_{60})$.

Figure 5.3 shows the results of the RT_{60} estimation performed using Prego's method, on measured RIRs. Results for the first-order polynomial are shown in orange and for the second-order in green. We observe that the second-order polynomial shows a more accurate estimation of the RT_{60} (Pearson correlation between ground-truth and estimated RT_{60}: $\rho = 0.88$) than the first-order polynomial ($\rho = 0.85$).

For the DRR estimation, we adopt the BiLSTM model of [182], which predicts two time-frequency domain energy masks. The ratio between the masked spectrograms is used to estimate the DRR.

5.3.5 Reverberation matching loss

Formulation

In this section, we detail the *Reverberation Matching (RM)* loss, which quantifies the distance between the ground-truth reverberant STFT $\boldsymbol{Y}$ and the estimated reverberant STFT produced by convolving the estimated dry signal $\hat{\boldsymbol{S}}$ with an RIR sampled by the reverberation sampler $\hat{h}$.

The loss inside the expectation in Eq. (5.8) assumes IID complex additive noise on each bin of the reverberant STFT $\boldsymbol{Y}$. We denote this loss as:

$$\mathcal{L}_{\mathbb{C}}(\hat{\boldsymbol{S}}, \hat{h}, \boldsymbol{Y}) \triangleq \left\| \mathcal{C}(\boldsymbol{S}, \hat{h}) - \boldsymbol{Y} \right\|_F^2. \tag{5.22}$$

We also consider an additional term, denoted $\mathcal{L}_{\text{MAG}}$, that computes a distance in the log-magnitude space as in [268]:

$$\mathcal{L}_{\text{MAG}}(\hat{\boldsymbol{S}}, \hat{h}, \boldsymbol{Y}) \triangleq \left\| \log \frac{1 + |\boldsymbol{Y}|}{1 + \left| \mathcal{C}(\hat{\boldsymbol{S}}, \hat{h}) \right|} \right\|_F^2. \tag{5.23}$$

The total per-sample reverberation matching loss $\mathcal{L}$ combines these two terms as the weighted sum

$$\mathcal{L} = \mathcal{L}_{\mathbb{C}} + \lambda \mathcal{L}_{\text{MAG}}. \tag{5.24}$$

Preliminary experiments showed that adding the log-magnitudes loss term $\mathcal{L}_{\text{MAG}}$ enabled faster convergence of our proposed framework. Combining these two losses with the expectation in Eq. (5.9), our training loss is the expectation

$$\mathbb{E}_{p(\hat{h}|\mathcal{A}_{w_A}(\boldsymbol{Y}))} \mathcal{L}(\hat{\boldsymbol{S}}, \hat{h}, \boldsymbol{Y}). \tag{5.25}$$

We now detail some sampling strategies used to compute this expectation.

Loss variants

The expectation in Eq. (5.25) is computed via Monte Carlo sampling. Depending on the sampling strategy adopted, the loss has different physical interpretations regarding the underlying reverberant scene. In particular, multiple draws of the RIR from the sampler $\mathcal{R}(\Theta)$ can be viewed as simulating different virtual microphone positions within a room characterized by Θ. Variability between these draws reflects potential variations in the observed reverberant signal due to changes in microphone

placement, even though the target signal $\boldsymbol{Y}$ corresponds to a single physical microphone. In this work, we consider three sampling strategies:

— **Single**: a single RIR $\hat{h}$ is drawn from $\mathcal{R}(\Theta)$ for each reverberant spectrogram $\boldsymbol{Y}$. Across different epochs, different RIR draws may be associated with the same $\boldsymbol{Y}$. Physically, this corresponds to simulating one virtual microphone in a room with parameters Θ, while acknowledging that the exact microphone position may not match that of $\boldsymbol{Y}$. The loss is computed as:

$$\mathcal{L}_{\text{single}} = \mathcal{L}(\hat{\boldsymbol{S}}, \hat{h}, \boldsymbol{Y}). \tag{5.26}$$

— **Average**: multiple RIRs $\{\hat{h}^{(i)}\}_{i=1}^{I}$ are drawn from $\mathcal{R}(\Theta)$, and the loss is computed as the mean across these draws. This corresponds to simulating I virtual microphones in the same room and averaging their contribution to the loss. In our experiments, $I = 10$ draws is used:

$$\mathcal{L}_{\text{avg}} = \frac{1}{I} \sum_{i=1}^{I} \mathcal{L}(\hat{\boldsymbol{S}}, \hat{h}^{(i)}, \boldsymbol{Y}). \tag{5.27}$$

— **Best**: as in the "Average" strategy, multiple RIRs are drawn, but only the draw yielding the lowest loss is used for backpropagation. This can be interpreted as searching for the virtual microphone position in the room that best explains $\boldsymbol{Y}$, and encouraging the model to match this optimal configuration:

$$\mathcal{L}_{\text{Best}} = \min_{i} \mathcal{L}(\hat{\boldsymbol{S}}, \hat{h}^{(i)}, \boldsymbol{Y}). \tag{5.28}$$

Balancing the loss terms

To ensure that $\mathcal{L}_{\mathbb{C}}$ and $\mathcal{L}_{\text{MAG}}$ contribute equally during training, we adopt the GradNorm method [261] to automatically adjust the weight λ such that the Frobenius norm of the gradients of both losses with respect to the model weights are equal.

In the case of the $\mathcal{L}_{\text{Best}}$ strategy, λ and the selected index of the optimal RIR draw i are interdependent, which results in a bi-level optimization problem. To circumvent this issue, we approximate the GradNorm method by solving for each sampled RIR i:

$$\left\| \frac{\partial \mathcal{L}_{\mathbb{C}}(\hat{\boldsymbol{S}}, \hat{h}^{(i)}, \boldsymbol{Y})}{\partial \mathcal{C}(\hat{\boldsymbol{S}}, \hat{h})} \right\| = \left\| \lambda_i \frac{\partial \mathcal{L}_{\text{MAG}}(\hat{\boldsymbol{S}}, \hat{h}^{(i)}), \boldsymbol{Y}}{\partial \mathcal{C}(\hat{\boldsymbol{S}}, \hat{h})} \right\|. \tag{5.29}$$

We then select the RIR draw $\hat{i}$ that yields the lowest loss, and backpropagate gradients only through $\mathcal{C}(\hat{\boldsymbol{S}}, \hat{h}^{(i)})$ with the corresponding optimal weight $\lambda_{\hat{i}}$.

5.4 Experimental setup

We tested our proposed dereverberation guided by a reverberation model framework under several supervision scenarios on datasets of both real and synthetic RIRs. This section presents our experimental protocol.

5.4.1 Datasets

We evaluate our procedure on both synthetic and real reverberation. The following datasets are considered:

1. *EARS-Reverb*: the Expressive Anechoic Recordings of Speech (EARS) dataset [96] is composed of high-quality dry speech signals recorded from various speakers and diverse content. We use its dereverberation benchmark, EARS-Reverb, generated by convolving anechoic speech from EARS with RIRs sampled from publicly-available datasets. This dataset is provided at a sampling rate of 48 kHz. Our task being dereverberation at 16 kHz, we resample both clean and reverberant signals to this sampling rate.
2. *EARS-ISM*: we consider a simpler reverberant dataset, composed of anechoic audio from EARS and simulated reverberation synthesized using the ISM method as in Section 4.2.3.
3. *Out-Of-Domain*: we also evaluate the generalization performance on models trained on synthetic RIRs and tested on real RIRs.

5.4.2 Neural dereverberators

Our proposed dereverberation supervision settings are, in theory, adaptable to any speech model. In practice, such deep-learning-based models are very diverse and their performance is variable. Indeed, it is relevant to consider how the reverberation-aware training interacts with various dry speech priors. We present dereverberation results for three neural dereverberators:

— BiLSTM [234]: this simple model consists of a 2-layer bidirectional LSTM model followed by a linear layer, performing sub-band recursive processing of the STFT magnitudes. Since this model is only able to process magnitude masks, it can be considered as phase-agnostic, and the underlying speech model is that both dry and reverberant speech are Gaussian complex noise, with circular invariance.

— FSN [235]: this more powerful LSTM called FullSubNet is capable of cross-band processing and, since it is able to generate a complex-valued mask, it can be considered as a phase-aware model. This model has also been used in reverberation-aware training [255, 50].

— TFL [236]: TF-Locoformer represents a state-of-the-art model for speech enhancement and dereverberation, which is very powerful and expressive thanks to its self-attention module capable of global modeling and convolution handling local modeling.

5.4.3 Misc settings

During training, 4-second excerpts are processed at 16 kHz. STFT processing is done using a 512-sample Hann window with an overlap of 50 %. All DNNs are trained using Adam optimizer, with a learning rate of 10^{-4}, and early stopping based on the ESTOI metric on a validation set is used.

5.5 Results

In this section, we present the results of our experiments to train our proposed framework for the tasks of acoustic parameter estimation and dereverberation under several supervision scenarios.

The performance is evaluated using the SI-SDR [245], ESTOI [247], WB-PESQ [250] and SRMR [251] metrics, reviewed in Section 3.3.5.

5.5.1 Dereverberation with strong supervision

Figure 5.4 reports the performance of the strong supervision implemented as our proposed reverberation matching loss with the ground-truth RIR (defined in Eq. (5.11)). Specifically, the baseline training losses are as follows: for the BiLSTM model, MSE between the ground-truth and estimated magnitudes; for FullSubNet, MSE between the ground-truth ideal and estimated complex ratio mask; and for the TF-Locoformer (TFL), a combination of time-domain loss and a multiresolution spectral loss between the ground-truth and estimated dry signals. On synthetic RIRs, training with supervision from the ground-truth dry signal generally outperforms supervision via the proposed reverberation matching loss. The only exceptions occur with FullSubNet which yields higher performance on the SISDR metric when trained with the reverberation matching loss, and with TF-Locoformer which performs better on the ESTOI metric under the same conditions. In contrast, on real RIRs, an opposite trend is observed. For all metrics except SRMR, models trained with supervision from the exact RIR either achieve better performance or exhibit differences that are not statistically significant, according to a Wilcoxon non-parametric test (p-value < 0.001). One explanation for this behavior is that supervision with the reverberation matching loss offers a more balanced optimization of magnitude and phase information, leading to faster convergence, which is crucial on this harder dataset.

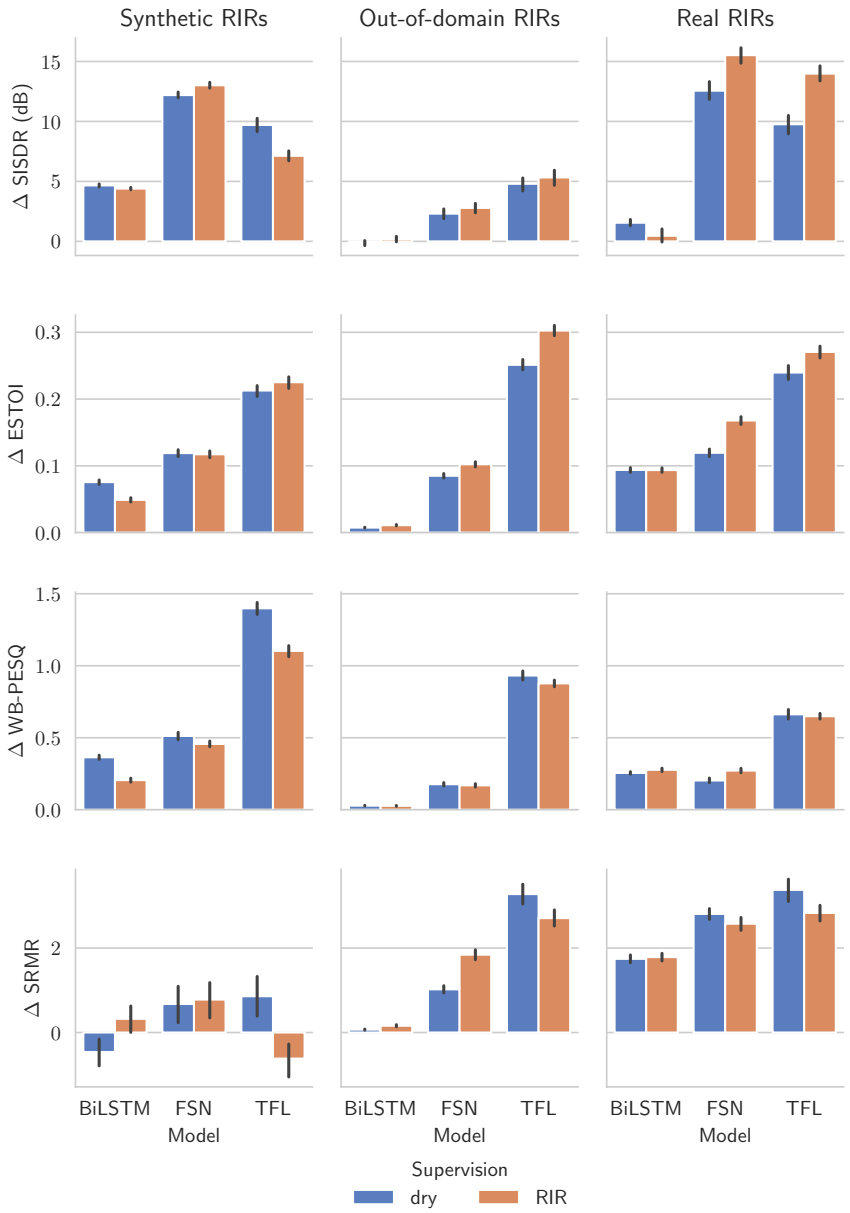


Figure 5.4 – Dereverberation with strong supervision: Comparison of the proposed training loss (supervision by the RIR) and the baseline training loss (supervision by the dry signal). Results are presented as the relative improvement compared to the reverberant input. The 95 % confidence intervals are indicated by black lines.

