Data-driven room-acoustic modelling

Georg Götz
Data-driven room-acoustic modelling

Georg Götz

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall Jeti of the school on 4 July 2024 at 12:00.

Aalto University
School of Electrical Engineering
Department of Information and Communications Engineering
Communication Acoustics
Supervising professor
Prof. Ville Pulkki, Aalto University, Finland

Thesis advisor
Prof. Sebastian J. Schlecht, Aalto University, Finland

Preliminary examiners
Prof. Zühre Sü Gül, Bilkent University, Turkey
Prof. Elias Zea, KTH Royal Institute of Technology, Sweden

Opponent
Prof. Cheol-Ho Jeong, Technical University of Denmark, Denmark

Aalto University publication series
DOCTORAL THESES 133/2024

© 2024 Georg Götz

ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2024

Finland
Abstract

Author
Georg Götz

Name of the doctoral thesis
Data-driven room-acoustic modelling

Publisher
School of Electrical Engineering

Unit
Department of Information and Communications Engineering

Series
Aalto University publication series DOCTORAL THESES 133/2024

Field of research
Acoustics and Audio Signal Processing

Manuscript submitted 7 March 2024

Date of the defence 4 July 2024

Permission for public defence granted (date) 26 April 2024

Language
English

Monograph Article thesis Essay thesis

Abstract
The study of room acoustics has traditionally been of interest in architectural planning and design. With the spread of virtual- and augmented-reality technology, room-acoustic modelling has also become increasingly relevant for audio engines. The dynamic and fast-paced nature of such applications requires rendering systems to operate in real-time. However, accurate state-of-the-art room-acoustic-simulation technology is often computationally expensive, limiting its use for audio engines. Data-driven methods offer the potential to bypass expensive simulations, while ensuring convincing perceptual experiences. This dissertation works towards data-driven audio engines by exploring the interaction between room-acoustic modelling and data-driven methods. It comprises five peer-reviewed publications that investigate automatic data acquisition, robust room-acoustic analysis in complex environments, and data-driven room-acoustics rendering.

As sound propagates through a room, it interacts with various surfaces, leading to a gradual energy decay over time. The properties of this energy decay significantly influence the acoustic impression evoked by a room, making it a widely studied topic in room-acoustic research. The first part of this thesis provides an overview of sound-energy decay, its analysis, and challenges associated with complex geometries featuring multiple rooms and non-uniform absorption-material distributions. To this end, it introduces a neural network for multi-exponential sound-energy-decay analysis. Moreover, spatial and directional variations of sound-energy decay are investigated, and a compact representation to model them is proposed.

The second part of this thesis is centred around data-driven methods and explores how they can be applied to room-acoustics research. After elaborating on the properties of room-acoustic data, techniques for its large-scale acquisition are investigated. Two of the contained publications describe autonomous robot systems for conducting room-acoustic measurements. While the first one describes the general idea and the design constraints of a practical system, the second one extends the measurement strategy to complex geometries featuring multiple connected rooms. An overview of commonly used machine-learning concepts is provided, focusing on the ones relevant for the included publications. Finally, several applications of data-driven methods in room-acoustics research are described, including a summary of a late-reverberation rendering system proposed in one of the appended publications.

Keywords Room acoustics, sound-energy decay, inhomogeneous and anisotropic late reverberation, autonomous measurement robots, automatic data acquisition, machine learning, late-reverberation rendering


ISSN (printed) 1799-4934 ISSN (pdf) 1799-4942

Location of publisher Helsinki Location of printing Helsinki Year 2024

Preface

The research presented in this dissertation was conducted at the Acoustics Laboratory of the Department of Information and Communications Engineering at Aalto University between November 2019 and March 2024, where it received funding from the Academy of Finland (project no. 317341) and the School of Electrical Engineering of Aalto University. Additionally, I want to thank the Nokia Foundation for awarding me the Nokia Scholarship. I also acknowledge the computational resources provided by the Aalto Science-IT project, which were crucial for the work presented here.

This work was also supported by many brilliant people I had the opportunity to work with over the last years. First and foremost, I want to express my gratitude to my supervisor Prof. Ville Pulkki. I was extremely lucky that I got to meet you during my exchange studies, and that you gave me the opportunity to conduct my own little research project as a summer job, which would later become a long summer of 830 measurements. Thank you for the constant source of inspiration, and showing me the creative side of the research process. Many times, your out-of-the-box thinking guided me towards the last missing puzzle piece before completing a project. I also want to thank my advisor Prof. Sebastian J. Schlecht. Your guidance along my research journey was invaluable. I am very thankful that you sharpened my critical thinking skills, kept me on the right track, and taught me about the intricacies of academic life and the writing process. Your effort of bringing many doctoral candidates together to our weekly round table helped me tremendously to learn about various topics and stay sane during the last four years.

I also want to thank the preliminary examiners, Prof. Zühre Sü Gül and Prof. Elias Zea, for their valuable, constructive, and prompt feedback that helped to improve this manuscript, and Prof. Cheol-Ho Jeong for agreeing to act as my opponent.

Being surrounded by so many brilliant and like-minded people at the Aalto Acoustics Laboratory was inspiring and lots of fun. From day one, I already felt like I was halfway there, although I had just started with my research. I want to thank Chris, Pedro, and Stefan for not only being
my colleagues, but also my good friends, for joining me during countless adventures across Finland, and learning a thing or two about microbiology with me. I am also grateful to all members of the weekly “Sound in VR” round table – Andrea, Gian Marco, Gloria, Jana, Janis, Jon, Karolina, Kyung, Nils, and Tom – for the many insightful discussions, and the weekly reminder that the doctoral research journey consists of many ups- and downs that are better to celebrate and easier to digest together. I further want to thank Abraham for building the first prototype of the measurement robots, Aleksi for being such a great help with all practical things around the lab, Ricardo for sharing his vast machine learning knowledge with me, Teodors for the many fruitful discussions about how to make practical use of the common-slope model, and Alec, Antti, Eloi, Henri, Jackie, Janani, Julie, Leo, Leonardo, Maddie, Otto, Petteri, Raimundo, Taeho, and Vasileios for being fantastic colleagues.

Before coming to Aalto, I conducted my first research experiments during my Bachelor’s and Master’s studies in Ilmenau. I am incredibly grateful to my advisors at the time, Stephan Werner and Florian Klein, who sparked my fascination for scientific work and encouraged me to introduce myself at the Aalto Acoustics Lab during my exchange studies. I am very thankful that you opened this door for me, and that our collaboration and scientific discussions continued across institutions.

During my doctoral research, I was lucky enough to complete a research internship at Meta Reality Labs Research. The research I conducted there also became a part of this dissertation, and I want to thank Paul Calamia, Sebastià V. Amengual Garí, and Ishwarya Ananthabhotla for their mentoring and guidance.

Beyond work, I am lucky to have many wonderful and supportive people in my life. I am grateful to my friends in Germany, Finland, and abroad for your friendship over the years and across borders. Even if there are occasionally periods of radio silence, I am lucky to know that everything is like usual when we meet again. I also want to thank my family for supporting me and believing in me throughout my whole life, especially during the last years. No matter where I go or what I do, I am always happy and grateful to return to such a loving home, and share the latest adventures with you. Finally, I am incredibly grateful to Jamila for your kind and warm heart, your unconditional and endless support, and for being the best company I could wish for during the last bit of this journey.

Helsinki, June 9, 2024,

Georg Götz
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>1</td>
</tr>
<tr>
<td>Contents</td>
<td>3</td>
</tr>
<tr>
<td>List of publications</td>
<td>5</td>
</tr>
<tr>
<td>Author’s contribution</td>
<td>7</td>
</tr>
<tr>
<td>List of figures</td>
<td>9</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>11</td>
</tr>
<tr>
<td>Symbols</td>
<td>13</td>
</tr>
<tr>
<td><strong>1. Introduction</strong></td>
<td>15</td>
</tr>
<tr>
<td><strong>2. Sound-energy decay in enclosed spaces</strong></td>
<td>19</td>
</tr>
<tr>
<td>2.1 Sound propagation in rooms</td>
<td>19</td>
</tr>
<tr>
<td>2.2 Energy-decay analysis</td>
<td>21</td>
</tr>
<tr>
<td>2.3 Multi-exponential decay</td>
<td>22</td>
</tr>
<tr>
<td>2.4 Inhomogeneous and anisotropic decay</td>
<td>25</td>
</tr>
<tr>
<td><strong>3. Data-driven room-acoustic modelling</strong></td>
<td>29</td>
</tr>
<tr>
<td>3.1 Room-acoustic data</td>
<td>29</td>
</tr>
<tr>
<td>3.2 Large-scale data acquisition</td>
<td>30</td>
</tr>
<tr>
<td>3.3 Machine learning in room acoustics</td>
<td>32</td>
</tr>
<tr>
<td>3.3.1 Input features</td>
<td>33</td>
</tr>
<tr>
<td>3.3.2 Loss functions</td>
<td>34</td>
</tr>
<tr>
<td>3.3.3 Commonly used architectures</td>
<td>35</td>
</tr>
<tr>
<td>3.4 Applications to room-acoustic rendering</td>
<td>38</td>
</tr>
<tr>
<td><strong>4. Summary of contributions</strong></td>
<td>41</td>
</tr>
<tr>
<td><strong>5. Conclusions and future work</strong></td>
<td>45</td>
</tr>
</tbody>
</table>
Contents

References 47
Publications 61
This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.