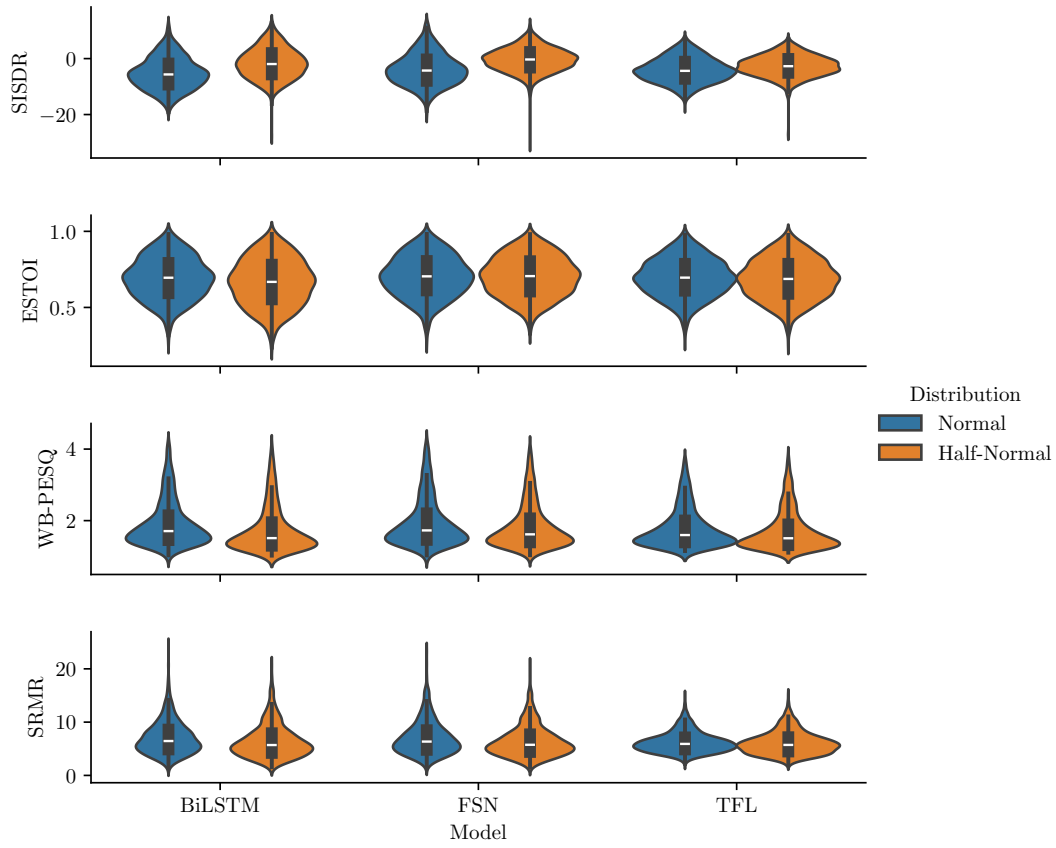


Figure 5.5 – Dereverberation with weak supervision: comparison between drawing the late reverberation using a normal or half-normal distribution on the EARS-Reverb dataset.

5.5.2 Weak supervision for dereverberation

Preliminary experiment: synthetic RIRs distribution

In this experiment, we measure the benefit of using a half-normal distribution in our proposed RIR sampler for synthetic RIRs.

Figure 5.5 measures the influence of the late reverberation model on weakly-supervised dereverberation. Two settings for the RIR sampler on the EARS-ISM dataset described in Eq. (5.13) are compared: one using a normal distribution for the noise $b(n) \sim \mathcal{N}(0, \sigma^2)$ and one using a half-normal distribution $b(n) \sim \left|\mathcal{N}(0, \sigma^2)\right|$, as described in Section 5.3.3. The sampling strategy used is the *single*, described in Section 5.3.5

Drawing the late reverberation from a half-normal distribution seems to be very beneficial to the dereverberation performance on the SI-SDR metric at the cost of a slight degradation of other metrics. The effects on the ESTOI and WB-PESQ are however quite limited: two one-sided paired T-Tests with the null hypothesis that the mean of the results are less than a given interval apart are performed. These

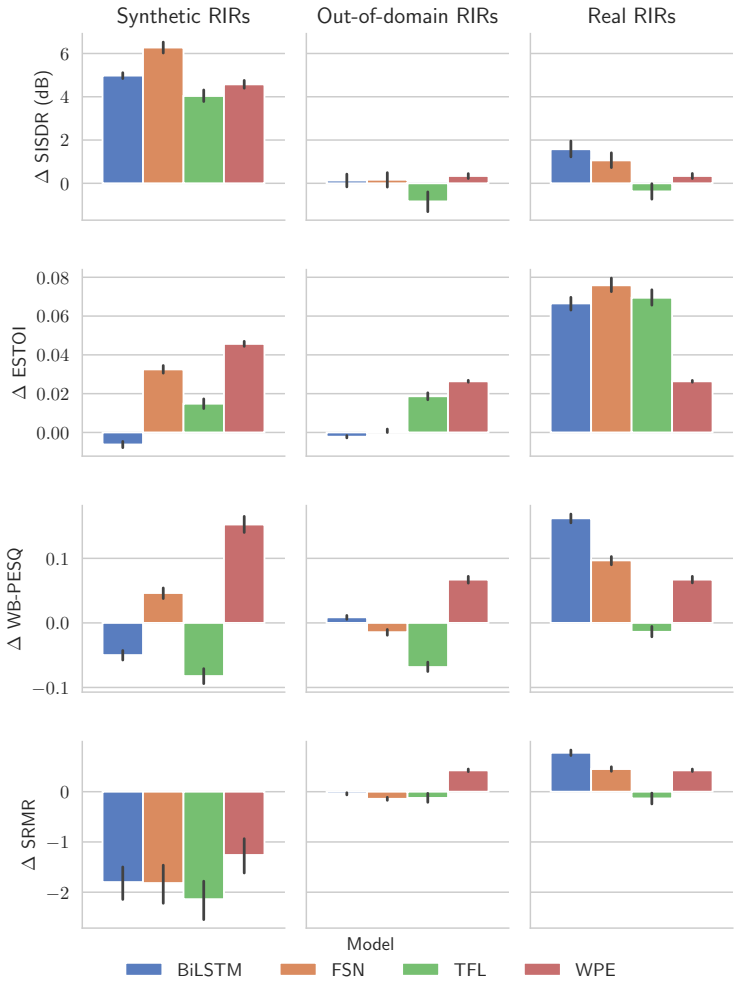


Figure 5.6 – Weak supervision for dereverberation: Comparison of the weakly-supervised methods with the WPE baseline. All models (BiLSTM, FSN, TFL) are trained with the "Single" reverberation matching loss variant. Results are presented as the relative improvement compared to the reverberant input. The 95 % confidence intervals are indicated by black lines.

tests show that (with p-value $<1 \times 10^{-4}$) for all source models, the mean WB-PESQ difference is less than 0.2, and the ESTOI difference is lower than 0.02 for all models except the BiLSTM.

Influence of the dereverberation model

Figure 5.6 summarizes the performance of dereverberation models trained under weak supervision with the "Single" variant. The WPE baseline is provided for comparison. In comparison to the WPE baseline, few models achieve significant improvements. On real RIRs, only the BiLSTM model consistently outperforms WPE across all metrics, while FullSubNet performs better on all metrics except SRMR. When comparing models, FullSubNet achieves equal or superior performance to BiL-

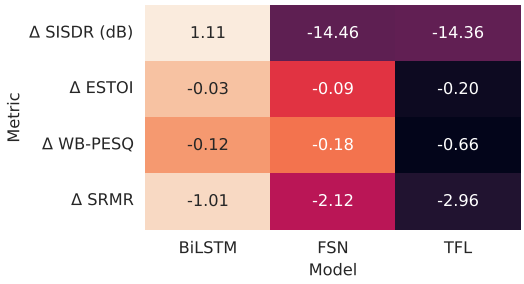


Figure 5.7 – Weak dereverberation: Degradation caused by training using weak supervision compared to strong supervision on the EARS-Reverb dataset, using the "Single" loss variant.

STM on synthetic RIRs, while BiLSTM outperforms FSN on real RIRs for SISDR, WB-PESQ, and SRMR, with FullSubNet performing better only on ESTOI. The TFL consistently underperforms under the weak supervision regime, likely due to its high model capacity relative to the simplicity of the reverberation model.

Comparison with strong supervision

Figure 5.7 shows the degradation of using our proposed reverberation matching loss in a weak supervision setting using the oracle reverberation parameters, compared to the strong supervision of the exact RIR (from the EARS-Reverb dataset). The results vary greatly from model to model, and remain consistent between metrics: BiLSTM shows less degradation when going from strong to weak supervision than FSN, which in turns outperforms TFL. The surprisingly good performance of BiLSTM compared to other models can be explained by the fact that this model is optimized for spectral magnitude masking and not for phase estimation. Consequently, reverberation by Polack's model, which greatly perturbates the phase of the reverberant signal, better matches the underlying phase-agnostic assumption of the BiLSTM model than reverberation using ground-truth RIRs. This underlies the need to have consistency between the deep-learning-based speech model and reverberation sampler underlying priors.

Influence of the loss variant

Figure 5.8 compares the improvement of each loss variant over the reverberant input, for each source model on the EARS-Reverb dataset (since it is only single domain where our methods outperform WPE). Across all models, datasets, and metrics, no supervision variant ("Single", "Average", or "Best" microphone strategy) consistently outperforms the others. In most cases, the differences are not statistically significant. The limited effectiveness of the "Best microphone" strategy may be attributed to an insufficiently constrained training objective. For this reason, throughout the remain-

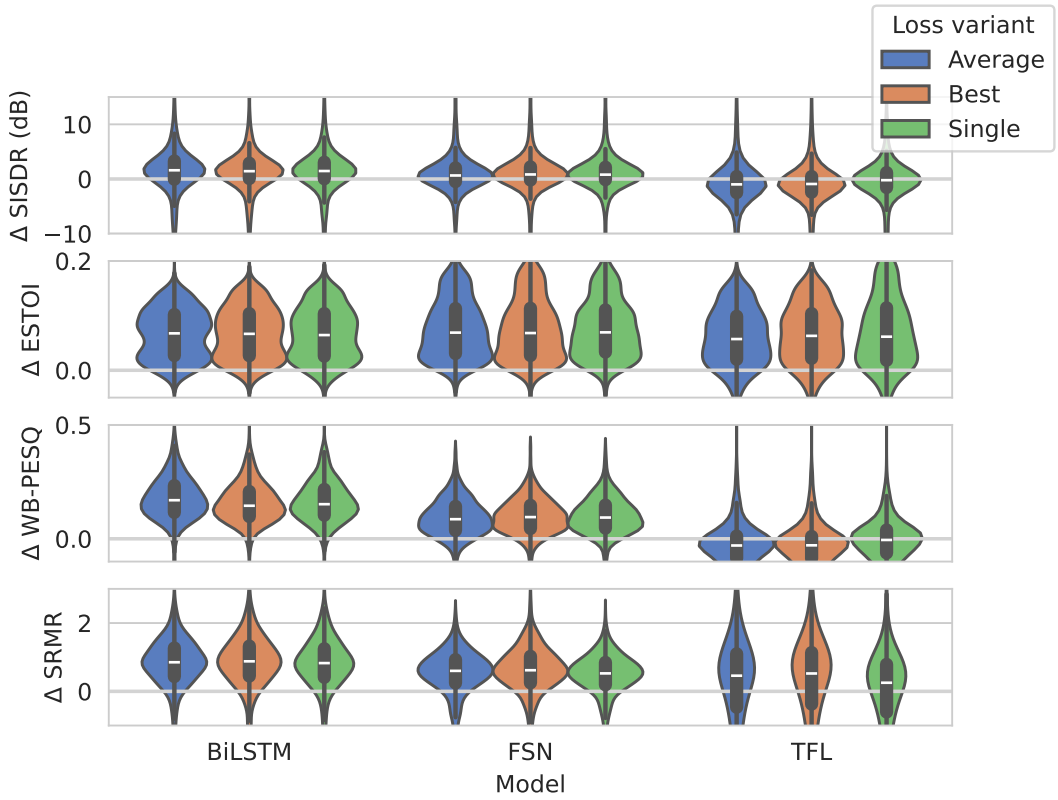


Figure 5.8 – Weak dereverberation: Comparison of the RM loss variants on the EARS-Reverb dataset. Results are presented as the relative improvement compared to the reverberant input.

der of our experiments, we adopt the *single* reverberation matching loss variant since it represents the lowest computational cost compared to other variants, at no cost in performance.

5.5.3 Acoustic parameter estimation with various supervision

Influence of the optimization objective and dataset size

Figure 5.9 presents the performance of the acoustic analyzer $\mathcal{A}_w$ for DRR estimation under two supervision strategies: (i) the proposed RM loss, which leverages paired reverberant and dry signals, and (ii) the Parameter Matching (PM) loss, which requires DRR annotations for each reverberant signal.

The evaluation considers three training data regimes (full dataset, 5 % subset, and 100 training samples), across synthetic, out-of-domain, and real RIRs. Performance is reported using both the PM loss (i.e., MSE between estimated and ground-truth DRR) and the RM loss as evaluation metrics, to analyze cross-objective consistency.

Models trained to minimize the PM loss perform well when sufficient data is available, but their accuracy degrades significantly as training data is reduced. In contrast,

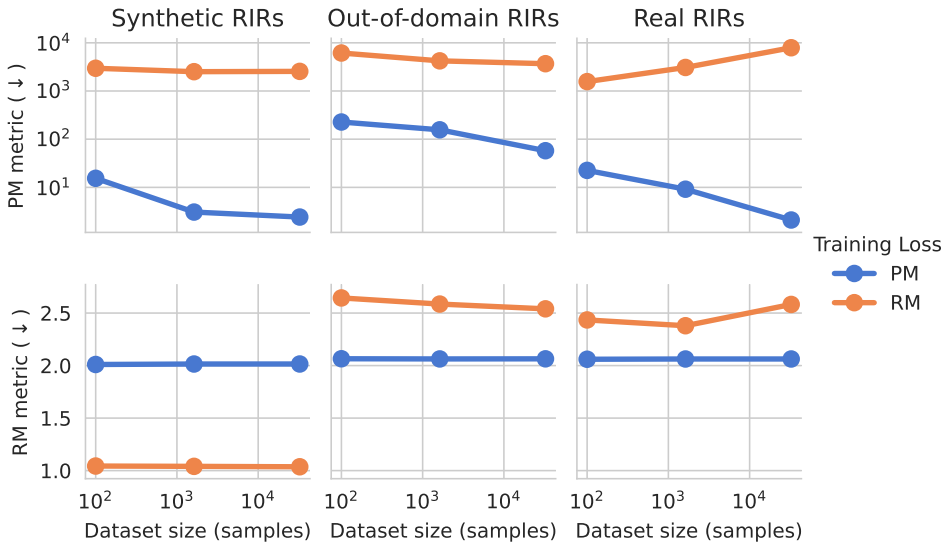


Figure 5.9 – Acoustic parameter estimation: Comparison of the Parameter matching (PM) and our proposed Reverberation Matching (RM) loss for DRR estimation for various training set sizes.

RM loss optimization is more challenging on real RIRs, though in-domain training still outperforms training on mismatched (out-of-domain) data. Notably, on real RIRs, models trained with RM loss increasingly diverge from their PM-trained counterparts as more data is provided. This suggests an inconsistency between the two objectives, likely due to late-reverberation noise in real RIRs, which affects the resynthesized target in RM but is not modeled by the RIR sampler. The resulting mismatch leads the model to overestimate reverberant energy, thereby distorting the inferred DRR. However, on synthetic RIRs, both models trained with PM and RM objectives yield similar results on synthetic RIRs regardless of data size.