Author’s contribution

Publication I: “Autonomous robot twin system for room acoustic measurements”

VP initiated the study. The original idea was further refined by the author and AMO. AMO built a first prototype of the system, which was further developed by the author. The author conducted all experiments under the guidance of SJS and VP. The author wrote the manuscript and produced all figures and tables. The co-authors reviewed the manuscript and commented on it.

Publication II: “Autonomous room acoustic measurements using rapidly-exploring random trees and Gaussian processes”

The initial idea originated from discussions between the author and the co-authors. The proposed measurement strategy was conceptualised and implemented by the author, under the guidance from the co-authors. The author conducted all experiments under the guidance of the co-authors. The author wrote the manuscript and produced all figures and tables. The co-authors reviewed the manuscript and commented on it.

Publication III: “Neural network for multi-exponential sound energy decay analysis”

The initial idea originated from discussions between the author and SJS. The neural network was conceptualised and implemented by the author under the guidance of SJS, taking into account suggestions from RFP and VP. The author simulated the synthetic dataset and conducted all experiments under the guidance of SJS. RFP developed the initial version
of the toolbox, which has been maintained and further developed by the author since then. The author wrote the manuscript, except Sec. III.B and VII., which were written by RFP. The author produced all figures and tables, except Fig. 4, which was produced by RFP. The co-authors reviewed the manuscript and commented on it.

**Publication IV: “Common-slope modeling of late reverberation”**

The initial idea originated from discussions between the author and the co-authors. The detailed approach for determining common slopes from RIR sets was conceptualised by the author under the guidance of SJS. The author simulated the synthetic dataset and conducted all experiments under the guidance of SJS. The author wrote the manuscript and produced all figures and tables. The co-authors reviewed the manuscript and commented on it.

**Publication V: “Dynamic late reverberation rendering using the common-slope model”**

The initial idea originated from discussions between the author, SJS, and VP. The author detailed the rendering system in discussions with TK and SJS. A first prototype of the system was implemented by TK, and further improved by the author to its published state, under the guidance of SJS. The author simulated the synthetic dataset used in the evaluation. The author produced the demo renderings, taking into account suggestions from SJS. The author produced all figures and wrote the manuscript. The co-authors reviewed the manuscript and commented on it.
List of figures

2.1 Different graphical representations of sound-energy decay in rooms: (a) RIR, (b) squared RIR, and (c) EDC from the Schroeder backwards integration. 22

2.2 Example of a multi-exponential sound-energy decay featuring $\kappa = 2$ decay rates. The dashed and dash-dotted line depict the two exponentials, and the dotted line corresponds to the background noise term. 24

2.3 Inhomogeneity of the decaying sound field in a coupled-room environment consisting of an acoustically treated meeting room and a reverberant hallway. The plots show EDF values over the entire scene at several time instances. 26

3.1 Different commonly used deep-learning components. 36
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Augmented reality</td>
</tr>
<tr>
<td>BEM</td>
<td>Boundary element method</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian information criterion</td>
</tr>
<tr>
<td>CAPR</td>
<td>Common-acoustical-pole and residue</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional neural network</td>
</tr>
<tr>
<td>DeepONet</td>
<td>Deep operator network</td>
</tr>
<tr>
<td>DGFEM</td>
<td>Discontinuous Galerkin finite element method</td>
</tr>
<tr>
<td>DOA</td>
<td>Direction of arrival</td>
</tr>
<tr>
<td>DOD</td>
<td>Direction of departure</td>
</tr>
<tr>
<td>EDC</td>
<td>Energy-decay curve</td>
</tr>
<tr>
<td>EDF</td>
<td>Energy-decay function</td>
</tr>
<tr>
<td>FDTD</td>
<td>Finite-difference time-domain</td>
</tr>
<tr>
<td>FEM</td>
<td>Finite element method</td>
</tr>
<tr>
<td>GA</td>
<td>Geometrical acoustics</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian process</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean absolute error</td>
</tr>
<tr>
<td>MLP</td>
<td>Multilayer perceptron</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean squared error</td>
</tr>
<tr>
<td>PINN</td>
<td>Physics-informed neural network</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified linear unit</td>
</tr>
<tr>
<td>RIR</td>
<td>Room impulse response</td>
</tr>
<tr>
<td>RT</td>
<td>Reverberation time</td>
</tr>
<tr>
<td>RRT</td>
<td>Rapidly-exploring random tree</td>
</tr>
<tr>
<td>SEM</td>
<td>Spectral element method</td>
</tr>
<tr>
<td>ULI</td>
<td>Upper limit of integration</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual reality</td>
</tr>
<tr>
<td>1D, 3D</td>
<td>One-dimensional, three-dimensional</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$A$</td>
<td>decay amplitude</td>
</tr>
<tr>
<td>$b$</td>
<td>bias term in neural network</td>
</tr>
<tr>
<td>$d$</td>
<td>energy-decay function</td>
</tr>
<tr>
<td>$f_s$</td>
<td>sampling frequency</td>
</tr>
<tr>
<td>$h$</td>
<td>room impulse response</td>
</tr>
<tr>
<td>$k$</td>
<td>exponential-decay index, mode-group index, decay-slope index</td>
</tr>
<tr>
<td>$L$</td>
<td>upper limit of integration in Schroeder backwards integration</td>
</tr>
<tr>
<td>$N$</td>
<td>number of neural-network inputs</td>
</tr>
<tr>
<td>$M$</td>
<td>number of neurons in a linear layer</td>
</tr>
<tr>
<td>$N_0$</td>
<td>noise amplitude</td>
</tr>
<tr>
<td>$t$</td>
<td>discrete time index</td>
</tr>
<tr>
<td>$\hat{t}$</td>
<td>time</td>
</tr>
<tr>
<td>$T$</td>
<td>decay time</td>
</tr>
<tr>
<td>$w$</td>
<td>weight in neural network</td>
</tr>
<tr>
<td>$x$</td>
<td>neural-network input</td>
</tr>
<tr>
<td>$x_s$, $x_r$</td>
<td>Cartesian $x$-coordinate of source and receiver, respectively</td>
</tr>
<tr>
<td>$\mathbf{x}$</td>
<td>source-receiver configuration</td>
</tr>
<tr>
<td>$\mathbf{x}_s$</td>
<td>source position</td>
</tr>
<tr>
<td>$\mathbf{x}_r$</td>
<td>receiver position</td>
</tr>
<tr>
<td>$y$</td>
<td>neural-network output</td>
</tr>
<tr>
<td>$y_s$, $y_r$</td>
<td>Cartesian $y$-coordinate of source and receiver, respectively</td>
</tr>
<tr>
<td>$z_s$, $z_r$</td>
<td>Cartesian $z$-coordinate of source and receiver, respectively</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>kernel length in 1D convolutional layer</td>
</tr>
<tr>
<td>$\theta$</td>
<td>elevation angle</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>number of exponential decays, number of mode groups, decay-model order</td>
</tr>
</tbody>
</table>
Symbols

$\sigma$  neural-network activation function

$\phi$  azimuth angle

$\Psi$  decay kernel

$\Omega_s$  direction of departure from the source

$\Omega_r$  direction of arrival at the receiver
1. Introduction

Understanding and manipulating the acoustic properties of rooms is highly relevant to modern society. People spend a significant amount of their time in indoor environments, and the prevailing acoustic conditions can influence their comfort [1, 2], and even their health and well-being [3]. Environmental noise protection in residential dwellings is a significant aspect of public health, prompting the World Health Organization (WHO) to publish guidelines for protecting human health from environmental noise exposure [3].

As a research field, room-acoustic modelling has a long tradition, dating back to early works by Sabine [4], Eyring [5, 6], and others [7]. Sabine's pioneering work was data-driven, involving many reverberation measurements in the same hall to investigate the relationship between material absorption and sound decay in rooms [4]. Many subsequent room-acoustics studies follow a similarly empirical paradigm. However, our present room-acoustics knowledge is equally grounded on a strong theoretical foundation [8, 9].

Combining data-driven and theory-driven research can significantly advance scientific knowledge, possibly surpassing what might be achieved through either approach independently [10]. Identifying patterns within datasets and exploiting them to solve room-acoustic problems can contribute to the development of new theories. For instance, in his work on coupled rooms, Eyring showed measured data, explained that sound-energy decays can feature multiple distinct slopes, and derived a theoretical basis for his observations [6]. Until the work of Xiang and Goggans, there were no appropriate models and tools to quantitatively analyse such energy decays [11]. Since then, many further studies advanced the theory of multi-exponential sound-energy decays and the corresponding sound fields [12, 13, 14, 15, 16].

On the other hand, a solid theoretical base can guard against the misinterpretation of empirical data. For example, the ISO standard 3382-2 includes a cautionary note recommending not to use a single-slope model to determine the reverberation time of a room with a multi-exponential
sound-energy decay [17]. Applying a single-slope model in such situations may lead to flawed estimates. Such a model mismatch can have particularly severe implications for large-scale data analyses [18]. Similarly, the robustness of room-acoustic analysis tools for extracting ground truth labels must be considered when developing data-driven room-acoustic methods. Flawed training labels may result in data-driven systems that perform well on the chosen metrics, but fail when considering a more complete theoretical model, a lower data abstraction level, or corrected training labels. Therefore, choosing appropriate theoretical models and robust analysis methods is crucial for data-driven room-acoustic modelling.

Early room-acoustic studies were inherently constrained by the prevailing measurement technology of their time, such as chronographs [4] or magnetic tape recorders [19]. Nowadays, room-acoustic-simulation technology is widely available, and measurements can be conducted with merely a loudspeaker, a microphone, and a laptop, making large room-acoustic datasets ubiquitous. Nevertheless, both approaches for room-acoustic data acquisition have drawbacks. While room-acoustic simulations still exhibit algorithmic deficiencies and suffer from uncertainties [20, 21], room-acoustic measurements are time-consuming and tedious. Furthermore, spatial or directional measurements require large microphone arrays and considerable installation effort. As data-driven methods are often trained on large datasets, many room-acoustic studies increase the amount of data by combining multiple measured and simulated datasets. However, a combination of multiple datasets may lead to inconsistencies regarding measurement and simulation parameters, and potentially complicate training and evaluation. Consequently, data-driven room-acoustic approaches must carefully select the appropriate data and determine suitable tasks and targets to obtain meaningful results.

This thesis was shaped by the interaction between data-driven and theory-driven research. It contains this introductory part and five peer-reviewed publications. The publications can be divided into three groups.

i) Publication I and Publication II focused on efficient room-acoustic data acquisition. Although room acoustic simulations offer a flexible tool to obtain large quantities of room-acoustic data, their accuracy can be inferior to room-acoustic measurements, with simulated data deviating considerably from measurements in certain controlled benchmark tests [22]. However, room-acoustic measurements are tedious and time-consuming, and previous efforts to automate them were limited to measurements over small spatial areas or measurement systems that require considerable installation effort. Therefore, Publication I proposed a mobile autonomous robot system to measure large room-acoustic datasets over extended spatial regions, requiring only minimal planning, setup, and supervision. As research for this thesis progressed, room-acoustic modelling in complex environments...
Introduction

became increasingly important. Therefore, Publication II further explored autonomous room-acoustic measurement robots and presented a new measurement strategy suitable for multi-room environments.

ii) Publication III and Publication IV are centred around robust room-acoustic analysis in complex environments. Over the past years, machine-learning approaches gained popularity in acoustic research [23, 24], necessitating robust analysis tools to process large room-acoustic datasets. Therefore, Publication III proposed a lightweight neural network to analyse multi-exponential energy decays. Additionally, the publication evaluated the robustness of the proposed architecture and a state-of-the-art Bayesian decay-analysis approach on a large-scale dataset. The efficiency of the neural-network-based analysis enabled thorough analyses of large room-acoustic datasets, exposing spatial and directional sound-energy-decay variations. Consequently, Publication IV introduced the common-slope model that describes spatial and directional late-reverberation variations using a compact representation. Room-acoustic analyses using the common-slope model yield interpretable results that offer new insights on spatially and directionally varying late reverberation. For example, the reverberation fade between two coupled rooms can be explicitly described using the common-slope model.

iii) Finally, Publication V introduced a new approach for data-driven room-acoustic rendering, based upon the theoretical foundations outlined in Publication III and Publication IV. Room-acoustic renderings aims to imprint a room-acoustic sensation on a dry source signal in an audio engine. Virtual- and augmented-reality (VR/AR) applications and computer games often feature complex room geometries with interconnected rooms and non-uniform absorption. Efficient and accurate rendering of such scenes can be challenging. Publication V leverages the potential of the compact representation proposed in Publication IV, requiring only few reverberators to render spatially and directionally varying late reverberation.

The remainder of this thesis is structured as follows. Chapter 2 focuses on the sound-energy decay in enclosed spaces, addressing multi-exponential decays and the intricacies of spatially and directionally varying sound-energy decay. Chapter 3 provides an overview of data-driven room-acoustic modelling, elaborating on large-scale room-acoustic data collection, commonly used machine-learning concepts, and possible applications. Chapter 4 summarises the contributions of the publications included in this thesis. Chapter 5 presents the conclusions drawn from the research conducted for this thesis.
2. Sound-energy decay in enclosed spaces

Room-acoustics research studies sound propagation and the resulting sound field in enclosed spaces. Sound waves interact with the enclosure boundaries, such as its walls, floor, and ceiling, as well as with other surfaces like furniture. Each interaction affects the sound energy in the enclosure, causing different fractions of the incoming energy to be reflected, absorbed, or transmitted to the outside [8]. The combined effects of these interactions cause the total sound energy within the enclosure to decay over time.

This chapter describes the sound-energy decay in enclosed spaces. The first section elaborates on fundamentals of sound propagation in rooms, followed by a section exploring sound-energy-decay analysis. The subsequent section addresses multi-exponential decays, why they occur, and how they can be analysed. Finally, the last section covers inhomogeneous and anisotropic sound fields, and their implications on the sound-energy decay.

2.1 Sound propagation in rooms

When sound propagates between a sound source and a receiver inside a room, its combined acoustic transmission path is typically given by the room impulse response (RIR) $h(x,t)$. Throughout this thesis, $t$ is the discrete time index at sample rate $f_s$, whereas $\tilde{t} = t/f_s$ is the time in seconds. The source-receiver configuration $\mathbf{x} = (\mathbf{x}_s, \mathbf{x}_r, \Omega_s, \Omega_r)$ characterises the spatial and directional properties of the transmission path. More specifically, $\mathbf{x}_s = (x_s, y_s, z_s)$ and $\mathbf{x}_r = (x_r, y_r, z_r)$ denote the source and receiver position in three Cartesian coordinates, respectively. The direction of departure (DOD) from the source $\Omega_s = (\phi_s, \theta_s)$ and direction of arrival (DOA) at the receiver $\Omega_r = (\phi_r, \theta_r)$ are required for directional RIRs, i.e. for cases where the source or receiver deviates from an omnidirectional directivity pattern. Both directions are given in terms of the azimuth angle $\phi$ and the elevation angle $\theta$ of a right-handed spherical coordinate system as defined in ISO 80000-2 [25].
RIRs can be obtained from room-acoustic measurements or computer simulations. In RIR measurements, a sound source excites the room either with an impulse [26, 27, 28, 29] or by playing a continuous excitation signal, like a noise sequence [30, 31] or sine sweep [32, 33] using a loudspeaker. The receiver records the resulting sound field with a microphone. Lastly, the recording is deconvolved with the excitation signal to obtain the RIR [34]. For RIR measurements covering multiple source-receiver configurations \( x \), the sound source and microphone must be physically moved to different positions in the room. This manual work can become tedious and time-consuming when measuring large datasets with many source-receiver configurations. Additionally, multi-channel microphone or loudspeaker arrays often require a significant amount of cables and additional hardware, further complicating large-scale measurements. Alternatively, measurement setups can contain multiple loudspeakers and microphones, allowing for subsequent measurements of multiple source-receiver configurations without moving the individual components. However, this measurement paradigm requires multiples loudspeakers and microphones, making it more expensive.

In computer simulations, the sound propagation from source to receiver can be computed in various ways using computer models of the room. Geometrical acoustics (GA) methods assume that sound propagates as rays, thus neglecting its wave properties [35]. While this assumption is valid for high frequencies, i.e. short wavelengths compared to the surfaces in the simulation, it can be challenging to model wave effects like diffraction at lower frequencies [36, 37, 38]. In contrast, wave-based simulations solve the wave equation and inherently model wave phenomena like diffraction [39]. Therefore, they potentially yield physically more accurate results, at the expense of higher computational cost than GA methods. Various wave-based simulation approaches exist, including the finite-difference time-domain (FDTD) method [39, 40], the boundary element method (BEM) [41], the finite element method (FEM) [42], the spectral element method (SEM) [43], and the discontinuous Galerkin finite element method (DGFEM) [44, 45]. GA and wave-based methods can also be combined into hybrid simulations to accurately simulate RIRs over the entire audible bandwidth without spending excessive computational resources [46].