Robustness of the proposed framework to acoustic parameter estimation errors

Figure 5.10 provides a deeper analysis of the relationship between PM and RM metrics on the EARS-Reverb dataset. Interestingly, RM loss performance appears largely independent of DRR estimation accuracy. Even models that overfit under PM supervision (e.g., when trained on small datasets) exhibit comparable RM performance. This indicates that optimizing the RM objective does not require precise DRR estimation and suggests that $\mathcal{A}_w$ need not be finely tuned to be effective. Finally, we observe that $\mathcal{A}_w$ achieves strong performance even with limited data, supporting our design choice to pre-train this module independently before dereverberation model training.

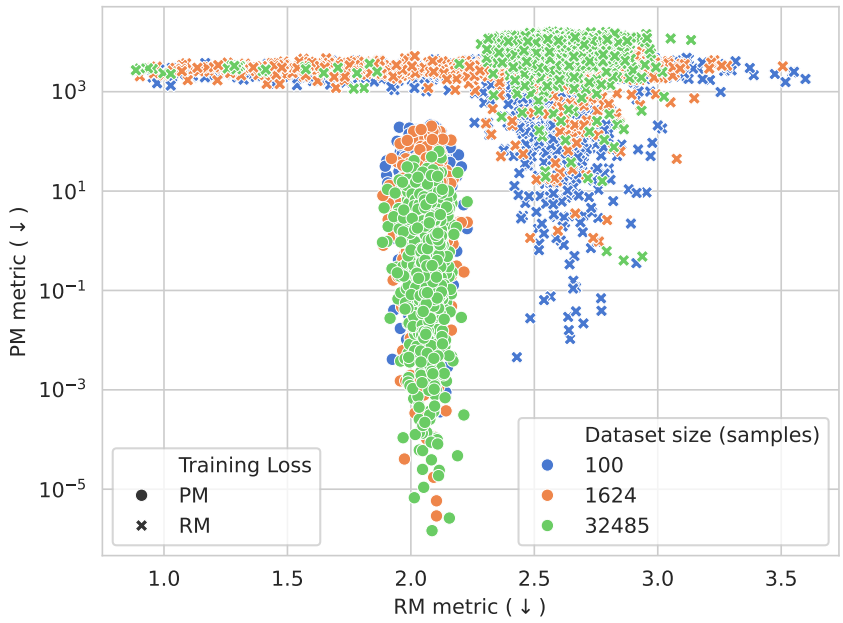


Figure 5.10 – Acoustic parameter estimation: Relative performance of each estimated blind DRR estimation sample from EARS-Reverb on the PM and RM metrics

5.5.4 Unsupervised dereverberation

In this Section, we evaluate the impact of the blind acoustic analyzer module on the final performance of the dereverberation model trained using our proposed framework.

Unsupervised training of the neural dereverberator

While the dereverberation network is retrained from scratch, the reverberation analyzers are reused from the previous experiment and remain frozen. This allows us to examine how the dereverberation performance varies as a function of both the loss used to train the reverberation model and the quantity of supervision available during its training. The dereverberation model itself is always trained on the full set of reverberant-only data, which reflects realistic deployment scenarios where access to clean signals or acoustic parameters is limited but not reverberant signals. We focus on the BiLSTM model trained on real RIRs with the "single" dereverberation loss variant, as it achieved the best performance on SISDR and WB-PESQ among models that significantly outperformed the WPE baseline. Results are presented in Figure 5.11. The best results are obtained when the reverberation analyzer was trained using the parameter matching (PM) loss, consistent with prior findings where PM minimized both PM and RM metrics during DRR estimation. As expected, dereverberation performance improves with increased training data for the reverberation model. A notable exception is observed when using a reverberation model trained with PM loss

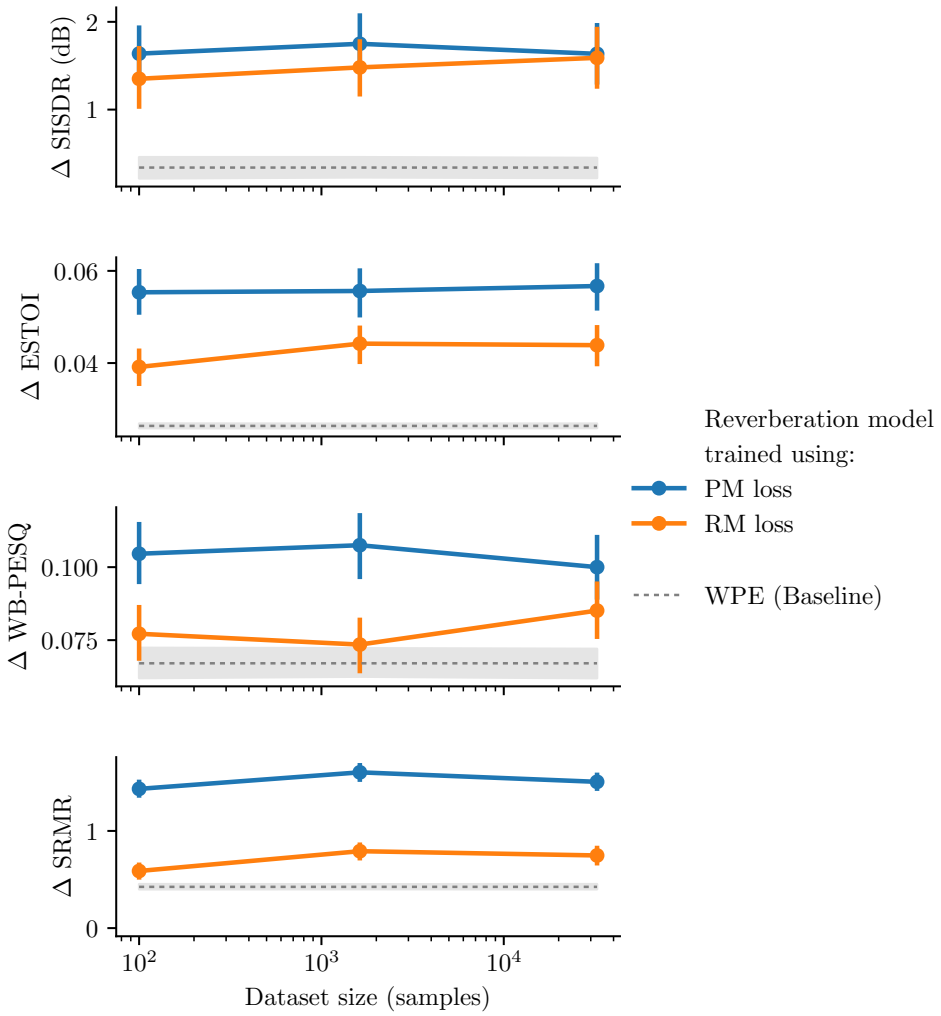


Figure 5.11 – Unsupervised dereverberation: Improvement over the reverberant input for different annotation dataset sizes. For the unsupervised methods, the x-axis represents the quantity of data used to pre-train the acoustic analyzer before training the dereverberation model.

on a large dataset, which underperforms on the SISDR and WB-PESQ metrics due to a very small degradation of the performance of the acoustic analyzer on the RM metric when the dataset size increases (as seen in the previous experiment). Despite this, we observe strong dereverberation performance when the analyzer is trained using the PM loss, even when using only 100 samples of acoustic parameters, demonstrating the robustness of our proposed framework in low-resource conditions, and outperforming the unsupervised baseline of WPE on all metrics.

5.5.5 Training-less variant

In this part, we study the direct optimization of the dry signal given acoustic parameters, described in Section 5.3.2.

Results in an unsupervised setting

Figure 5.12 shows the performance of the training-less variant compared to the BiLSTM variant (that is trained using the framework described in the previous section), on the EARS-Reverb dataset. Both variants use the Acoustic analyzer trained using 100 samples of paired wet signals and acoustic parameters. One hundred gradient steps are performed at inference for the training-less variant. The training-less variant fails to properly dereverberate the input signal, in the sense that no significant (p-value $<1\times10^{-4}$) improvement over the reverberant input can be produced. This indicates that training a dereverberation model on a dataset of reverberant signals is still necessary for effective unsupervised dereverberation. The time complexity required to perform the 100 gradient descent steps required at inference for the training-less method makes it less effective than the WPE baseline, that already shows a good dereverberation performance after only 3 iterations.

Analysis of a gradient step

We can confirm these results by observing how the optimization of the training-less variant loss function defined in Eq. (5.12) influences the distance of the dereverberated signal with both the reverberant input and the dry signal. Several key criteria are supposed to be met for a meaningful optimization:

— The loss between the synthesized reverberant signal and the reverberant input should be lower than the loss between the dry signal and the reverberant input:

$$\mathcal{L}(\boldsymbol{S}, \hat{h}, \boldsymbol{Y}) < \mathcal{L}(\boldsymbol{S}, \delta, \boldsymbol{Y}). \tag{5.30}$$

— The synthesized reverberant signal should be closer to the reverberant input

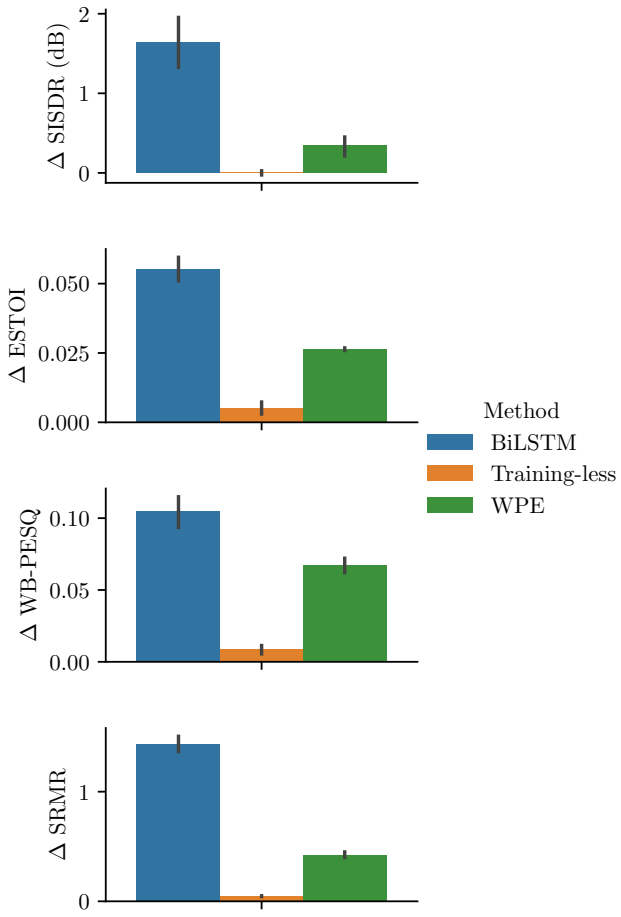


Figure 5.12 – Comparison of the BiLSTM and training-less variants.

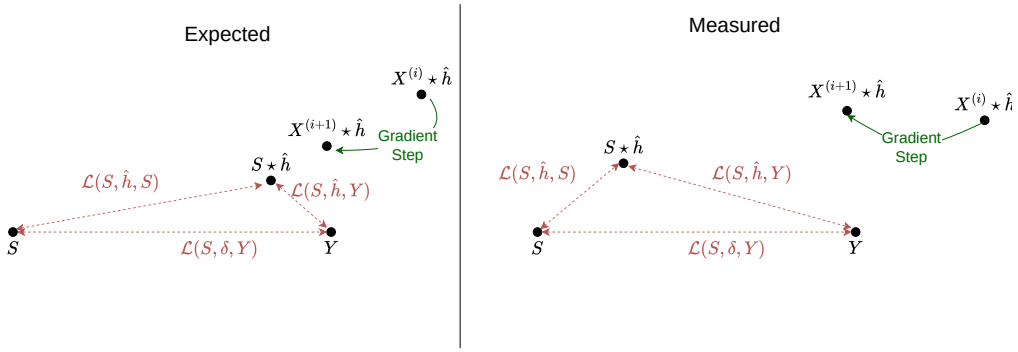


Figure 5.13 – Schematic view of the expected and measured reverberation losses.

than to the source signal:

$$\mathcal{L}(\boldsymbol{S}, \hat{h}, \boldsymbol{Y}) < \mathcal{L}(\boldsymbol{S}, \hat{h}, \boldsymbol{S}). \tag{5.31}$$

— Between two iterations of the reverberation-matching loss optimization for dereverberation, the dereverberated signal should go closer to the dry signal and farther from the reverberant signal:

$$\mathcal{L}(\boldsymbol{X}^{(i)}, \hat{h}, \boldsymbol{Y}) \leq \mathcal{L}(\boldsymbol{X}^{(i+1)}, \hat{h}, \boldsymbol{Y}) \Rightarrow \mathcal{L}(\boldsymbol{X}^{(i)}, \delta, \boldsymbol{S}) \geq \mathcal{L}(\boldsymbol{X}^{(i+1)}, \delta, \boldsymbol{S}), \tag{5.32}$$

where $\boldsymbol{X}$ is the signal being optimized.

These proposed criteria can be visualized on Figure 5.13.

We simulate the first gradient step of the training-less variant of our method, with $\boldsymbol{X}$ initialized at $\boldsymbol{Y}$, on the test dataset of EARS-Reverb. We sample the RIR using oracle parameters, in order to alleviate RIR parameter estimation errors. As in the previous experiments, we use the *single* loss variant.

We perform a non-parametric Wilcoxon paired signed-rank test, with a two sided alternative. The first criterion is met, indicating that the proposed reverberation sampler brings the dry signal closer to the reverberant input. The second criterion is not met, meaning that the reverberation synthesized using our proposed sampler is still very weak compared to adding reverberation by convolving from a measured impulse response. Our proposed reverberation procedure cannot even bring a ground-truth dry signal halfway towards its reverberant counterpart. While this result does not mean the proposed procedure is unable to perform dereverberation, it indicates that more optimization steps might need to be performed in order to obtain a satisfying dereverberated signal. The third criterion is not met, indicating that optimizing the proposed reverberation-matching loss fails to produce a meaningful dry signal. In other words, minimizing the reverberation matching loss fails to optimize towards a dry signal The proposed reverberation matching loss is not a good proxy in order to dereverberate a signal. This objective misalignment can be explained by the inaccu-

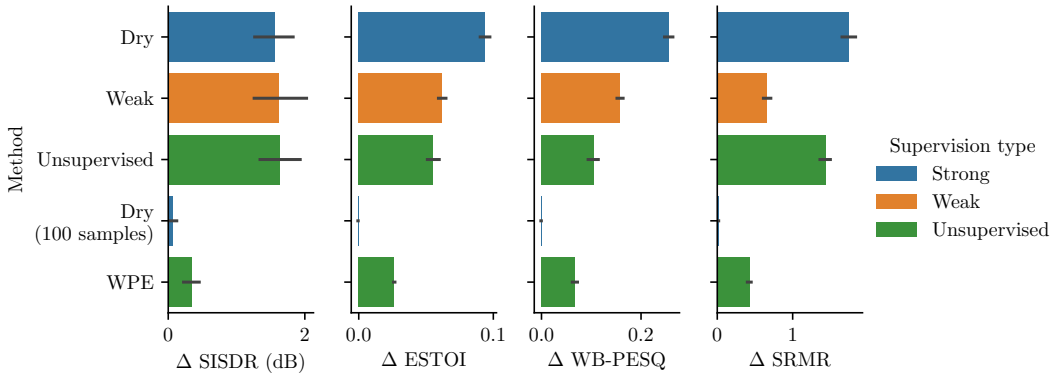


Figure 5.14 – Summary of the results for the BiLSTM model. Improvement of the reverberant input for all methods.

racy of our proposed reverberation sampling procedure. Indeed, all criteria are met when using ground-truth RIRs instead of sampled ones .