Despite the considerable advancement of room acoustic simulations over the past decades, the accuracy and perceptual quality of simulated RIRs is still inferior to those obtained from room-acoustic measurements [22]. Apart from algorithmic deficiencies, uncertainties related to input material properties and geometry also influence the simulation results [20, 21]. In contrast, room-acoustic measurements can be noisy or faulty, necessitating the detection of flawed RIRs [47, 48]. Nevertheless, given the current state of measurement and simulation technology, measurements typically
capture the prevailing room-acoustic conditions of real rooms better than simulations, because they do not suffer from the simulation uncertainties described above. RIR simulations can still be highly valuable for architectural planning, where the investigated rooms might not physically exist yet.

2.2 Energy-decay analysis

The ISO 3382 standard describes two methods for determining the sound-energy decay in enclosed spaces [49, 17]. In the interrupted noise method, the sound source excites the room until a steady state is reached. Subsequently, the source is turned off, and the resulting energy decay is recorded [49, 17, 8]. Random fluctuations in the measured decay can be mitigated by averaging multiple decay recordings. The integrated impulse response method, proposed by Schroeder [19], is commonly referred to as Schroeder backwards integration. Schroeder discovered that integrating the time-reversed RIR is equivalent to averaging multiple decay recordings [19].

The Schroeder backwards integration method yields the energy-decay function (EDF), or the energy-decay curve (EDC) when presented graphically. In discrete time, the integration turns into a sum, and the EDF for source-receiver configuration \( \mathbf{x} \) can be determined by [49, 17, 19]

\[
d(\mathbf{x}, t) = \sum_{l=t}^{L} h^2(\mathbf{x}, l),
\]

where \( L \) is called upper limit of integration (ULI). The ULI considerably influences the resulting EDF due to the background noise inherent in room-acoustic measurements [50, 8], thus prompting many studies on its optimal choice [51, 52, 53].

Figure 2.1 illustrates different graphical representations of sound-energy decay in rooms. An RIR \( h(\mathbf{x}, t) \), like the one depicted in Figure 2.1a, is a typical result from a room-acoustic measurement or simulation. However, a representation on a linear scale is not suitable for analysing the full dynamic range of the sound-energy decay. In contrast, a squared RIR \( h^2(\mathbf{x}, t) \) plotted on a logarithmic scale (see Figure 2.1b) provides insights on the entire energy decay including the noise floor. Figure 2.1c shows the corresponding EDF, illustrating that the backwards integration eliminates the random fluctuations of the squared RIR.

Due to frequency-dependent attenuation, sound-energy-decay analysis is typically carried out in frequency bands, such as octave or one-third-octave bands [49, 17]. Therefore, a filter bank processes either the excitation noise (in the interrupted noise method) or the RIR (in the Schroeder backwards integration) before proceeding with the decay analysis [49, 17].
Sound-energy decay in enclosed spaces

Figure 2.1. Different graphical representations of sound-energy decay in rooms: (a) RIR, (b) squared RIR, and (c) EDC from the Schroeder backwards integration.

The utilised filter bank should be highly selective, because leakage between bands affects the decay analysis results [54].

The reverberation time (RT) is a commonly used metric to describe the sound-energy decay in rooms. It is typically determined in frequency bands, denoted by $T_{60}$, and defined as the duration until the sound energy in a room has decayed by 60 dB from the steady-state after the source has been turned off [49, 17]. With sound energy decaying exponentially over time, the RT can be estimated by fitting a straight line to the EDF on a logarithmic scale [49, 17]. For EDFs with dynamic ranges smaller than 60 dB, a decay of 20 dB or 30 dB can be evaluated and extrapolated to 60 dB instead, corresponding to the measures $T_{20}$ and $T_{30}$, respectively [49, 17]. Noise subtraction [55, 53] or RIR truncation [51, 52, 53] mitigate the influence of background noise on the RT estimate. Alternatively, non-linear models with a dedicated noise term can fit the entire EDF including background noise to obtain unbiased RT estimates $T$ [56, 57].

2.3 Multi-exponential decay

Sound-energy-decay analysis only yields meaningful RT estimates if the investigated frequency band exhibits a single decay rate. Otherwise, the decay cannot be characterised using only one exponential, thus requiring a more elaborate model featuring multiple exponentials with different decay rates. Such multi-exponential decays were already described as early as 1931 by Eyring when he investigated the energy decay in two rooms coupled together through an open window [6]. Since then, many studies explored multi-exponential decays in the context of coupled rooms [8, 11, 58, 13, 59, 60, 61, 62, 63], multi-volume structures [12, 14, 16, 64], and rooms with nonuniform absorption distributions [65, 66, 67, 68, 69].
In coupled rooms, multi-exponential decays arise from the energy exchange between both rooms [6, 8, 58]. The size of the aperture between the rooms has a significant effect on the energy flow [13], and various models relate the coupled-room decay to the decays of the individual rooms [8, 58, 60, 61, 59] and the distance between source, receiver, and aperture [60, 61, 59]. Recent research has demonstrated that these coupling models are valuable for developing efficient reverberation algorithms in coupled-room environments [70].

The emergence of multi-exponential decays in rooms with nonuniform absorption can be explained by dividing the room modes into groups [65, 68, 69]. In a room with nonuniform absorption, where some walls are highly absorptive while the others are reflective, certain modes arise from standing waves that interact with the absorptive surface (non-grazing incidence), whereas others run parallel to it and have only minimal interaction (grazing incidence). The decay behaviour of these mode groups differs considerably, because modes with non-grazing incidence will exhibit a smaller decay time than those with grazing incidence [65, 68]. As a consequence, the sum of all mode groups results in multi-exponential decay. In general, the EDF can be determined through a Laplace transform of the modal-damping density [8, 71, 72].

The concept of reverberation time can be easily extended to multi-exponential decays by using a model consisting of multiple superimposed exponentials. A model containing \( \kappa \) exponentials with different decay rates is given by [11, 73]

\[
d_{\kappa}(x,t) = N_{0,x}(L-t) + \sum_{k=1}^{\kappa} A_{k,x} \left[ \exp\left(-\frac{13.8}{f_s T_{k,x}} t\right) - \exp\left(-\frac{13.8}{f_s T_{k,x}} L\right) \right],
\]

where \( T_{k,x} \) is the decay time of the \( k \)th exponential term, and \( A_{k,x} \) is the corresponding decay amplitude. The model contains a dedicated term to account for the background noise, and \( N_{0,x} \) is the corresponding noise amplitude. Because of the constant \( -13.8 = \ln(10^{-6}) \), the decay time \( T_{k,x} \) quantifies the time in seconds until the \( k \)th exponential has decayed by 60 dB, thus resembling the definition of RT. The number of exponentials \( \kappa \) is also called model order. For \( \kappa = 1 \), the model corresponds to the single-rate decay model with background noise for estimating RT [56]. The second term in the square brackets is a constant accounting for the finite ULI [52, 74]. It can be neglected for large \( L \) [74].

The model can also be written in a more concise form,

\[
d_{\kappa}(x,t) = N_{0,x} \Psi_{0,x}(t) + \sum_{k=1}^{\kappa} A_{k,x} \left[ \Psi_{k,x}(t) - \Psi_{k,x}(L) \right],
\]
Figure 2.2. Example of a multi-exponential sound-energy decay featuring $\kappa = 2$ decay rates. The dashed and dash-dotted line depict the two exponentials, and the dotted line corresponds to the background noise term.

by combining the exponential decays and noise term into the decay kernel

$$
\Psi_{k,x}(t) = \begin{cases} 
L - t, & \text{if } k = 0 \\
\exp\left(-13.8 t \frac{1}{T_k x}\right), & \text{if } k > 0
\end{cases}.
$$

(2.4)

Figure 2.2 illustrates an example of a multi-exponential decay from a coupled-room environment. The decay can be described using two exponential decays, which are depicted as the dashed and dash-dotted line. Both exponentials exhibit considerably different slopes. Sound-energy decays exhibiting multiple slopes when plotted on a logarithmic scale are also called multi-slope decays.

The sound-energy decay described by an RIR can be fully characterised after estimating the model parameters $T_{k,x}$, $A_{k,x}$, and $N_{0,x}$. For single-rate decays without background noise, the parameter estimation corresponds to fitting a straight line to the logarithmic EDC. For noisy decays or decays with multiple decay rates, the parameter estimation requires more advanced approaches. Although the decay amplitudes $A_{k,x}$ and noise amplitude $N_{0,x}$ occur as linear coefficients in Equation (2.3), nonlinear parameter estimation approaches are required for fitting the model, because the decay times $T_{k,x}$ appear in the exponent [56, 57]. To this end, Xiang and Goggans proposed a Bayesian formalism for sound-energy-decay analysis [11], which can also be used to determine the model order [75, 76]. Handling the large, multivariate search space associated with Bayesian sound-energy-decay analysis can be computationally challenging. However, the computational load can be reduced by using various advanced sampling approaches [11, 74, 77, 78, 79] or by approximating the Bayesian evidence via the Bayesian information criterion (BIC) [76].