While the training-less dereverberation procedure fails to produce a meaningful dereverberated signal, the data-driven procedure still manages to enhance its reverberant input after being trained with our proposed reverberation-matching loss as an objective. This can be explained by the fact that, the deep-learning-based optimization procedure is trained to optimize a nonlinear mapping from the reverberant signal to the dry signal, rather than optimizing each sample individually. The DNN hence learns the reverberation-matching objective on a full dataset of reverberant signals, and can therefore act as a learnt prior over the distribution of dry signals. This regularization explains the satisfactory results of our proposed hybrid dereverberation framework.

5.5.6 Summary of the best-performing method

For comparison, we also evaluate strongly supervised variants of our proposed method in data-limited scenarios.

Figure 5.14 summarizes the performance of the BiLSTM model in each supervision setting detailed above on the EARS-Reverb dataset. We also compare the results of this method when the model is trained using only 100 pairs of dry and reverberant signals.

When trained with only 100 paired dry and reverberant samples, the BiLSTM dereverberation model exhibits markedly poor performance, failing to generalize effectively. These findings suggest that leveraging a small number of in-domain annotations to pre-train the acoustic analyzer, followed by unsupervised training of the dereverberation model, yields superior performance compared to directly training a strongly supervised dereverberation model under data-limited or mismatched conditions.

5.6 Conclusion

In this chapter, he dereverberation problem is formulated as a maximum likelihood estimation of the dry signal and acoustic parameters. We propose to solve this problem using a hybrid approach that is adaptable to several supervision settings. Our most data-efficient method outperforms state-of-the-art unsupervised dereverberation methods by leveraging only 100 samples of acoustic parameters such as direct-to-reverberant ratio and reverberation time. Our experiments show that, although our approach can be generalized to any dereverberation and reverberation models, their underlying priors should match.

Chapter 6

Conclusion and Perspectives

Chapter abstract

In this chapter, we summarize our findings, in order to answer the research questions. We provide an overview of the tools we developed throughout this research. Additionally, by acknowledging the limitations of our approaches, we suggest future research directions.

6.1 Summary of our findings

In Chapter 4, we explored the intersection of dereverberation and acoustic system identification, leveraging deconvolution techniques to establish a novel framework for non-blind acoustic system identification. Given a reverberant measured signal and a dry source estimated by a neural dereverberator, we analyzed whether it is possible to determine the acoustic system in which the reverberant signal was recorded. In such a blind setting, the acoustic system identification problem boils down to a deconvolution problem. However, this approach introduces a new challenge: the neural dereverberator's output is prone to estimation artifacts due to imperfect dereverberation, which complicates the deconvolution process.

We evaluated existing dereverberation techniques in this non-ideal scenario and found that they are not robust to noise. Specifically, deconvolved RIRs exhibit noisy tails, which contradicts the expected exponentially decaying late reverberation described by acoustic models. To address this issue, we introduced a novel non-blind acoustic system identification technique, that incorporates a regularization on the tail of the deconvolved RIRs, ensuring it has a physically plausible decay.

Furthermore, we put forward a novel physical training loss for neural dereverberators, aiming to train the network in a way that ensures physical consistency, thus improving its performance in non-blind system identification tasks. Through this approach, we evaluated several neural dereverberator architectures and observed that

not all were capable of optimizing the proposed physical loss. In fact, improvements in the physical loss did not always translate to significant gains in dereverberation performance, highlighting a key challenge: most neural architectures struggle to jointly optimize for reverberation matching and dereverberation. This setting enables us to evaluate several architectures of neural dereverberators on their compatibility with room acoustics.

To overcome this limitation, we proposed a regularization strategy that combines the physical loss with a traditional dereverberation loss enabling the neural dereverberator to simultaneously perform acoustic system identification and dereverberation.

While the proposed method demonstrates the feasibility of linking dereverberation and deconvolution-based acoustic system identification, its performance remains bounded by the inherent instability of deconvolution and the use of a regularization technique that only imposes a soft constraint. In Chapter 5, we address these limitations by introducing a probabilistic framework that models the reverberant signal as the convolution of a deterministic source and a stochastic RIR parametrized by a small set of deterministic acoustic parameters. Within this formulation, both the deconvolution and the acoustic parameter estimation problems are cast as a maximum-likelihood estimation problem. Instead of training the DNN using a physics-driven loss regularization, we define a novel training framework in which the loss is computed as a distance between the observed reverberant and a synthesized reverberant signal sampled from the acoustic model. This design incorporates physical consistency directly into the loss function through explicit modeling of the reverberation process, rather than through indirect regularization. A key advantage of this approach lies in its flexibility regarding supervision. The proposed framework supports a wide range of supervision settings, starting from strongly supervised training, using the exact RIR as characterization of the acoustic system, towards an unsupervised regime, including a variety of weakly-supervised configurations.

6.2 Answer to the research questions

After investigating both sides of dereverberation and acoustic system identification, we are now able to answer the two research questions that drive our study:

1. To what extent are existing neural dereverberators consistent with room acoustics?
2. How to leverage room acoustics models in data-driven methods for dereverberation?

We have shown that the ability of dereverberation models to yield a sufficiently accurate sound source estimate to perform acoustic system identification is hindered by the low robustness of existing deconvolution techniques. In practice, pre-trained dereverberation models generally fall short in this regard. However, by jointly optimizing

the neural networks to perform both dereverberation and acoustic system identification, we demonstrated that certain neural architectures could learn to retrieve realistic RIRs at no cost on their dereverberation performance. This finding establishes that some neural dereverberators, when carefully trained, can be made consistent with fundamental physical properties of room acoustics.

Throughout this thesis, we explored two complementary strategies for integrating room acoustics models with data-driven dereverberation: in Chapter 4 we introduced *soft constraints* on the neural dereverberator, by regularizing its training loss using physical priors. In Chapter 5, we proposed to integrate *hard constraints* on the training framework of a neural dereverberator, structuring it around an explicit reverberation synthesis process. These two approaches respectively embody the dual operators at the core of reverberation modeling: deconvolution for matching RIRs, and convolution for matching reverberant signals. Our results demonstrate that hybrid models combining data-driven dereverberation with model-based reverberation leverage the performance of data-driven methods and the interpretability and adaptability to limited data scenarios provided by model-based approaches.

In other words, we have proposed several hybrid deep-learning approaches that could foster understanding of both sonic *spaces* and sonic *sources*.

6.3 List of our contributions

Throughout this research, we have developed three novel hybrid approaches for dereverberation and acoustic system identification:

— **A non-blind acoustic system identification method**: this method yields physically realistic RIRs by enforcing a constraint on their energy decay. We have shown that this approach remains effective when the dry source is estimated from the reverberant signal by a neural dereverberator.

— **A joint training method for blind acoustic system identification and dereverberation**: this approach enables a neural dereverberator to simultaneously perform both tasks, compensating for the lack of robustness of deconvolution techniques. It successfully produces physically plausible RIRs without compromising dereverberation performance, and can be applied to out-of-domain sources such as singing voice.

— **A flexible training strategy for dereverberation**: by formulating dereverberation in a maximum-likelihood framework, we enable a neural dereverberator to be trained under the supervision of a stochastic and parametric reverberation model. This novel framework does not only train the neural dereverberator in a physically consistent manner, but also allows to use less data for training, harnessing the true power of hybrid approaches.

6.4 Limitations of our contributions

While our contributions represent significant advancements, several limitations need to be addressed in future research.

One key limitation arises from our approach to modeling deconvolution. Although initially promising, this strategy proved to be less robust than anticipated. The inherent instability of deconvolution techniques, particularly when the dry source is estimated from the reverberant signal, hindered the overall effectiveness of the method.

Moreover, further investigation is needed into the expressiveness of different DNNs architectures for acoustic modeling. While some of the architectures we explored showed potential, we acknowledge that the diversity of available models means that a more exhaustive analysis is necessary to draw solid conclusions on the compatibility of data-driven methods and model-based reverberation. Additionally, we did not test state-of-the-art dereverberation methods on their ability to perform acoustic system identification, although it could offer valuable insights and potentially enhance the performance of our proposed methods.

Lastly, our model-based hybrid training techniques caused the computational requirements and resource usage to increase compared to the baseline supervised training. While we ensured that the computational requirements for inference remained the same as the baseline, such a computational cost remains an important consideration.

6.5 Perspectives

Among our contributions, the framework for dereverberation formulated in Chapter 5 offers particularly promising perspectives for further exploration. This framework can be extended in several ways, accounting for generative models of varied audio sources, a wider range of reverberation models, and other inverse problems in signal processing.

6.5.1 Generative source models

The framework proposed in Chapter 5 could be framed around a probabilistic generative source model rather of a deterministic one, for better interpretability and performance. Indeed, while a deterministic source can be considered as having a deterministic position, a generative source can be interpreted as having a stochastic position within a room, thereby naturally aligning with the assumption of the SWFT used in this work. A straightforward implementation would be to define a maximum a posteriori estimation framework accounting for a stochastic sound source, instead of the maximum likelihood estimation of the deterministic dry source proposed in Chapter 5. Another advantage of such an approach is that a generative neural dereverberator would be able to model the joint distribution of dry and reverberant signals,

and would hence be better suited to a stochastic reverberant signal. Such a framework could also be used to model the mapping between a reverberant signal distribution and a joint source and reverberation distribution as being bijective, calling for flow-matching [269] techniques to invert reverberation and explore the continuous field between dry and reverberant signals.

Furthermore, generative source models are able to model more complex data distributions than discriminative methods, making this extension directly applicable to a wider range of sources, such as singing voices or musical instruments. This extension is also meaningful as our proposed framework is adapted to domains where anechoic data is hardly available.

6.5.2 Improvements of the reverberation model

The stochastic reverberation model used in this work could be further refined to improve realism and generalization. Recent advances in blind acoustic system identification and data-driven reverberation modeling, particularly those based on DDSP frameworks, could be leveraged to increase the physical accuracy of the model. Moreover, a deterministic representation of early reflections could be incorporated by enforcing sparsity constraints, following approaches such as [270].

An additional extension would be to move towards a multichannel formulation. While the proposed framework can already generate virtual microphone channels by sampling multiple RIRs from the stochastic reverberation model, we have shown that this strategy yields no significant improvement in the single-channel case. In a true multichannel context, however, spatial consistency constraints could be explicitly modeled, potentially leading to more coherent and robust dereverberation.

Such extensions are expected to enhance the performance of the unsupervised and weakly supervised variants of the proposed framework, whose effectiveness is currently limited by the simplicity of the underlying Polack model.

6.5.3 Application to other inverse problems

The proposed framework, by explicitly disentangling the sound source from reverberation effects, naturally extends to a broader class of inverse problems. A first direction is acoustic matching, where the ability of the model to jointly capture a source and room characteristics could be rigorously evaluated. More generally, the concept of supervising a source reconstruction model through a differentiable degradation process can be applied beyond dereverberation to a wide range of signal restoration tasks, including image deconvolution, bandwidth extension, and audio declipping. Such extensions would further demonstrate the versatility of the proposed flexible supervision training framework across domains.

References

[1] Heinrich Kuttruff and Michael Vorländer. *Room Acoustics*. 7th ed. Boca Raton: CRC Press, June 2024 (cit. on pp. 1, 30, 31).

[2] Steven J. Waller. « Hear Here: Prehistoric Artists Preferentially Selected Reverberant Spaces and Choice of Subject Matter Underscores Ritualistic Use of Sound ». In: *Between Worlds: Understanding Ritual Cave Use in Later Prehistory*. Ed. by Lindsey Büster, Eugène Warmenbol, and Dimitrij Mlekuž. Cham: Springer International Publishing, 2019, pp. 251–264 (cit. on p. 1).

[3] Nico F. Declercq et al. « A Theoretical Study of Special Acoustic Effects Caused by the Staircase of the El Castillo Pyramid at the Maya Ruins of Chichen-Itza in Mexico ». In: *The Journal of the Acoustical Society of America* 116.6 (Dec. 2004), pp. 3328–3335 (cit. on p. 1).

[4] Andrea Farnetani, Nicola Prodi, and Roberto Pompoli. « On the Acoustics of Ancient Greek and Roman Theaters ». In: *The Journal of the Acoustical Society of America* 124.3 (Sept. 2008), pp. 1557–1567 (cit. on p. 1).

[5] Jürgen Meyer. *Acoustics and the Performance of Music*. Frankfurt/Main: Verlag Das Musikinstrument, 1978 (cit. on p. 1).

[6] M. R. Schroeder, D. Gottlob, and K. F. Siebrasse. « Comparative Study of European Concert Halls: Correlation of Subjective Preference with Geometric and Acoustic Parameters ». In: *The Journal of the Acoustical Society of America* 56.4 (Oct. 1974), pp. 1195–1201 (cit. on p. 1).

[7] Kurt Blaukopf. *Musical Life in a Changing Society: Aspects of Music Sociology*. Portland, Or: Amadeus Press, 1992 (cit. on p. 1).

[8] Michael Barron. *Auditorium Acoustics and Architectural Design*. 2nd ed. London: Spon Press, Sept. 2009 (cit. on p. 1).

[9] Carl C. Crandell and Joseph J. Smaldino. « Classroom Acoustics for Children With Normal Hearing and With Hearing Impairment ». In: *Language, Speech, and Hearing Services in Schools* 31.4 (Oct. 2000), pp. 362–370 (cit. on p. 1).

[10] Maria Klatte et al. « Effects of Classroom Acoustics on Performance and Well-Being in Elementary School Children: A Field Study ». In: *Environment and Behavior* 42.5 (Sept. 2010), pp. 659–692 (cit. on p. 1).

[11] Yoichi Ando. *Concert Hall Acoustics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985 (cit. on p. 1).

[12] Leo Beranek. *Concert Halls and Opera Houses*. New York, NY: Springer, 2004 (cit. on p. 1).

[13] Brian F. G. Katz, Damian Murphy, and Angelo Farina. « Exploring Cultural Heritage through Acoustic Digital Reconstructions ». In: *Physics Today* 73.12 (Dec. 2020), pp. 32–37 (cit. on p. 1).

[14] Hannes Rosseel and Toon van Waterschoot. « A State-of-the-Art Review on Acoustic Preservation of Historical Worship Spaces through Auralization ». In: *Signal Processing* 234 (Sept. 2025), p. 109992 (cit. on p. 1).

[15] Vern Oliver Knudsen. *Architectural Acoustics*. J. Wiley & sons, Incorporated, 1932 (cit. on p. 1).

[16] R. H. Bolt and A. D. MacDonald. « Theory of Speech Masking by Reverberation ». In: *The Journal of the Acoustical Society of America* 21.6 (Nov. 1949), pp. 577–580 (cit. on p. 1).

[17] David B. Hawkins and William S. Yacullo. « Signal-to-Noise Ratio Advantage of Binaural Hearing Aids and Directional Microphones under Different Levels of Reverberation ». In: *Journal of Speech and Hearing Disorders* 49.3 (Aug. 1984), pp. 278–286 (cit. on p. 2).

[18] Yi Hu and Kostas Kokkinakis. « Effects of Early and Late Reflections on Intelligibility of Reverberated Speech by Cochlear Implant Listeners ». In: *The Journal of the Acoustical Society of America* 135.1 (Dec. 2013), EL22–EL28 (cit. on p. 2).

[19] Emanuël Habets et al. « Joint Dereverberation and Residual Echo Suppression of Speech Signals in Noisy Environments ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.8 (Nov. 2008), pp. 1433–1451 (cit. on p. 2).

[20] Marco Jeub. « Do We Need Dereverberation for Hand-Held Telephony? » In: (2010) (cit. on p. 2).

[21] Takuya Yoshioka et al. « Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition ». In: *IEEE Signal Processing Magazine* 29.6 (Nov. 2012), pp. 114–126 (cit. on p. 2).

[22] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, eds. *Audio Source Separation and Speech Enhancement*. Hoboken, NJ: John Wiley & Sons, 2018 (cit. on pp. 2, 16, 28).

[23] Jacob Benesty et al. *Advances in Network and Acoustic Echo Cancellation*. Digital Signal Processing. Berlin Heidelberg: Springer, 2001 (cit. on p. 2).

[24] Toon van Waterschoot and Marc Moonen. « Fifty Years of Acoustic Feedback Control: State of the Art and Future Challenges ». In: *Proceedings of the IEEE* 99.2 (Feb. 2011), pp. 288–327 (cit. on p. 2).

[25] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. « Acoustic Matching By Embedding Impulse Responses ». In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2020, pp. 426–430 (cit. on pp. 2, 42).

[26] Jacques Hadamard. *Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques*. Reproduction en fac-similé. Paris: J. Gabay, 1932 (cit. on p. 2).

[27] Marina Krémé. « Modification Locale et Consistance Globale Dans Le Plan Temps-Fréquence ». These de Doctorat. Aix-Marseille, July 2021 (cit. on pp. 2, 14).

[28] Patrick A. Naylor and Nikolay D. Gaubitch, eds. *Speech Dereverberation*. Signals and Communication Technology. London: Springer London, 2010 (cit. on pp. 2, 29, 83).

[29] D. Kundur and D. Hatzinakos. « Blind Image Deconvolution ». In: *IEEE Signal Processing Magazine* 13.3 (May 1996), pp. 43–64 (cit. on p. 2).

[30] J. Mourjopoulos. « On the Variation and Invertibility of Room Impulse Response Functions ». In: *Journal of Sound and Vibration* 102.2 (Sept. 1985), pp. 217–228 (cit. on pp. 3, 38).

[31] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. 3rd ed. Chicago, IL: University of Chicago Press, 1996 (cit. on p. 3).

[32] Léna Soler. *Introduction à l'épistémologie*. Philo. Paris: Ellipses, 2000 (cit. on p. 3).

[33] Tomohiro Nakatani et al. « Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (Sept. 2010), pp. 1717–1731 (cit. on pp. 3, 38, 76).

[34] Nikolay D. Gaubitch, Mark R. P. Thomas, and Patrick A. Naylor. « Dereverberation Using LPC-based Approaches ». In: *Speech Dereverberation*. Ed. by Patrick A. Naylor and Nikolay D. Gaubitch. London: Springer London, 2010, pp. 95–128 (cit. on pp. 3, 38).

[35] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016 (cit. on p. 3).

[36] Zhong-Qiu Wang et al. « TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 3221–3236 (cit. on pp. 4, 29, 42, 53).

[37] Bertrand Russell, Jean-Michel Roy, and Elizabeth Ramsden Eames. *Théorie de la connaissance: le manuscrit de 1913*. Textes philosophiques. Paris: J. Vrin, 2002 (cit. on p. 4).

[38] Danilo de Oliveira et al. « The PESQetarian: On the Relevance of Goodhart's Law for Speech Enhancement ». In: *Proc. Interspeech 2024*. 2024, pp. 3854–3858 (cit. on p. 4).

[39] Roman Frigg and Stephan Hartmann. *Models in Science*. Feb. 2006 (cit. on p. 5).

[40] Mary S. Morgan and Margaret Morrison, eds. *Models as Mediators: Perspectives on Natural and Social Science*. Ideas in Context. Cambridge: Cambridge University Press, 1999 (cit. on p. 5).

[41] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. « Multilayer Feedforward Networks Are Universal Approximators ». In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366 (cit. on p. 5).

[42] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown, 2016 (cit. on p. 5).

[43] Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press, 2021 (cit. on p. 5).

[44] Constance Douwes et al. « Is Quality Enoughf Integrating Energy Consumption in a Large-Scale Evaluation of Neural Audio Synthesis Models ». In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–5 (cit. on p. 5).

[45] Lynn H. Kaack et al. « Aligning Artificial Intelligence with Climate Change Mitigation ». In: *Nature Climate Change* 12.6 (June 2022), pp. 518–527 (cit. on p. 5).

[46] Cynthia Rudin. « Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead ». In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215 (cit. on p. 6).

[47] Nir Shlezinger, Yonina C. Eldar, and Stephen P. Boyd. « Model-Based Deep Learning: On the Intersection of Deep Learning and Optimization ». In: *IEEE Access* 10 (2022), pp. 115384–115398 (cit. on p. 6).

[48] Stanley H. Chan, Xiran Wang, and Omar A. Elgendy. « Plug-and-Play ADMM for Image Restoration: Fixed-Point Convergence and Applications ». In: *IEEE Transactions on Computational Imaging* 3.1 (Mar. 2017), pp. 84–98 (cit. on pp. 6, 23).

[49] Vishal Monga, Yuelong Li, and Yonina C. Eldar. « Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing ». In: *IEEE Signal Processing Magazine* 38.2 (Mar. 2021), pp. 18–44 (cit. on p. 6).

[50] Louis Bahrman, Mathieu Fontaine, Jonathan Le Roux, and Gaël Richard. « Speech Dereverberation Constrained on Room Impulse Response Characteristics ». In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 622–626 (cit. on pp. 7, 50, 87).

[51] Louis Bahrman, Mathieu Fontaine, and Gaël Richard. « A Hybrid Model for Weakly-Supervised Speech Dereverberation ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5 (cit. on pp. 8, 76).

[52] Louis Bahrman, Mathieu Fontaine, and Gaël Richard. « Déréverbération Non-Supervisée de La Parole Par Modèle Hybride ». In: *XXXe Colloque Francophone de Traitement Du Signal et Des Images*. Strasbourg, France: GRETSI, Aug. 2025 (cit. on pp. 8, 76).

[53] Louis Bahrman, Marius Rodrigues, Mathieu Fontaine, and Gaël Richard. « U-DREAM: Unsupervised Dereverberation Guided by a Reverberation Model ». In: *IEEE Transactions on Audio, Speech and Language Processing* 34 (2026), pp. 1552–1563 (cit. on pp. 8, 76).

[54] D. Griffin and Jae Lim. « Signal Estimation from Modified Short-Time Fourier Transform ». In: *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 8. Boston, MASS, USA: Institute of Electrical and Electronics Engineers, 1983, pp. 804–807 (cit. on pp. 14, 42).

[55] Félix Mathieu. « Traitement de La Phase Des Signaux Audio Dans Les Réseaux de Neurones Profonds ». These de Doctorat. Institut polytechnique de Paris, Nov. 2023 (cit. on pp. 14, 42).

[56] L. Cohen. « Time-Frequency Distributions-a Review ». In: *Proceedings of the IEEE* 77.7 (July 1989), pp. 941–981 (cit. on p. 14).

[57] Chester H. Page. « Instantaneous Power Spectra ». In: *Journal of Applied Physics* 23.1 (Jan. 1952), pp. 103–106 (cit. on pp. 14, 30).

[58] F. Hlawatsch and François Auger. *Temps-fréquence: concepts et outils*. Paris: Hermès Science Publications, 2005 (cit. on pp. 14, 30).

[59] Martin Vetterli, Jelena Kovačević, and Vivek K. Goyal. *Foundations of Signal Processing*. Cambridge: Cambridge university press, 2014 (cit. on p. 15).

[60] « Adaptive Filtering in Sub-Bands with Critical Sampling: Analysis, Experiments, and Application to Acoustic Echo Cancellation ». In: *IEEE Transactions on Signal Processing* 40.8 (Aug. 1992). Ed. by André Gilloire and Martin Vetterli, pp. 1862–1875 (cit. on p. 17).

[61] A. Gilloire and M. Vetterli. « Adaptive Filtering in Sub-Bands ». In: *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. Apr. 1988, 1572–1575 vol.3 (cit. on p. 17).

[62] Yekutiel Avargel and Israel Cohen. « System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.4 (May 2007), pp. 1305–1319 (cit. on pp. 17, 22, 52, 72, 131).

[63] Alan V. Oppenheim and Ronald W. Schafer. *Digital Signal Processing*. Englewood Cliffs, N.J: Prentice-Hall, 1975 (cit. on p. 19).

[64] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Version 29. Cambridge New York Melbourne New Delhi Singapore: Cambridge University Press, 2023 (cit. on pp. 20, 21, 59).

[65] William Karush. « Minima of Functions of Several Variables with Inequalities as Side Constraints ». In: *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago* (1939) (cit. on p. 20).

[66] H. W. Kuhn and A. W. Tucker. « Nonlinear Programming ». In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 2. University of California Press, Jan. 1951, pp. 481–493 (cit. on p. 20).

[67] M. Zuhair Nashed and University of Wisconsin–Madison, eds. *Generalized Inverses and Applications: Proceedings of an Advanced Seminar*. Publication of the Mathematics Research Center, University of Wisconsin–Madison ; No. 32. New York: Academic Press, 1976 (cit. on p. 21).

[68] Ulf Grenander and Gábor Szegö. *Toeplitz Forms and Their Applications*. 2., (textually unaltered) ed., repr. Providence, RI: AMS, 2001 (cit. on p. 21).

[69] Jérôme Idier. *Approche bayésienne pour les problèmes inverses*. Paris: Hermès Science publications : Lavoisier, 2001 (cit. on p. 21).

[70] Ali Baghaki, M. Omair Ahmad, and M. N. S. Swamy. « A New Two-Stage Method for Single-Microphone Speech Dereverberation ». In: *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. May 2016, pp. 778–781 (cit. on p. 21).

[71] Neal Parikh and Stephen Boyd. « Proximal Algorithms ». In: *Foundations and Trends in Optimization* 1.3 (Jan. 2014), pp. 127–239 (cit. on p. 22).

[72] Amir Beck and Marc Teboulle. « A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems ». In: *SIAM Journal on Imaging Sciences* 2.1 (Jan. 2009), pp. 183–202 (cit. on p. 22).

[73] Stephen Boyd. « Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers ». In: *Foundations and Trends® in Machine Learning* 3.1 (2010), pp. 1–122 (cit. on p. 22).

[74] Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. « Plug-and-Play Priors for Model Based Reconstruction ». In: *2013 IEEE Global Conference on Signal and Information Processing*. Dec. 2013, pp. 945–948 (cit. on p. 22).

[75] Per Christian Hansen. « Deconvolution and Regularization with Toeplitz Matrices ». In: *Numerical Algorithms* 29.4 (Apr. 2002), pp. 323–378 (cit. on p. 23).

[76] Ina Kodrasi, Timo Gerkmann, and Simon Doclo. « Frequency-Domain Single-Channel Inverse Filtering for Speech Dereverberation: Theory and Practice ». In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2014, pp. 5177–5181 (cit. on p. 23).

[77] Haonan Hu et al. « Speech Dereverberation with Deconvolution Regularized by Denoising ». In: *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Dec. 2024, pp. 1–6 (cit. on pp. 23, 38).

[78] K.S. Tang et al. « Genetic Algorithms and Their Applications ». In: *IEEE Signal Processing Magazine* 13.6 (Nov. 1996), pp. 22–37 (cit. on p. 23).

[79] Michael Chemistruck, Kyle Marcolini, and Will Pirkle. « Generating Matrix Coefficients for Feedback Delay Networks Using Genetic Algorithm ». In: *Audio Engineering Society Convention*. Vol. 133. Oct. 2012 (cit. on pp. 23, 35).

[80] Laurent Girin et al. « Dynamical Variational Autoencoders: A Comprehensive Review ». In: *Foundations and Trends® in Machine Learning* 15.1-2 (2021), pp. 1–175 (cit. on p. 24).

[81] Jean-Marie Lemercier et al. « Diffusion Models for Audio Restoration: A Review ». In: *IEEE Signal Processing Magazine* 41.6 (Nov. 2024), pp. 72–84 (cit. on pp. 24, 40).

[82] Vesa Valimaki et al. « Fifty Years of Artificial Reverberation ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.5 (July 2012), pp. 1421–1448 (cit. on pp. 27, 33).

[83] L. A. Ostrovsky. « Wave Processes in Media with Strong Acoustic Nonlinearity ». In: *The Journal of the Acoustical Society of America* 90.6 (Dec. 1991), pp. 3332–3337 (cit. on p. 28).

[84] Allan D. Pierce. *Acoustics: An Introduction to Its Physical Principles and Applications*. Cham: Springer International Publishing, 2019 (cit. on p. 28).

[85] Roland Badeau. *General Stochastic Reverberation Model*. Research Report. Télécom ParisTech, Feb. 2019 (cit. on p. 28).

[86] Vesa Välimäki et al. « More Than 50 Years of Artificial Reverberation ». In: *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*. Audio Engineering Society, Jan. 2016 (cit. on p. 28).

[87] Jean-Dominique Polack. « La Transmission de l'energie Sonore Dans Les Salles ». PhD thesis. Université du Maine, 1988 (cit. on pp. 29, 33, 80).

[88] W. B. Joyce. « Sabine's Reverberation Time and Ergodic Auditoriums ». In: *The Journal of the Acoustical Society of America* 58.3 (Sept. 1975), pp. 643–655 (cit. on pp. 29, 33).

[89] Rebecca Stewart and Mark Sandler. « Statistical Measures of Early Reflections of Room Impulse Responses ». In: (2007) (cit. on p. 29).

[90] James A. Moorer. « About This Reverberation Business ». In: *Computer Music Journal* 3.2 (1979), pp. 13–28 (cit. on pp. 29, 32, 33).