Publication III proposes the DecayFitNet, an alternative decay-analysis approach based on a neural network. In contrast to the previously de-
scribed iterative Bayesian approaches, the neural network is deterministic during inference. Numerical difficulties and the appropriate choice of hyperparameters have moved from the decay analysis to the network training stage. Despite being lightweight and computationally efficient, the experiments outlined in Publication III showed that the decay-analysis performance of the neural network is comparable to state-of-the-art Bayesian approaches. The proposed neural network architecture is especially appealing for analysing large decay datasets, or implementing it in environments with limited computational resources, such as mobile devices.

2.4 Inhomogeneous and anisotropic decay

By definition, RIRs describe the acoustic transmission path for a specific source-receiver configuration $x$. The diffuse sound field is a popular model that assumes uniform average energy density for all positions and uniform probability of energy flow in all directions [80, 8]. These idealised conditions are rarely met in practice, and the diffuseness of a sound field can be quantified using a variety of methods [81, 82, 83, 84, 85]. Sound fields in many environments exhibit positional and directional variations, which are also referred to as sound field inhomogeneity and anisotropy, respectively. Consequently, in a non-diffuse sound field, the RIR changes when either the source or the receiver moves or rotates, and these changes also affect the EDF, as seen from Equation (2.1).

Figure 2.3 illustrates an example of inhomogeneous sound-energy decay in a coupled-room environment consisting of an acoustically treated meeting room and a reverberant hallway. The RIRs that were used to produce this figure were simulated using a wave-based approach with a spatial resolution of 0.2 m on the $xy$-plane at 1.55 m height. More details on the dataset can be found in Publication IV. The figure shows the EDF value at several time instances plotted over the entire scene. At $\tilde{t} = 0 \text{ms}$, the EDF corresponds to the steady-state energy, and Figure 2.3a reveals a highly inhomogeneous energy distribution with higher sound-energy levels in the meeting room than in the hallway. A similar energy distribution is observed at $\tilde{t} = 10 \text{ms}$. At $\tilde{t} = 100 \text{ms}$, the energy levels of both rooms approach each other, resulting in an almost homogeneous energy distribution. However, after a while, the energy levels diverge again due to the higher absorption in the meeting room, as illustrated by Figure 2.3d and 2.3e. After $\tilde{t} = 200 \text{ms}$, the hallway exhibits higher energy levels than the meeting room for the remainder of the decay.

Various studies investigated the sound-energy decay in inhomogeneous and anisotropic sound fields. Many of them focused on RT or other ISO 3382 parameters [49, 17], and studied their variations with respect to position [86, 62, 87, 88, 89] and direction [90, 91, 92, 93]. In the context
Figure 2.3. Inhomogeneity of the decaying sound field in a coupled-room environment consisting of an acoustically treated meeting room and a reverberant hallway. The plots show EDF values over the entire scene at several time instances.

of coupled rooms, some studies also explored spatial variations of multi-exponential decay parameters [13, 94]. Other research also addressed directional and spatial variations of EDFs [95, 59] or non-standardised quantities derived from them [96, 97].

The inhomogeneity of decaying sound fields can be predicted with different models, taking into account the source-receiver distance [98, 60, 59, 99] or the receiver-aperture distance in coupled rooms [59]. Publication IV proposed the common-slope model of late reverberation, a descriptive model for inhomogeneous and anisotropic sound-energy decays. In contrast to the previously described models, the common-slope model does not aim to predict inhomogeneity and anisotropy from source-receiver configurations. Instead, it uses a dataset containing multiple EDFs from the same environment to determine a set of common decay slopes, and their amplitude variations describe how the sound-energy decay varies spatially and
Sound-energy decay in enclosed spaces
directionally. Consequently, Equation (2.3) becomes

\[
d_{k,cs}(x, t) = \left( \frac{N_{0,x}}{\Psi_{0,cs}(t)} \Psi_{0,cs}(t) + \sum_{k=1}^{\kappa} A_{k,x} \left[ \Psi_{k,cs}(t) - \Psi_{k,cs}(L) \right] \right),
\]

where the decay kernel \( \Psi_{k,cs}(t) \) is given by

\[
\Psi_{k,cs}(t) = \begin{cases} 
L - t, & \text{if } k = 0 \\
\exp\left( -\frac{13.8}{T_k} t \right), & \text{if } k > 0 
\end{cases}
\]

The decay kernel does not depend on the source-receiver configuration \( x \) anymore, because the decay times \( T_k \) are spatially and directionally invariant. In other words, the inhomogeneity and anisotropy of the energy decay is fully described in terms of the decay amplitudes \( A_{k,x} \), and \( N_{0,x} \) allows for describing localised noise sources. This property of the common-slope model was inspired by the common-acoustical-pole and residue (CAPR) model proposed by Haneda et al. [100], which exploits that room-mode decay times are independent of the source-receiver configuration. While the CAPR model assumes common modes, the common-slope model assumes common EDF slopes. Therefore, the common-slope model enables a compact representation of inhomogeneous and anisotropic energy decay, and can be used to obtain interpretable room-acoustic analyses.

The common-slope model is also valid for dynamic environments, i.e., those where either the sound source or the receiver moves. Position or orientation changes of the source or the receiver are modelled as varying decay amplitudes \( A_{k,x} \). Furthermore, the model can also be applied to environments consisting of multiple incoherent sound sources. In such cases, the energy contributions of all sources add up incoherently, and the corresponding decay amplitudes \( A_{k,x} \) are summed analogously.
3. Data-driven room-acoustic modelling

The previous chapter illustrated the intricacies of room-acoustic modelling and sound-energy-decay analysis, particularly when the simplifying assumptions of single-rate decays or diffuse sound fields are no longer applicable. Despite the well-researched theoretical base outlined in the previous section, data-driven methods may help to refine, explain, and interpret theory [10]. Therefore, the work presented in this thesis used data-driven methods to advance room-acoustic modelling.

This chapter elaborates how data-driven methods can be used for room-acoustic modelling. The first section describes room-acoustic data and addresses difficulties associated with collecting the large amounts required for data-driven approaches. The subsequent section addresses machine learning in room acoustics, covering popular loss functions and architectures. Lastly, several applications of data-driven methods in room acoustics will be outlined.

3.1 Room-acoustic data

Data-driven methods analyse and model data to identify underlying patterns and relationships, interpret these findings, and possibly answer a given research question [101, 10]. In the context of room acoustics, the underlying data is fairly high-dimensional. With movable sound sources and receivers, RIRs have ten degrees of freedom: six for the source and receiver position, and four for the DOD and DOA, as outlined in Section 2.1. A dataset covering each degree of freedom with a fine resolution may quickly exceed the budget regarding data storage and acquisition effort.

For example, when considering the task of sound-field reconstruction, the required number of RIRs can be determined from sampling theory. Analogously to the Nyquist-Shannon sampling theorem [102], sound fields can be reconstructed from discrete measurements if the spatial sampling is sufficiently dense [103]. More precisely, the spatial sampling frequency must be at least twice the spatial bandwidth of the sound field [103].
For three-dimensional sampling using an omnidirectional source and receiver, the number of RIRs consequently grows polynomially with the spatial bandwidth to the power of three. Therefore, sound-field reconstruction over large volumes can be difficult with respect to data acquisition and storage. However, several advanced reconstruction approaches have been proposed over the past years, offering means to reduce the number of required RIRs [104, 105, 106, 107, 108, 109, 110, 111, 112]. Such approaches typically use compressed sensing [104, 105, 106], deep neural networks [107, 109, 111, 112], or more traditional machine-learning methods like Gaussian process (GP) regression or kernel-ridge regression [108, 110]. In GP regression, scalar fields and their spatial correlation are modelled in a probabilistic way. This modelling paradigm enables data-driven interpolation and extrapolation from few data points, simultaneously quantifying the uncertainty of the resulting estimation [113]. In addition to its application in sound-field reconstruction, GP regression can also be used in automatic data-acquisition systems to progressively merge measured data into a model of the environment. This approach was also pursued in Publication II, which will be described more thoroughly in the following section.

Data-driven applications typically require diverse datasets to generalise to various room-acoustic conditions. The required degree of diversity depends on the target application and the desired generalisation power. A plethora of measured and simulated room-acoustic datasets have been published over the past years. In addition to varied source-receiver configurations [114, 115, 116, 117, 118, 119, 120, 18, 121, 122, 123, 124, 125], datasets can also cover different room shapes and sizes [114, 116, 126, 117, 121, 122, 123, 125]. Datasets with diverse decay characteristics can also be collected in specialised rooms with variable acoustics, avoiding the effort of moving the measurement setup between rooms [115, 118, 127, 124]. The complexity of the environment can also be influenced by adding, moving, or removing scattering objects like furniture [18] or humans [125], and by coupling multiple rooms or sub-volumes [121, 122, 123].