[91] Eric A. Lehmann and Anders M. Johansson. « Diffuse Reverberation Model for Efficient Image-Source Simulation of Room Impulse Responses ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6 (Aug. 2010), pp. 1429–1439 (cit. on pp. 29, 33).

[92] Maximilian Schäfer et al. « Distribution of Modal Damping in Absorptive Shoebox Rooms ». In: *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2023, pp. 1–5 (cit. on p. 29).

[93] Matti Karjalainen et al. « Estimation of Modal Decay Parameters from Noisy Response Measurements ». In: *Journal of the Audio Engineering Society* 50.11 (2002), pp. 867–878 (cit. on pp. 29, 31).

[94] Sungho Lee, Hyeong-Seok Choi, and Kyogu Lee. « Differentiable Artificial Reverberation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 2541–2556 (cit. on pp. 29, 35, 37).

[95] J. Eaton et al. « The ACE Challenge — Corpus Description and Performance Evaluation ». In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2015, pp. 1–5 (cit. on p. 29).

[96] Julius Richter et al. « EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation ». In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 4873–4877 (cit. on pp. 29, 87).

[97] James Eaton et al. « Estimation of Room Acoustic Parameters: The ACE Challenge ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.10 (Oct. 2016), pp. 1681–1693 (cit. on p. 29).

[98] M. R. Schroeder. « Complementarity of Sound Buildup and Decay ». In: *The Journal of the Acoustical Society of America* 40.3 (Sept. 1966), pp. 549–551 (cit. on p. 30).

[99] Dennis R. Morgan. « A Parametric Error Analysis of the Backward Integration Method for Reverberation Time Estimation ». In: *The Journal of the Acoustical Society of America* 101.5 (May 1997), pp. 2686–2693 (cit. on p. 30).

[100] Laurent Faiget, Claude Legros, and Robert Ruiz. « Optimization of the Impulse Response Length: Application to Noisy and Highly Reverberant Rooms ». In: *Journal of the Audio Engineering Society* 46.9 (1998), pp. 741–750 (cit. on pp. 30, 31).

[101] J.-M. Jot. « An Analysis/Synthesis Approach to Real-Time Artificial Reverberation ». In: *Acoustics, Speech, and Signal Processing, IEEE International Conference On*. IEEE Computer Society, Mar. 1992, pp. 221–224 (cit. on pp. 30, 35, 69).

[102] Gloria Dal Santo et al. « Similarity Metrics for Late Reverberation ». In: *2024 58th Asilomar Conference on Signals, Systems, and Computers*. Oct. 2024, pp. 1409–1413 (cit. on p. 30).

[103] Wallace Clement Sabine. *Collected Papers on Acoustics*. New York: Dover Press, 1922 (cit. on p. 31).

[104] Carl F. Eyring. « Reverberation Time in "Dead" Rooms ». In: *The Journal of the Acoustical Society of America* 1.2A_Supplement (Jan. 1930), pp. 168–168 (cit. on p. 31).

[105] Roland Badeau. « Statistical Wave Field Theory ». In: *The Journal of the Acoustical Society of America* 156.1 (July 2024), pp. 573–599 (cit. on pp. 31, 33, 80, 81).

[106] Ning Xiang. « Evaluation of Reverberation Times Using a Nonlinear Regression Approach ». In: *The Journal of the Acoustical Society of America* 98.4 (Oct. 1995), pp. 2112–2121 (cit. on p. 31).

[107] Sylvio R. Bistafa and John S. Bradley. « Reverberation Time and Maximum Background-Noise Level for Classrooms from a Comparative Study of Speech Intelligibility Metrics ». In: *The Journal of the Acoustical Society of America* 107.2 (Feb. 2000), pp. 861–875 (cit. on pp. 31, 40).

[108] Lauri Savioja and U. Peter Svensson. « Overview of Geometrical Room Acoustic Modeling Techniques ». In: *The Journal of the Acoustical Society of America* 138.2 (Aug. 2015), pp. 708–730 (cit. on p. 31).

[109] Lauri Savioja, Timo J. Rinne, and Tapio Takala. « Simulation of Room Acoustics with a 3-D Finite Difference Mesh ». In: *International Computer Music Conference Proceedings* 1994 (1994) (cit. on p. 31).

[110] D. Botteldooren. « Finite-difference Time-domain Simulation of Low-frequency Room Acoustic Problems ». In: *The Journal of the Acoustical Society of America* 98.6 (Dec. 1995), pp. 3302–3308 (cit. on p. 31).

[111] Lonny L. Thompson. « A Review of Finite-Element Methods for Time-Harmonic Acoustics ». In: *The Journal of the Acoustical Society of America* 119.3 (Mar. 2006), pp. 1315–1330 (cit. on p. 31).

[112] Stephen Kirkup. « The Boundary Element Method in Acoustics: A Survey ». In: *Applied Sciences* 9.8 (Jan. 2019), p. 1642 (cit. on p. 31).

[113] Jont B. Allen and David A. Berkley. « Image Method for Efficiently Simulating Small-room Acoustics ». In: *The Journal of the Acoustical Society of America* 65.4 (Apr. 1979), pp. 943–950 (cit. on p. 31).

[114] Jeffrey Borish. « Extension of the Image Model to Arbitrary Polyhedra ». In: *The Journal of the Acoustical Society of America* 75.6 (June 1984), pp. 1827–1836 (cit. on p. 31).

[115] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. « Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms ». In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2018, pp. 351–355 (cit. on pp. 32, 53, 81).

[116] Yi Luo and Jianwei Yu. « FRA-RIR: Fast Random Approximation of the Image-source Method ». In: *Proc. Interspeech 2023*. 2023, pp. 3884–3888 (cit. on p. 32).

[117] Bowen Zhi et al. « A Differentiable Image Source Model for Room Acoustics Optimization ». In: *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2023, pp. 1–5 (cit. on pp. 32, 35).

[118] Liam Kelley et al. « RIR-in-a-Box: Estimating Room Acoustics from 3D Mesh Data through Shoebox Approximation ». In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 3255–3259 (cit. on p. 32).

[119] Achille Aknin, Théophile Dupré, and Roland Badeau. « Evaluation of a Stochastic Reverberation Model Based on the Image Source Principle ». In: *International Conference on Digital Audio Effects*. Sept. 2020 (cit. on pp. 32, 33).

[120] A. Krokstad, S. Strom, and S. Sørsdal. « Calculating the Acoustical Room Response by the Use of a Ray Tracing Technique ». In: *Journal of Sound and Vibration* 8.1 (July 1968), pp. 118–125 (cit. on p. 32).

[121] John Kenneth Haviland and Balakrishna D. Thanedar. « Monte Carlo Applications to Acoustical Field Solutions ». In: *The Journal of the Acoustical Society of America* 54.6 (Dec. 1973), pp. 1442–1448 (cit. on p. 32).

[122] Thomas Funkhouser et al. « A Beam Tracing Approach to Acoustic Modeling for Interactive Virtual Environments ». In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. New York, NY, USA: Association for Computing Machinery, July 1998, pp. 21–32 (cit. on p. 32).

[123] T. Lewers. « A Combined Beam Tracing and Radiatn Exchange Computer Model of Room Acoustics ». In: *Applied Acoustics* 38.2 (Jan. 1993), pp. 161–178 (cit. on p. 32).

[124] Bengt-Inge L. Dalenbäck. « Room Acoustic Prediction Based on a Unified Treatment of Diffuse and Specular Reflection ». In: *The Journal of the Acoustical Society of America* 100.2 (Aug. 1996), pp. 899–909 (cit. on p. 32).

[125] Alex Southern et al. « Room Impulse Response Synthesis and Validation Using a Hybrid Acoustic Model ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.9 (Sept. 2013), pp. 1940–1952 (cit. on p. 32).

[126] Wouter Wittebol et al. « A Hybrid Room Acoustic Modeling Approach Combining Image Source, Acoustic Diffusion Equation, and Time-Domain Discontinuous Galerkin Methods ». In: *Applied Acoustics* 223 (July 2024), p. 110068 (cit. on p. 32).

[127] Michael Vorländer. « Simulation of the Transient and Steady-state Sound Propagation in Rooms Using a New Combined Ray-tracing/Image-source Algorithm ». In: *The Journal of the Acoustical Society of America* 86.1 (July 1989), pp. 172–178 (cit. on p. 32).

[128] Jean-Dominique Polack. « Playing Billiards in the Concert Hall: The Mathematical Foundations of Geometrical Room Acoustics ». In: *Applied Acoustics* 38.2 (Jan. 1993), pp. 235–244 (cit. on p. 33).

[129] D. Middleton. « A Statistical Theory of Reverberation and Similar First-Order Scattered Fields–I: Waveforms and the General Process ». In: *IEEE Transactions on Information Theory* 13.3 (July 1967), pp. 372–392 (cit. on p. 33).

[130] Roland Badeau. « Statistical Wave Field Theory: Special Polyhedra ». In: *The Journal of the Acoustical Society of America* 157.3 (Mar. 2025), pp. 2263–2278 (cit. on p. 33).

[131] Achille Aknin and Roland Badeau. « Stochastic Reverberation Model with a Frequency Dependent Attenuation ». In: *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2021, pp. 351–355 (cit. on p. 33).

[132] Per Rubak and Lars G. Johansen. « Artificial Reverberation Based on a Pseudo-Random Impulse Response ». In: *journal of the audio engineering society* 4725 (May 1998) (cit. on p. 33).

[133] Per Rubak and Lars G. Johansen. « Artificial Reverberation Based on a Pseudo-Random Impulse Response Ii ». In: *journal of the audio engineering society* 4900 (May 1999) (cit. on p. 33).

[134] Matti Karjalainen and Hanna Järveläinen. « Reverberation Modeling Using Velvet Noise ». In: *Journal of the Audio Engineering Society* 7 (Mar. 2007) (cit. on p. 33).

[135] M.R. Schroeder and B.F. Logan. « "Colorless" Artificial Reverberation ». In: *IRE Transactions on Audio* AU-9.6 (Nov. 1961), pp. 209–214 (cit. on p. 34).

[136] Sebastian J. Schlecht and Emanuël A. P. Habets. « Feedback Delay Networks: Echo Density and Mixing Time ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.2 (Feb. 2017), pp. 374–383 (cit. on p. 34).

[137] Damian Murphy et al. « Acoustic Modeling Using the Digital Waveguide Mesh ». In: *IEEE Signal Processing Magazine* 24.2 (Mar. 2007), pp. 55–66 (cit. on p. 34).

[138] Hequn Bai, Gaël Richard, and Laurent Daudet. « Late Reverberation Synthesis: From Radiance Transfer to Feedback Delay Networks ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.12 (Dec. 2015), pp. 2260–2271 (cit. on p. 34).

[139] Anton Ratnarajah et al. « Fast-Rir: Fast Neural Diffuse Room Impulse Response Generator ». In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2022, pp. 571–575 (cit. on pp. 34, 37).

[140] Silvia Arellano et al. *Room Impulse Response Generation Conditioned on Acoustic Parameters*. July 2025 (cit. on p. 34).

[141] Sheng Lyu, Yuemin Yu, and Chenshu Wu. « Temporal Modeling of Room Impulse Response Generation via Multi-Scale Autoregressive Learning ». In: *Proc. Interspeech 2025*. 2025, pp. 923–927 (cit. on p. 34).

[142] Kun Su, Mingfei Chen, and Eli Shlizerman. « INRAS: Implicit Neural Representation for Audio Scenes ». In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 8144–8158 (cit. on p. 34).

[143] I. Martin et al. « Predicting Room Impulse Responses Through Encoder-Decoder Convolutional Neural Networks ». In: *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. Sept. 2023, pp. 1–6 (cit. on p. 34).

[144] Tom Sprunck. « Peut-on Entendre La Forme d'une Pièce ? : Reconstruction de La Géométrie d'une Salle à Partir de Mesures Acoustiques Par Super-Résolution et Optimisation de Forme ». These de Doctorat. Strasbourg, Dec. 2024 (cit. on p. 35).

[145] Alastair H. Moore, Mike Brookes, and Patrick A. Naylor. « Room Geometry Estimation from a Single Channel Acoustic Impulse Response ». In: *21st European Signal Processing Conference (EUSIPCO 2013)*. Sept. 2013, pp. 1–5 (cit. on p. 35).

[146] Tom Sprunck et al. « Fully Reversing the Shoebox Image Source Method: From Impulse Responses to Room Parameters ». In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 1023–1033 (cit. on p. 35).

[147] Wangyang Yu and W. Bastiaan Kleijn. « Room Acoustical Parameter Estimation From Room Impulse Responses Using Deep Neural Networks ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 436–447 (cit. on p. 35).

[148] Martin Kuster. « Reliability of Estimating the Room Volume from a Single Room Impulse Response ». In: *The Journal of the Acoustical Society of America* 124.2 (Aug. 2008), pp. 982–993 (cit. on p. 35).

[149] Cédric Foy, Antoine Deleforge, and Diego Di Carlo. « Mean Absorption Estimation from Room Impulse Responses Using Virtually Supervised Learninga) ». In: *The Journal of the Acoustical Society of America* 150.2 (Aug. 2021), pp. 1286–1299 (cit. on p. 35).

[150] Dejan Markovic´ et al. « Estimation of Room Dimensions from a Single Impulse Response ». In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Oct. 2013, pp. 1–4 (cit. on p. 35).

[151] Bo Holm-Rasmussena, Heidi-Maria Lehtonen, and Vesa Välimäki. « A New Reverberator Based on Variable Sparsity Convolution ». In: *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*. Vol. 5. 2013, pp. 7–8 (cit. on p. 35).

[152] Vesa Välimäki et al. « Late Reverberation Synthesis Using Filtered Velvet Noise ». In: *Applied Sciences* 7.5 (May 2017), p. 483 (cit. on p. 35).

[153] Jay Coggin and Will Pirkle. « Automatic Design of Feedback Delay Network Reverb Parameters for Impulse Response Matching ». In: *Journal of the Audio Engineering Society* 9666 (Sept. 2016) (cit. on p. 35).

[154] Ilias Ibnyahya and Joshua D Reiss. « A Method for Matching Room Impulse Responses with Feedback Delay Networks ». In: (2022) (cit. on p. 35).

[155] Jesse Engel et al. *DDSP: Differentiable Digital Signal Processing*. Jan. 2020 (cit. on p. 35).

[156] Gloria Dal Santo et al. « Differentiable Feedback Delay Network For Colorless Reverberation ». In: *Proceedings of the 26th International Conference on Digital Audio Effects (DAFx23)*. Aalborg University, Sept. 2023, pp. 244–251 (cit. on p. 35).

[157] Orchisama Das et al. *Differentiable Grouped Feedback Delay Networks for Learning Coupled Volume Acoustics*. Aug. 2025 (cit. on p. 35).

[158] Christian J. Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. « Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech ». In: *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2021, pp. 221–225 (cit. on pp. 35, 37).

[159] Alessandro Ilic Mezza et al. « Differentiable Scattering Delay Networks for Artificial Reverberation ». In: *Proceedings of the 28-Th Int. Conf. on Digital Audio Effects (Dafx25)* (Sept. 2–5, 2025). Ed. by L. Gabrielli and S. Cecchi. Ancona, Italy, Sept. 2025 (cit. on p. 35).

[160] *ISO 3382-1:2009* (cit. on p. 35).

[161] Angelo Farina. « Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique ». In: *Audio Engineering Society Convention 108*. Audio Engineering Society, Feb. 2000 (cit. on p. 35).

[162] Jeffrey Borish and James B. Angell. « An Efficient Algorithm for Measuring the Impulse Response Using Pseudorandom Noise ». In: *Journal of the Audio Engineering Society* 31.7/8 (1983), pp. 478–488 (cit. on p. 36).

[163] Martin Jälmby, Filip Elvander, and Toon van Waterschoot. « Fast Low-Latency Convolution by Low-Rank Tensor Approximation ». In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–5 (cit. on p. 36).

[164] Yuanqing Lin and D.D. Lee. « Bayesian Regularization and Nonnegative Deconvolution for Room Impulse Response Estimation ». In: *IEEE Transactions on Signal Processing* 54.3 (Mar. 2006), pp. 839–847 (cit. on p. 36).

[165] Alexis Benichoux et al. « Convex Regularizations for the Simultaneous Recording of Room Impulse Responses ». In: *IEEE Transactions on Signal Processing* 62.8 (Apr. 2014), pp. 1976–1986 (cit. on pp. 36, 59).

[166] Toon van Waterschoot, Geert Rombouts, and Marc Moonen. « Optimally Regularized Adaptive Filtering Algorithms for Room Acoustic Signal Enhancement ». In: *Signal Processing* 88.3 (Mar. 2008), pp. 594–611 (cit. on p. 36).

[167] Louis Lalay, Mathieu Fontaine, and Roland Badeau. « Unified Variational and Physics-aware Model for Room Impulse Response Estimation ». In: *Proc. Interspeech 2025*. 2025, pp. 3818–3822 (cit. on p. 36).

[168] Côme Peladeau and Geoffroy Peeters. « Blind Estimation of Audio Effects Using an Auto-Encoder Approach and Differentiable Digital Signal Processing ». In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2024, pp. 856–860 (cit. on p. 36).

[169] David Sundström, Filip Elvander, and Andreas Jakobsson. « Estimation of Impulse Responses for a Moving Source Using Optimal Transport Regularization ». In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2024, pp. 921–925 (cit. on p. 36).

[170] Gongping Huang, Jacob Benesty, and Jingdong Chen. « Dimensionality Reduction of Room Acoustic Impulse Responses and Applications to System Identification ». In: *IEEE Signal Processing Letters* 30 (2023), pp. 1107–1111 (cit. on p. 36).

[171] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. « Deep Impulse Responses: Estimating and Parameterizing Filters with Deep Networks ». In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2022, pp. 3209–3213 (cit. on p. 36).

[172] Jean-Marc Jot, Laurent Cerveau, and Olivier Warusfel. « Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model ». In: *AES: Convention of the Audio Engineering Society*. 1997 (cit. on p. 36).

[173] Seongrae Kim, Jae-hyoun Yoo, and Jung-Woo Choi. « Echo-Aware Room Impulse Response Generation ». In: *The Journal of the Acoustical Society of America* 156.1 (July 2024), pp. 623–637 (cit. on p. 36).

[174] Jackie Lin, Georg Götz, and Sebastian J. Schlecht. *Deep Room Impulse Response Completion*. Feb. 2024 (cit. on p. 36).

[175] Rama Ratnam et al. « Blind Estimation of Reverberation Time ». In: *The Journal of the Acoustical Society of America* 114.5 (Oct. 2003), pp. 2877–2892 (cit. on p. 36).

[176] Heinrich W Lollmann and Peter Vary. « Estimation of the Reverberation Time in Noisy Environments ». In: *Proceedings of the International Workshop on Acoustic Echo and Noise Control*. Seattle, USA, Sept. 2008 (cit. on p. 36).

[177] Thiago Prego et al. « Blind Estimators for Reverberation Time and Direct-to-Reverberant Energy Ratio Using Subband Speech Decomposition ». In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2015, pp. 1–5 (cit. on pp. 36, 84).

[178] Paul Kendrick et al. « Monaural Room Acoustic Parameters from Music and Speech ». In: *The Journal of the Acoustical Society of America* 124.1 (July 2008), pp. 278–287 (cit. on p. 36).

[179] Feifei Xiong et al. « Exploring Auditory-Inspired Acoustic Features for Room Acoustic Parameter Estimation From Monaural Speech ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (Oct. 2018), pp. 1809–1820 (cit. on p. 36).

[180] Hannes Gamper and Ivan J. Tashev. « Blind Reverberation Time Estimation Using a Convolutional Neural Network ». In: *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Sept. 2018, pp. 136–140 (cit. on p. 36).

[181] Chunxi Wang et al. « Exploring the Power of Pure Attention Mechanisms in Blind Room Parameter Estimation ». In: *EURASIP Journal on Audio, Speech, and Music Processing* 2024.1 (Apr. 2024), p. 23 (cit. on p. 36).

[182] Wolfgang Mack, Shuwen Deng, and Emanuël A.P. Habets. « Single-Channel Blind Direct-to-Reverberation Ratio Estimation Using Masking ». In: *Interspeech 2020*. ISCA, Oct. 2020, pp. 5066–5070 (cit. on pp. 36, 85).

[183] Pablo Peso Parada et al. « A Single-Channel Non-Intrusive C50 Estimator Correlated With Speech Recognition Performance ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.4 (Apr. 2016), pp. 719–732 (cit. on p. 37).

[184] Samir Sadok et al. « AnCoGen: Analysis, Control and Generation of Speech with a Masked Autoencoder ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5 (cit. on pp. 37, 41).

[185] Michael Neri et al. « Single-Channel Speaker Distance Estimation in Reverberant Environments ». In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2023) (cit. on p. 37).

[186] Andrea F. Genovese et al. « Blind Room Volume Estimation from Single-channel Noisy Speech ». In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2019, pp. 231–235 (cit. on p. 37).

[187] Nicholas J. Bryan. « Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation ». In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2020, pp. 1–5 (cit. on p. 37).

[188] Prerak Srivastava, Antoine Deleforge, and Emmanuel Vincent. « Realistic Sources, Receivers and Walls Improve The Generalisability of Virtually-Supervised Blind Acoustic Parameter Estimators ». In: *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. Sept. 2022, pp. 1–5 (cit. on p. 37).

[189] Lijun Wang, Suradej Duangpummet, and Masashi Unoki. « Blind Estimation of Speech Transmission Index and Room Acoustic Parameters by Using Extended Model of Room Impulse Response Derived From Speech Signals ». In: *IEEE Access* 11 (2023), pp. 49431–49444 (cit. on p. 37).

[190] Nils Peters, Jaeyoung Choi, and Howard Lei. « Matching Artificial Reverb Settings to Unknown Room Recordings: A Recommendation System for Reverb Plugins ». In: *Journal of the Audio Engineering Society*. San Francisco, USA, Oct. 2012 (cit. on p. 37).

[191] Sungho Lee, Hyeong-Seok Choi, and Kyogu Lee. « Yet Another Generative Model for Room Impulse Response Estimation ». In: *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2023, pp. 1–5 (cit. on p. 37).

[192] Zhiheng Liao et al. « Blind Estimation of Room Impulse Response from Monaural Reverberant Speech with Segmental Generative Neural Network ». In: *Interspeech 2023*. ISCA, Aug. 2023, pp. 2723–2727 (cit. on p. 37).

[193] Francesc Lluís and Nils Meyer-Kahlen. « Blind Spatial Impulse Response Generation from Separate Room- and Scene-Specific Information ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5 (cit. on p. 37).

[194] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. « Comparison of Different Impulse Response Measurement Techniques ». In: *Journal of the Audio Engineering Society* 50.4 (Apr. 2002) (cit. on p. 38).

[195] Stephen T. Neely and J. B. Allen. « Invertibility of a Room Impulse Response ». In: *The Journal of the Acoustical Society of America* 66.1 (July 1979), pp. 165–169 (cit. on p. 38).

[196] J. Mourjopoulos, P. Clarkson, and J. Hammond. « A Comparative Study of Least-Squares and Homomorphic Techniques for the Inversion of Mixed Phase Signals ». In: *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 7. May 1982, pp. 1858–1861 (cit. on p. 38).

[197] Andrea De Giusti and Matteo Romanin. « Speech Dereverberation Using U-Net and Minimum-Phase Inverse Prefiltering ». In: *2024 32nd Telecommunications Forum (TELFOR)*. Nov. 2024, pp. 1–4 (cit. on p. 38).

[198] Jean-Marie Lemercier, Simon Welker, and Timo Gerkmann. « Diffusion Posterior Sampling for Informed Single-Channel Dereverberation ». In: *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2023, pp. 1–5 (cit. on p. 38).

[199] Keisuke Kinoshita et al. « Neural Network-Based Spectrum Estimation for Online WPE Dereverberation ». In: *Interspeech 2017*. ISCA, Aug. 2017, pp. 384–388 (cit. on p. 39).

[200] Jahn Heymann et al. « Frame-Online DNN-WPE Dereverberation ». In: *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Sept. 2018, pp. 466–470 (cit. on p. 39).

[201] Hao Li, Xueliang Zhang, and Guanglai Gao. « Robust Speech Dereverberation Based on WPE and Deep Learning ». In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Dec. 2020, pp. 52–56 (cit. on p. 39).

[202] Petko N. Petkov et al. « An Unsupervised Learning Approach to Neural-net-supported Wpe Dereverberation ». In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2019, pp. 5761–5765 (cit. on p. 39).

[203] Meihuang Wang et al. « A Deep Proximal-Unfolding Method for Monaural Speech Dereverberation ». In: *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Nov. 2022, pp. 324–329 (cit. on p. 39).

[204] John R. Hershey, Jonathan Le Roux, and Felix Weninger. *Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures*. Nov. 2014 (cit. on p. 39).

[205] Ziye Yang et al. « Integrating Data Priors to Weighted Prediction Error for Speech Dereverberation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024), pp. 1–16 (cit. on p. 39).

[206] Koichi Saito et al. « Unsupervised Vocal Dereverberation with Diffusion-Based Generative Models ». In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–5 (cit. on p. 39).

[207] Joon-Young Yang and Joon-Hyuk Chang. « VACE-WPE: Virtual Acoustic Channel Expansion Based on Neural Networks for Weighted Prediction Error-Based Speech Dereverberation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 174–189 (cit. on p. 39).

[208] Rintaro Ikeshita, Naoyuki Kamo, and Tomohiro Nakatani. « Blind Signal Dereverberation Based on Mixture of Weighted Prediction Error Models ». In: *IEEE Signal Processing Letters* 28 (2021), pp. 399–403 (cit. on p. 39).

[209] Mahdi Parchami, Wei-Ping Zhu, and Benoit Champagne. « Speech Dereverberation Using Weighted Prediction Error with Correlated Inter-Frame Speech Components ». In: *Speech Communication* 87 (Mar. 2017), pp. 49–57 (cit. on p. 39).

[210] Deepak Baby and Hugo Van hamme. « Supervised Speech Dereverberation in Noisy Environments Using Exemplar-Based Sparse Representations ». In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2016, pp. 156–160 (cit. on p. 40).

[211] Zhong-Qiu Wang, Gordon Wichern, and Jonathan Le Roux. « Convolutive Prediction for Monaural Speech Dereverberation and Noisy-Reverberant Speaker Separation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3476–3490 (cit. on p. 40).

[212] Zhong-Qiu Wang. « USDnet: Unsupervised Speech Dereverberation via Neural Forward Filtering ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 3882–3895 (cit. on pp. 40, 44).

[213] Xiaoyu Bie et al. *A Benchmark of Dynamical Variational Autoencoders Applied to Speech Spectrogram Modeling*. June 2021 (cit. on p. 40).

[214] Deepak Baby and Hervé Bourlard. « Speech Dereverberation Using Variational Autoencoders ». In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2021, pp. 5784–5788 (cit. on p. 40).

[215] Pengyu Wang and Xiaofei Li. « RVAE-EM: Generative Speech Dereverberation Based On Recurrent Variational Auto-Encoder And Convolutive Transfer Function ». In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2024, pp. 496–500 (cit. on p. 40).

[216] K. Lebart, J. M. Boucher, and P. N. Denbigh. « A New Method Based on Spectral Subtraction for Speech Dereverberation ». In: *Acta Acustica united with Acustica* 87.3 (May 2001), pp. 359–366 (cit. on p. 40).

[217] Emanuël A. P. Habets, Sharon Gannot, and Israel Cohen. « Late Reverberant Spectral Variance Estimation Based on a Statistical Model ». In: *IEEE Signal Processing Letters* 16.9 (Sept. 2009), pp. 770–773 (cit. on p. 40).

[218] Simon Leglaive, Roland Badeau, and Gaël Richard. « Student's t Source and Mixing Models for Multichannel Audio Source Separation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.6 (June 2018), pp. 1154–1168 (cit. on p. 40).

[219] Jean-Marie Lemercier et al. « Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation With Diffusion Models ». In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 2244–2258 (cit. on pp. 40, 61).

[220] Eloi Moliner et al. *BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models*. May 2024 (cit. on p. 40).

[221] Bo Wu et al. « A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (Jan. 2017), pp. 102–111 (cit. on p. 41).

[222] Helin Wang et al. « TeCANet: Temporal-Contextual Attention Network for Environment-Aware Speech Dereverberation ». In: *Interspeech 2021*. ISCA: ISCA, Aug. 2021, pp. 1109–1113 (cit. on p. 41).

[223] Yuying Li, Yuchen Liu, and Donald S. Williamson. « A Composite T60 Regression and Classification Approach for Speech Dereverberation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023), pp. 1–11 (cit. on p. 41).

[224] Nagashree K. S. Rao et al. « Low-Complexity Neural Speech Dereverberation With Adaptive Target Control ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5 (cit. on p. 41).

[225] Yeonjong Choi, Chao Xie, and Tomoki Toda. « Reverberation-Controllable Voice Conversion Using Reverberation Time Estimator ». In: *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 2103–2107 (cit. on p. 41).

[226] Julius Richter et al. « Speech Enhancement and Dereverberation With Diffusion-Based Generative Models ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2351–2364 (cit. on p. 41).

[227] Naoki Murata et al. *GibbsDDRM: A Partially Collapsed Gibbs Sampler for Solving Blind Inverse Problems with Denoising Diffusion Restoration*. June 2023 (cit. on p. 41).

[228] Jean-Marie Lemercier et al. « StoRM: A Diffusion-Based Stochastic Regeneration Model for Speech Enhancement and Dereverberation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2724–2737 (cit. on p. 41).