3.2 Large-scale data acquisition

The previous remarks highlight the importance of large-scale room-acoustic datasets containing numerous RIRs. Room-acoustic measurements can be time-consuming and tedious, thus warranting the acceleration of the data-acquisition process. Additionally, accelerated data acquisition also benefits room-acoustic analysis of historical sites, where access and measurement permissions are often given only for limited time spans. Automated measurements and room-acoustic simulations are two approaches for quickly collecting large quantities of room-acoustic data that will be covered in the
Room-acoustic measurements can be automated using robots. Previous studies can be roughly divided into those that measure on a microscopic scale, densely covering the close vicinity to a point of interest [105, 128, 120, 112], and those that operate on a macroscopic scale, capturing the sound field over a larger part of the environment [13, 129, 130, 131, 132, 133]. Measurements on the microscopic scale typically yield dense RIR grids that have been used to reconstruct sound fields [105, 112], study isotropy [128], or evaluate array-processing techniques [120]. Measurements on the macroscopic scale can be used to study the sound field over extended areas [13, 129, 132], study acoustic materials at various positions in the room [133], or collect diverse datasets for speech processing or spatial audio applications [131, 130].

Many of the previously mentioned robot systems rely on fixed structures like robotic arms or trusses. However, data-driven approaches relying on diverse datasets motivate mobile robot systems that can quickly be set up in different rooms without installing large amounts of additional hardware. Inspired by the mobile systems described in [131, 132, 133], Publication I proposed a system containing two movable robots, one acting as a source and the other as a receiver. The robots move guided by a random-walk algorithm rather than executing a specified set of measurements. Using grid-based approaches to measure predefined source-receiver configurations may seem attractive, but manual grid and path planning can become complicated and time-consuming in irregular geometries or rooms with many obstacles.

In a random walk, each step involves moving with a random step size into a random direction [134]. Random walks have been extensively studied, enabling the analytical derivation of quantitative measures that describe path properties for scenarios with limited complexity [134, 135, 136]. For example, the cover time of a grid with \( N \) source-receiver-configurations is defined as the time required to measure all \( N \) configurations. Its lower bound is \( N \log(N) \) [136], indicating that random walks require more steps to achieve a target coverage than grid-based approaches. Publication I explored this relationship by investigating the cover time in a shoebox room when using a random-walk algorithm to control the robots.

Although Publication I showed promising results for shoebox rooms, realistic scenes are typically more complex. As an alternative to the random-walk approach, Publication II proposed using a more advanced path planning method known as rapidly-exploring random trees (RRTs). RRTs were proposed as a tool to quickly explore high-dimensional spaces including obstacles [137] and have since then found extensive use in robotic motion planning and information gathering [138, 139]. The trees uniformly cover a search space and preferably visit areas that have not been explored before [137, 140, 141], thus making them attractive for gathering diverse
Data-driven room-acoustic modelling

room-acoustic datasets. Publication II demonstrated their suitability for autonomous room-acoustic data acquisition in a multi-room environment, and compared them with the random-walk algorithm used in Publication I. The evaluation showed that the RRT-based strategy quickly explores the environment, outperforming the random-walk algorithm with respect to the number of steps required to cover the entire scene. Another recent publications also compared the RRT-based strategy against further acoustic and perceptual path- and grid-planning approaches [142].

While autonomous room-acoustic measurements with robots considerably accelerate the data acquisition process, simulations can be an alternative approach to quickly gather large amounts of room-acoustic data. As data-driven and learning-based methods are growing in popularity, multiple recent studies generated large-scale room-acoustic datasets using simulations [143, 122, 123, 144, 145, 146]. Although dataset generation with simulations is flexible and quick, the resulting RIRs can deviate significantly from measured data [22]. Simulation accuracy must be considered when generating large-scale room-acoustic datasets, because it influences the performance of deep-learning approaches relying on simulated data [147]. Furthermore, informed sampling and data acquisition techniques may still be beneficial, because simulations can also be time-consuming or costly.

3.3 Machine learning in room acoustics

Similar to other scientific disciplines, machine-learning approaches have also become ubiquitous tools in acoustics research [23, 24]. Machine learning describes the ability of computers to extract valuable information from raw data, find inherent patterns, and improve on a performance metric related to a given task [148, 149]. The following section gives a brief overview of several machine-learning components, namely input features, loss functions, and commonly used architectures.

Although being applicable to general machine-learning approaches, the exposition is centred around deep learning, i.e. machine learning using deep neural networks with many layers [150, 151]. Deep neural networks typically contain several linear layers combined with non-linear activation functions. Their performance is evaluated using loss functions, and optimised using stochastic gradient descent methods [150, 151]. The components of deep neural network are differentiable, facilitating the estimation of gradients via backpropagation [150, 151].

As deep learning has become a very broad field, this section focuses on applications in room-acoustics research and concepts that are relevant to the publications in this thesis.
3.3.1 Input features

Choosing input features, i.e. an appropriate input data format, is an important first step in machine-learning approaches. The algorithm can only exploit its full potential if the input data is in a suitable format and contains all the information that is required for the task. Acoustic input data can be presented to machine-learning algorithms in various ways, and their suitability depends on the specific use case.

Raw waveforms are the most straightforward format of presenting audio data to machine-learning algorithms, and several successful studies used this approach [152, 153, 154, 155, 156, 157, 158]. While the high dimensionality makes waveforms computationally too expensive as input features for classical machine-learning approaches, deep neural networks with appropriate architectures can leverage the full potential of this representation. Feeding raw waveforms to deep neural networks can be beneficial because it avoids lossy feature extraction and captures audio features both on the small and large time scale.

Alternatively, input features can be derived from transformations applied to the audio signal, for instance into the frequency, time-frequency, or wavenumber domain. Deep-learning approaches commonly use time-frequency representations like spectrograms as input features, because they can be represented as images. This image-like format allows for efficient processing using two-dimensional convolutional layers, which are well understood and widely used within the realm of image processing [159, 160, 161, 162, 163]. In acoustic applications, filter banks inspired by the auditory system can help to reduce the complexity of the neural networks by producing spectrograms with low spectral resolution [164]. Such spectrograms have successfully been used for room-acoustic tasks, such as blindly estimating room volume, room dimensions, or room-acoustic parameters [164, 165, 166, 167].

Publication III used neither of the previously described approaches. Instead, EDFs obtained from the Schroeder backwards integration procedure (see Section 2.2) were used as the input to the proposed neural network. EDFs are characterised by their smoothness and minimal high-frequency content. In contrast to the raw RIR waveforms or their time-frequency domain representations, a representation in terms of EDFs enables significant downsampling to lower the computational complexity of decay analysis methods. Therefore, EDFs were downsampled to 100 samples before feeding them to the neural network, making the resulting architecture lightweight and efficient.

In addition to processing audio signals, some room-acoustic tasks also require incorporating room geometry as input features. The complexity of three-dimensional (3D) room models can escalate rapidly for meshes with many surfaces. Moreover, modelling efforts typically extend beyond...
pure geometry to also include acoustic material properties, such as absorption characteristics. Various studies aimed at providing neural networks with room-geometry data, for instance, by utilising floor plans [144, 168], projecting the absorptive surfaces onto polyhedral feature maps [169], simplifying the mesh [170], or representing distances from a receiver to the closest surfaces as images [171].

In applications, where the room-acoustic effects are investigated over larger spatial extents, images of the sound pressure field can be used as inputs to neural networks [107, 145]. This input representation also allows leveraging tools from deep-learning-based image processing, such as convolutional layers.

### 3.3.2 Loss functions

Supervised machine-learning approaches calculate performance metrics based on their output data, and use them to improve on a given task. These performance metrics are also called loss functions, and they can be set up in various ways. Selecting the appropriate loss function requires identifying whether the problem should be treated as a classification or a regression task. In classification tasks, the goal is to select one or more classes based on the input, whereas in regression tasks aim to predict a continuous quantity.

In classification tasks, the output of machine-learning models can be represented as probabilities associated with predicting the different classes. A popular classification loss is the cross-entropy loss, which measures the discrepancy between predicted probabilities and actual class labels [150, 151].

Regression tasks in the field of room acoustics typically either output an audio signal or a room-acoustic parameter. The mean squared error (MSE) or mean absolute error (MAE) are commonly used to assess prediction accuracy of these outputs. When evaluating audio signals, loss functions operate in the time domain [152, 153, 170, 158] or frequency domain [172, 173, 156, 174]. Time-domain losses can be problematic, because they result in large errors for delayed or inverted signals, although such alterations do not affect the perceived similarity. Spectral losses can circumvent these issues, and hybrid time-frequency losses are widely used across applications. Such time-frequency losses benefit from simultaneously considering multiple resolutions to address the trade-off between time and frequency resolution [172, 173, 156, 174].