[229] Yi Luo and Nima Mesgarani. « Real-Time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network ». In: *Interspeech 2018*. ISCA, Sept. 2018, pp. 342–346 (cit. on p. 41).

[230] D. S. Wang, Y. X. Zou, and W. Shi. « A Deep Convolutional Encoder-Decoder Model for Robust Speech Dereverberation ». In: *2017 22nd International Conference on Digital Signal Processing (DSP)*. Aug. 2017, pp. 1–5 (cit. on pp. 41, 42).

[231] William Ravenscroft, Stefan Goetze, and Thomas Hain. « Receptive Field Analysis of Temporal Convolutional Networks for Monaural Speech Dereverberation ». In: *2022 30th European Signal Processing Conference (EUSIPCO)*. Aug. 2022, pp. 80–84 (cit. on pp. 41, 65).

[232] Vinay Kothapally and John H. L. Hansen. « Monaural Speech Dereverberation Using Deformable Convolutional Networks ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024), pp. 1–12 (cit. on p. 41).

[233] Lei Zhao et al. « Multi-Resolution Convolutional Residual Neural Networks for Monaural Speech Dereverberation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024), pp. 1–14 (cit. on p. 41).

[234] Felix Weninger et al. « Speech Enhancement with LSTM Recurrent Neural Networks and Its Application to Noise-Robust ASR ». In: *Latent Variable Analysis and Signal Separation*. Ed. by Emmanuel Vincent et al. Cham: Springer International Publishing, 2015, pp. 91–99 (cit. on pp. 41, 87).

[235] Xiang Hao et al. « Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement ». In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2021, pp. 6633–6637 (cit. on pp. 41, 42, 53, 65, 70, 87).

[236] Kohei Saijo et al. « TF-Locoformer: Transformer with Local Modeling by Convolution for Speech Separation and Enhancement ». In: *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Aalborg, Denmark: IEEE, Sept. 2024, pp. 205–209 (cit. on pp. 42, 88).

[237] Junghyun Koo, Seungryeol Paik, and Kyogu Lee. « Reverb Conversion Of Mixed Vocal Tracks Using An End-To-End Convolutional Deep Neural Network ». In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2021, pp. 81–85 (cit. on p. 42).

[238] Jaekwon Im and Juhan Nam. « DiffRENT: A Diffusion Model for Recording Environment Transfer of Speech ». In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2024, pp. 7425–7429 (cit. on p. 42).

[239] Wataru Nakata et al. *ReverbMiipher: Generative Speech Restoration Meets Reverberation Characteristics Controllability*. May 2025 (cit. on p. 42).

[240] Hakan Erdogan et al. « Phase-Sensitive and Recognition-Boosted Speech Separation Using Deep Recurrent Neural Networks ». In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2015, pp. 708–712 (cit. on p. 42).

[241] Donald S. Williamson, Yuxuan Wang, and DeLiang Wang. « Complex Ratio Masking for Monaural Speech Separation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.3 (Mar. 2016), pp. 483–492 (cit. on pp. 42, 70).

[242] Andong Li et al. « Two Heads Are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1829–1843 (cit. on p. 42).

[243] Ayal Schwartz, Sharon Gannot, and Shlomo E. Chazan. *Magnitude or Phase? A Two Stage Algorithm for Dereverberation*. Oct. 2022 (cit. on pp. 42, 65, 66).

[244] Emmanuel Vincent, Rémi Gribonval, and Cédric Fevotte. « Performance Measurement in Blind Audio Source Separation ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (July 2006), pp. 1462–1469 (cit. on p. 42).

[245] Jonathan Le Roux et al. « SDR – Half-baked or Well Done? » In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2019, pp. 626–630 (cit. on pp. 42, 88).

[246] Cees H. Taal et al. « An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (Sept. 2011), pp. 2125–2136 (cit. on p. 43).

[247] Jesper Jensen and Cees H. Taal. « An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.11 (Nov. 2016), pp. 2009–2022 (cit. on pp. 43, 88).

[248] A.W. Rix et al. « Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs ». In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Vol. 2. May 2001, 749–752 vol.2 (cit. on p. 43).

[249] Thomas Muller et al. « Speech Quality Evaluation of Neural Audio Codecs ». In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 1760–1764 (cit. on p. 43).

[250] ITUT Rec. « P. 862.2: Wideband Extension to Recommendation p. 862 for the Assessment of Wideband Telephone Networks and Speech Codecs ». In: *International Telecommunication Union, CH–Geneva* 41 (2005), pp. 48–60 (cit. on pp. 43, 88).

[251] Tiago H. Falk, Chenxi Zheng, and Wai-Yip Chan. « A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech ». In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (Sept. 2010), pp. 1766–1774 (cit. on pp. 44, 88).

[252] Szu-Wei Fu et al. « MetricGAN-U: Unsupervised Speech Enhancement/ Dereverberation Based Only on Noisy/ Reverberated Speech ». In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2022, pp. 7412–7416 (cit. on p. 44).

[253] Peter J. Huber. « Robust Estimation of a Location Parameter ». In: *The Annals of Mathematical Statistics* 35.1 (Mar. 1964), pp. 73–101 (cit. on p. 51).

[254] Garofolo, John et al. *CSR-I (WSJ0) Complete*. May 2007 (cit. on pp. 53, 70).

[255] Rui Zhou, Wenye Zhu, and Xiaofei Li. « Speech Dereverberation with a Reverberation Time Shortening Target ». In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–5 (cit. on pp. 53, 71, 87).

[256] Y. Bello-Cruz et al. « A Proximal Gradient Method with an Explicit Line Search for Multiobjective Optimization ». In: *Computational Optimization and Applications* (July 2025) (cit. on p. 60).

[257] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Jan. 2017 (cit. on p. 60).

[258] Jason Ansel et al. « PyTorch 2: Faster Machine Learning through Dynamic Python Bytecode Transformation and Graph Compilation ». In: *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, Apr. 2024 (cit. on pp. 61, 132).

[259] Ori Ernst et al. « Speech Dereverberation Using Fully Convolutional Networks ». In: *2018 26th European Signal Processing Conference (EUSIPCO)*. Sept. 2018, pp. 390–394 (cit. on p. 65).

[260] Prachi Sharma and Christian Kehling. « How Machines Perceive Rooms - Regions of Relevance in Room Impulse Responses ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5 (cit. on p. 65).

[261] Zhao Chen et al. « GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks ». In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, July 2018, pp. 794–803 (cit. on pp. 69, 86).

[262] Vassil Panayotov et al. « Librispeech: An ASR Corpus Based on Public Domain Audio Books ». In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2015, pp. 5206–5210 (cit. on p. 70).

[263] Rachel Bittner et al. « MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research ». In: *Proc. ISMIR*. 2014 (cit. on p. 73).

[264] Kevin Kilgour et al. *Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms*. Jan. 2019 (cit. on p. 73).

[265] Shawn Hershey et al. « CNN Architectures for Large-Scale Audio Classification ». In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2017, pp. 131–135 (cit. on p. 73).

[266] Kouhei Sekiguchi et al. « Autoregressive Moving Average Jointly-Diagonalizable Spatial Covariance Analysis for Joint Source Separation and Dereverberation ». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 2368–2382 (cit. on p. 76).

[267] Marius Rodrigues. *Déréverbération audio par apprentissage profond basé sur des modèles*. Rapport de stage M2. Palaiseau, France: Télécom Paris, 2025 (cit. on p. 80).

[268] Simon Schwär and Meinard Müller. « Multi-Scale Spectral Loss Revisited ». In: *IEEE Signal Processing Letters* 30 (2023), pp. 1712–1716 (cit. on p. 85).

[269] Roman Korostik, Rauf Nasretdinov, and Ante Jukić. « Modifying Flow Matching for Generative Speech Enhancement ». In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5 (cit. on p. 107).

[270] Constantinos Papayiannis, Christine Evers, and Patrick A. Naylor. « Sparse Parametric Modeling of the Early Part of Acoustic Impulse Responses ». In: *2017 25th European Signal Processing Conference (EUSIPCO)*. Aug. 2017, pp. 678–682 (cit. on p. 107).

Appendices

Appendix A

Proofs of Chapter 4

A.1 STFT reconstruction from the convolutive transfer function

From [62, Eq. (68)], let g_a and g_s the STFT analysis and synthsis windows respectively.

$$\mathcal{H}_{f,f',t} = \sum_l h[l] \sum_m g_a[m] e^{-j2\pi fm/N} g_s[tL - l + m] e^{j2\pi f'(tL-l+m)/N} \quad \text{(A.1)}$$

$$= \sum_l h[l] \sum_m g_a[m' - tL - l] e^{-j2\pi f(m'-tL-l)/N} g_s[m'] e^{j2\pi f'm'/N} \qquad (m' \leftarrow tL - l + m \quad \text{(A.2)}$$

By summing $(-1)^{f'}\mathcal{H}_{f,f',t}$ over f', we have:

$$\sum_{f'=0}^{N-1} (-1)^{f'} \mathcal{H}_{f,f',t} = \sum_{f'} \sum_l h[l] \sum_m g_a[m' - tL - l] e^{-j2\pi f(m'-tL-l)/N} g_s[m'] e^{j2\pi f'm'/N} \quad \text{(A.3)}$$

$$= \sum_l h[l] \sum_m g_a[m' - tL - l] e^{-j2\pi f(m'-tL-l)/N} g_s[m'] \sum_{f'=0}^{N-1} e^{j\pi(f'(2m'+1)/N} \quad \text{(A.4)}$$

Yet, $\sum_{f'=0}^{N-1} e^{j\pi(f'(2m'+1)/N} = 0$ unless $m' = N/2$, so:

$$\sum_{f'=0}^{N-1} (-1)^{f'} \mathcal{H}_{f,f',t} = g_s[N/2] \sum_l h(l) g_a[N/2 - tL - l] e^{-j2\pi f(N/2-tL-l)/N}. \quad \text{(A.5)}$$

The right-hand side is proportional to the STFT of h. Multiplying by $(-1)^{f'}$ enables to shift the analysis window by $-N/2$, and corresponds to the implementation of the

`pytorch` library [258], in which the first STFT analysis window is centered at the start of the signal.

A.2 Proximal gradient descent

A.2.1 Proximal operator

$$\text{prox}_{\lambda g}(x) \triangleq \underset{h}{\text{argmin}}\, \lambda g(h) + \frac{1}{2}\|h - x\|_2^2 \tag{A.6}$$

This proximal operator is separable, so it can be computed element-wise as:

$$\text{prox}_{\lambda g}(x)[n] = \underset{h[n]}{\text{argmin}}\, \frac{1}{2}(h[n] - x)^2 + \lambda \max\left(0, h[n]^2 - \sigma^2 e^{-2n/\tau}\right) \tag{A.7}$$

Two cases are straightforward:

— If $|x| \leq \sigma^2 e^{-2n/\tau}$:

$$\text{prox}_{\lambda g}(x)[n] = \underset{h[n]}{\text{argmin}}\, \frac{1}{2}(h[n] - x)^2 \tag{A.8}$$

$$= x \tag{A.9}$$

— If $|x| > \sigma^2 e^{-2n/\tau}$:

$$\text{prox}_{\lambda g}(x)[n] = \underset{h[n]}{\text{argmin}}\, \frac{1}{2}\|h[n] - x\|_2^2 + \lambda\left(h[n]^2 - \sigma^2 e^{-2n/\tau}\right) \tag{A.10}$$

$$= \frac{x}{2\lambda + 1} \tag{A.11}$$

The last situation arises when the proximal operator would bring $h[n]$ below the threshold of $\sigma^2 e^{-2n/\tau}$. In this case, $\text{prox}_{\lambda g}(x)[n] = \text{sign}(h[n])\,\sigma e^{-n/\tau}$

A.2.2 Exact line search

$$\eta = \underset{\eta \geq 0}{\text{argmin}}\, \mathcal{L}_d(h - \eta \nabla_{\mathcal{L}_d}(h)) \tag{A.12}$$

$$= \underset{\eta \geq 0}{\text{argmin}} \left\| \mathcal{T}(s)\left(h - 2\eta \mathcal{T}(s)^\mathsf{T}(\mathcal{T}(s)h - y)\right) - y \right\|_2^2 \tag{A.13}$$

$$= \underset{\eta \geq 0}{\text{argmin}} \left\| \mathcal{T}(s)h - y - 2\eta\left(\mathcal{T}(s)\mathcal{T}(s)^\mathsf{T}\mathcal{T}(s)h - \mathcal{T}(s)\mathcal{T}(s)^\mathsf{T} y\right) \right\|_2^2 \tag{A.14}$$

$$= \frac{a^\mathsf{T} b}{\|a\|_2^2}, \tag{A.15}$$

where $b = \mathcal{T}(s)h - y$ and $a = \mathcal{T}(s)\mathcal{T}(s)^{\mathsf{T}}b$. Using the property of Eq. (2.19), we have $b = s \star h - y$ and $a = 2s \star (s \circledast b)$

ECOLE
DOCTORALE

Titre: Déréverbération neuronale profonde hybride sensible à l'acoustique

Mots clés: Déréverbération; Apprentissage profond hybride; Réverbération; Traitement du signal audio

Résumé: Cette thèse porte sur la modélisation de l'acoustique des salles dans les approches d'apprentissage profond pour la déréverbération. Les signaux audio enregistrés sont souvent altérés par des effets de réverbération dûs à la propagation du son dans l'espace, ce qui nuit à leur intelligibilité et leur qualité. Cependant, la plupart des approches d'apprentissage profond développées pour atténuer ces effets restent en grande partie opaques et difficilement interprétables d'un point de vue physique. Après avoir étudié la compatibilité des réseaux neuronaux profonds existants avec des modèles d'acoustique des salles, nous proposons deux méthodes d'hybridation afin d'introduire des contraintes physiques dans leur apprentissage. La première régularise la fonction de perte d'entraînement d'un réseau neuronal profond pour encourager des solutions physiquement plausibles, et la seconde s'appuie sur une formulation en maximum de vraisemblance du problème de déréverbération et consiste en une stratégie d'apprentissage non supervisée intégrant un modèle de réverbération dans un réseau neuronal.

Title: Acoustics-aware hybrid deep neural dereverberation

Keywords: Dereverberation; Hybrid deep Learning; Reverberation; Audio signal processing

Abstract: The aim of this thesis is to leverage room acoustics models in deep-learning-based approaches for dereverberation. Audio signals are often altered by reverberation effects induced by objects and walls of the room in which they propagate, leading to a loss in intelligibility. However, most deep learning methods developed to tackle this problem can be considered as black-box systems, as they are purely data-driven and not interpretable from a physical perspective. After studying whether neural dereverberators are consistent with physical reverberation models, we propose two hybrid approaches to train a dereverberation model in a physically realistic manner. The first one regularizes the training loss to encourage a deep neural network to produce realistic solutions, and the second is motivated by a maximum-likelihood formulation of the problem and consists in an unsupervised learning strategy that integrates a reverberation model into a deep learning framework.

Institut Polytechnique de Paris
91120 Palaiseau, France