In some instances, regression losses are directly computed from the predicted values. Such losses have found application in room-acoustic parameter estimation [164, 169, 167], room-dimension and volume estimation [165, 154, 166, 175], boundary-condition inference [154, 155, 175], eigenfrequency estimation [144], in situ acoustic material property measurements [145], sound-field reconstruction [107, 112], or sound-field esti-
Data-driven room-acoustic modelling

More recently, several room-acoustic studies also incorporated additional domain knowledge into the losses to train deep neural networks. For instance, physics-informed neural networks (PINNs) include a loss term that quantifies how much the predicted data deviates from physical behaviour given by a governing differential equation [176], for instance the wave equation [177]. PINNs have successfully been used to predict the sound field by integrating initial conditions and boundary conditions given by the sound source and room surfaces, respectively [178]. They also found application in sound-field reconstruction [112, 179] and nearfield acoustic holography [180]. While the loss function in PINNs determines whether the predicted data satisfies the governing differential equation, deep operator networks (DeepONets) aim at directly learning the differential operator of the wave equation by evaluating whether the network processes input functions similarly as the operator [181]. DeepONets have successfully been used to estimate sound propagation in dynamic 3D scenes with complex geometry [182]. DeepONets belong to the larger class of neural operators, another example of which are Fourier neural operators, which have also found application in solving the acoustic wave equation [183].

In room-acoustic applications that rely on correctly estimating the decay characteristics of a room, incorporating a loss term that considers the full energy decay through EDFs can be advantageous [170, 174, 158]. Publication III also includes a loss computed directly from EDFs, rather than only relying on a loss derived from the parameters of the underlying exponential model. In addition to the EDF loss, the training includes a model order loss and a loss calculated from the estimated background noise parameter. This loss combination proved more effective in training a neural network for the decay analysis task than a single loss calculated from the model parameters, and it was chosen to make the network focus on accurately fitting the EDF rather than precisely matching decay parameters. The multi-exponential decay analysis problem is inherently very ill-conditioned [184]. Measurement inaccuracies can cause two models with significantly different decay times, decay amplitudes, and model orders to produce numerically equivalent fits to a given measurement. Generally, simpler models are favoured over more complex ones, and the use of the additional model-order loss encouraged the network to predict the correct number of slopes.

3.3.3 Commonly used architectures

Deep-learning methods are widely used in many scientific disciplines these days, prompting swift development and advancement of various neural network architectures. Generally, a neural network aims to approximate
Data-driven room-acoustic modelling

![Diagram of different commonly used deep-learning components.](image)

(a) Linear layer  
(b) ReLU activation  
(c) Multilayer perceptron  
(d) 1D convolutional layer

**Figure 3.1.** Different commonly used deep-learning components.

A function that maps a set of inputs to a set of outputs through multiple operations [150, 151].

A simple, yet powerful, neural network architecture for achieving this goal is called multilayer perceptron (MLP). MLPs contain alternating sequences of linear layers and nonlinear activation functions, followed by a final linear layer called output layer [150, 151]. Figure 3.1a depicts a linear layer that processes $N = 2$ input values using $M = 3$ neurons. A neuron is a unit that computes the following linear mapping [150, 151]

$$y = g(x) = \left( \sum_{i=0}^{N-1} w_i x_i \right) + b$$

(3.1)

from inputs $x$ to output $y$. The weights $w_i$ and the bias term $b$ are learnable parameters of the neuron that get adjusted during the neural network training. Each neuron features $N$ weights and 1 bias term. Consequently, a linear layer with $M$ neurons contains $M(N + 1)$ learnable parameters.

If neural networks only contained neurons, they would have limited use as function approximators, because they could not be used for functions other than the linear mappings given by Equation (3.1). However, by combining them with nonlinear activation functions, they become universal
Data-driven room-acoustic modelling

approximators, capable of approximating any function with arbitrarily small error as long as the linear layer contains enough neurons [150, 151]. Activation functions can have various shapes [185], and the rectified linear unit (ReLU) is among the most popular ones. Figure 3.1b depicts the ReLU activation function, given by [186]

$$\sigma_{\text{ReLU}}(x) = \max(0, x).$$  \hspace{1cm} (3.2)

Figure 3.1c shows an example MLP containing a single linear layer followed by a ReLU activation and an output layer. The layers are fully connected, each ReLU-processed linear layer output being an input to each neuron of the following layer. In the depicted scenario, there is only one layer following the linear layer, the output layer. In other words, the $M^{(1)}$ outputs of the linear layer followed by the ReLU activation are given by

$$y_j^{(0)} = \sigma_{\text{ReLU}} \left[ \sum_{i=0}^{N-1} w_{i,j}^{(0)} x_i + b_j^{(0)} \right], \quad \text{with} \quad 0 \leq j \leq M^{(0)} - 1,$$

and the outputs of the output layer are given by

$$y_j^{(\text{out})} = \left( \sum_{i=0}^{M^{(0)}-1} w_{i,j}^{(\text{out})} y_i^{(0)} \right) + b_j^{(\text{out})}, \quad \text{with} \quad 0 \leq j \leq M^{(\text{out})} - 1.$$  \hspace{1cm} (3.4)

As indicated by the superscripts, each linear layer has its own set of learnable weights and biases.

Although MLPs are simple and flexible function approximators, they can easily become impractically complex due to the large number of weights in linear layers. For example, applications concerned with audio signals or images typically process inputs with thousands or millions of values. An MLP with only few layers will feature several times as many parameters, even if the layers contain only a limited number of neurons. Additionally, MLPs treat each input feature independently, making it difficult to exploit correlations across them. Convolutional neural networks (CNNs) overcome these limitations and consider the spatial structure in the input, for example, by finding repetitive patterns in consecutive audio samples or areas of an image. CNNs produce output values by moving a kernel or filter across the input and multiplying it with the corresponding input values [159, 160, 150, 151]. This mechanism ensures that a pattern appearing in multiple locations of the input will generate identical output values at all relevant positions, independent of the pattern locations in the input. This concept is known as spatial invariance.

Figure 3.1d shows an example of a one-dimensional (1D) convolutional layer and demonstrates the calculation of its output. The output $y_j$ of a 1D convolutional layer is calculated using

$$y_j = \left( \sum_{\gamma=0}^{\Gamma-1} x_{j+\gamma} w_\gamma \right) + b,$$

\hspace{1cm} (3.5)
where $\Gamma$ is the length of the convolution kernel. For a convolutional layer with $N$ inputs, the resulting output has $N - \Gamma + 1$ elements. However, convolutional layers can also move multiple steps at a time (strided convolution), or process only every other element (dilated convolution), resulting in different output sizes [150, 151]. Equation (3.5) shows that the bias term $b$ does not depend on the input element, which is required to achieve spatial invariance. The number of learnable parameters is reduced considerably compared to MLPs and only amounts to $\Gamma + 1$ per convolution kernel, regardless of the input size. Convolutional layers can be extended to two or three dimensions, making them applicable to higher-dimensional input data such as images.

Convolutional layers have found many applications in room acoustics, for instance, in sound-field estimation and reconstruction [107, 171], eigenfrequency estimation [144], room-acoustic-parameter estimation [164, 154, 167], blind room-impulse-response estimation [156, 158], room-boundary-condition- and volume estimation [165, 155, 166, 175]. Furthermore, Publication III proposes a CNN-based architecture for the multi-exponential decay-analysis task.

The neural network outlined in Publication III consists of three 1D convolutional layers followed by three fully-connected linear layers and separate heads for all predicted parameters. Each head comprises two fully-connected linear layers. A ReLU activation function succeeds each layer, except for the output layer. The initial convolutional layers reduce the number of learnable parameters, focusing on the decay behaviour across multiple samples rather than the small-scale variations due to noise. The proposed network achieved a similar decay-analysis performance as a state-of-the-art Bayesian analysis approach, while being lightweight and processing several thousand EDFs in under a minute. This efficiency is crucial for making decay analysis available on mobile devices, such as head-mounted displays or smart phones. It is also beneficial for analysing large datasets, where systems with GPUs can leverage the full potential of the CNN-based architecture.

### 3.4 Applications to room-acoustic rendering

The previous section mentioned several room-acoustics studies relying on deep-learning architectures. Typical applications of data-driven room-acoustics methods include acoustic-material- and sound-field analysis, sound-field visualisation, sound-field control, and efficient room-acoustic rendering for VR/AR applications and computer games. This section focuses on the latter, as it is directly related to Publication V.

Scenes in VR/AR applications and computer games often include complex room geometries with non-uniform absorption and interconnected
rooms. These intricacies complicate efficient simulation in audio engines due to the reverberation being inhomogeneous and anisotropic. Several approaches for rendering such environments have been proposed. Parametric rendering methods follow a two-step process. First, they simulate the sound field with conventional methods, such as wave-based solvers. Subsequently, they extract various parameters from the simulated results and use the determined parameters to efficiently render the scene in a perceptually convincing way [187, 188, 189]. The first step can be bypassed through the use of simple room-acoustic models that identify relevant parameters at runtime [70], or by manually setting them according to the desired perceptual sensation [190, 191]. Instead of extracting parameters, simulations from the first step can also be used to train neural networks to efficiently interpolate to denser grids at runtime [178, 192, 193, 182, 112]. In scenarios, where computational complexity and storage demands are less critical, interpolation-based rendering techniques can be utilised [194, 195, 196, 197, 198]. Perceptual and room-acoustic considerations can significantly reduce the number of required responses in such approaches either by implementing informed interpolation and extrapolation methods [196, 198] or by using non-uniform simulation or measurement grids [199, 195].

The system proposed in Publication V follows a parametric rendering paradigm based on the common-slope model of late reverberation outlined in Publication IV. It can render inhomogeneous and anisotropic late reverberation for dynamic sources and receivers, using a set of exponentially decaying reverberators. Although the decay times of the reverberators remain fixed for all source-receiver configurations, inhomogeneity and anisotropy can be achieved by summing the reverberator outputs with gains that vary with position, direction, and frequency. The decay times and gains of the reverberators are determined from a collection of RIRs by adopting the common-slope model.
4. Summary of contributions

This chapter summarises the main contributions of the publications contained in this thesis.

**Publication I – “Autonomous robot twin system for room acoustic measurements”**

**Demo videos and a reference implementation are available at:**
http://research.spa.aalto.fi/publications/papers/artsram/
https://github.com/georg-goetz/ARTSRAM

Publication I proposes an autonomous robot system for room-acoustic measurements containing two mobile robots. One robot is equipped with a portable loudspeaker, while the other robot carries a microphone array, allowing for RIR measurements between them. After elaborating on the design constraints of a practical system, the study presents an example implementation. The outlined system can quickly measure large quantities of position-dependent RIRs at room scale, while being flexible, versatile, and cost-efficient. By employing the proposed random-walk procedure, the system can be used without prior manual path or grid planning. Publication I evaluates the random walk using a test measurement and simulations, showing that the random-walk procedure requires more steps than a grid-based approach. However, after a short time, the resulting measurements also cover the investigated shoebox environment densely with diverse source-receiver configurations.

**Publication II – “Autonomous room acoustic measurements using rapidly-exploring random trees and Gaussian processes”**

Publication II builds upon Publication I and introduces a new path-planning strategy based on RRTs and GP regression. The RRT-based strategy enables the robots to navigate complex geometries featuring multiple...
connected rooms, without requiring floor plans or predefined grids. Additionally, the system uses GP regression to progressively integrate new measurement data into an evolving model of the environment. After describing an example implementation of the robot system, the study evaluates the measurement strategy in an apartment-like environment consisting of three connected rooms. In this publication, the system was used to obtain maps of a late-reverberation quantity, namely, the common-slope amplitudes introduced in Publication IV. The evaluation indicates that the RRT-based path-planning strategy facilitates a quick exploration of all rooms, whereas the random-walk procedure proposed in Publication I had difficulties transitioning between rooms. GP regression performed similarly to simple interpolation strategies when modelling common-slope amplitudes, while being able to handle noisy observations and providing uncertainty estimates for the modelled value.

**Publication III – “Neural network for multi-exponential sound energy decay analysis”**

A toolbox including the pre-trained network and a reference implementation of the network training and dataset generation is available at: https://github.com/georg-goetz/DecayFitNet

Publication III proposes the DecayFitNet, a neural network that estimates the multi-exponential decay parameters [see Eq. (2.2)] from EDFs. Its architecture contains several 1D convolutional layers, followed by a multi-head MLP. A preprocessing step resamples input EDFs to a fixed length. This step enables the network to analyse inputs that have different sampling frequencies without requiring retraining. Since the network has been trained entirely on synthetic data, the study explores how to create such datasets for the investigated task. The evaluation demonstrates the performance of the network on two large-scale datasets containing measured EDFs, comparing it against a state-of-the-art Bayesian decay-analysis approach. The results show that both approaches robustly analyse multi-exponential decays on large datasets. Furthermore, the evaluation illustrates that the DecayFitNet performs just as well as the Bayesian approach, while being lightweight, computationally efficient, and deterministic during inference.
Publication IV – “Common-slope modeling of late reverberation”
A reference implementation of the common-slope analysis is available at:
https://github.com/georg-goetz/CommonSlopeAnalysis

Publication IV introduces the common-slope model, a compact representation of anisotropic and inhomogeneous late reverberation. Its fundamental idea is to model late reverberation as linear combinations of exponential decays with spatially and directionally invariant decay times called common slopes. Consequently, anisotropy and inhomogeneity are achieved by varying decay amplitudes. The study explores several approaches to determine the common slopes, identifying the k-means clustering method as the most general one. Finally, the common-slope model is evaluated on several datasets, showing that it can model inhomogeneous and anisotropic sound-energy decay, while requiring fewer parameters than the traditional multi-exponential model. The resulting amplitude maps offer interpretable insights on the spatial and directional late-reverberation variations.

Publication V – “Dynamic late reverberation rendering using the common-slope model”
Sound examples are available at:

Publication V presents a new late-reverberation rendering approach based on the common-slope model introduced in Publication IV. This new approach benefits from the compact late-reverberation representation enabled by the common-slope model, allowing the rendering system to operate with a minimal number of reverberators. Their frequency-dependent gains are adjusted according to the position and orientation of the source and the listener. The evaluation considers a multi-room environment that contains three interconnected rooms, each featuring considerably different absorption materials. It illustrates the decay amplitudes required for late-reverberation rendering within the investigated environment, highlighting the frequency-dependent inhomogeneous sound-energy decay.
5. Conclusions and future work

Overall, the work on this thesis was characterised by the continuous interaction between room acoustic theory and data-driven methods. The autonomous measurement robots described in Publications I and II aimed to provide the large and diverse datasets required by data-driven methods. The evaluations of both systems showed that room-acoustic measurements can be automated with robots and that the gathered data can be used to analyse and describe sound fields. Navigating unknown environments without prior path or grid planning can be challenging, but the proposed measurement strategies achieved promising results. Furthermore, Publication II showed that knowledge about sound fields in rooms can guide the utilised measurement strategies and prevent the excessive use of resources on data acquisition.

Room-acoustic theory on multi-exponential sound-energy decay in inhomogeneous and anisotropic sound fields supported the development of the room-acoustic analysis tools and models introduced in Publications III and IV. Publication III showed that neural networks can be used for multi-exponential decay analysis. They perform similarly to iterative state-of-the-art approaches, while being lightweight and deterministic during inference. Setting up a synthetic training dataset for this task was not straightforward, but the evaluation showed that the network performs well on large-scale datasets and generalizes to measured EDFs. Publication IV introduced the common-slope model, showing that the number of parameters to describe inhomogeneous and anisotropic late reverberation can be considerably reduced when assuming decay slopes with spatially and directionally invariant decay rates. The common-slope-model parameters can be determined from RIR sets of the environment. It was shown that averaging RIRs or RIR parameters may not yield the desired results, whereas clustering of decay times achieved a good fitting performance. Therefore, the reduced parameter set might be sufficient for many applications, including the late-reverberation rendering system described in Publication V.

Finally, the new insights gained in Publication IV prompted the work on
Conclusions and future work

data-driven late-reverberation rendering outlined in Publication V. The study showed that the reduced parameter set of the common-slope model is beneficial for an efficient late-reverberation rendering system. Instead of using one reverberator for each source-receiver configuration, the proposed approach only requires one reverberator per decay slope of the common-slope model. Therefore, the computational complexity can be reduced, while ensuring a convincing perceptual experience, as demonstrated by the provided audio examples.

Publications I and II showed that room-acoustic measurement robots can be used for automatic data acquisition. Their applicability was also demonstrated in more complex geometries involving multiple connected rooms. However, even advanced measurement strategies cannot bypass weaknesses of the underlying robot system. Commercially available mobile robots can be divided into two categories. The first category includes small and simple robots equipped with wheels, typically lacking mobility on uneven terrain. The second category includes larger quadruped robots. Despite featuring advanced movement capabilities, they can be bulky, hampering their ability to fit through narrow passages within environments. With improved robot systems becoming available in the future, automated room-acoustic measurements will also become more practical and flexible. Additionally, measurement systems involving drones could be a promising future avenue of research, provided that self-noise can be minimised.

The work on data-driven late-reverberation rendering presented in Publication V showed promising first results. So far, the required parameters have been extracted from measured or simulated responses in an offline step preceding the rendering. In the future, methods to estimate the parameters directly from the room geometry should be investigated. Machine-learning methods can be a suitable tool for this task.

Finally, future work should further investigate how room-acoustic modelling can leverage the full potential of data-driven methods. For instance, generalisation of data-driven approaches remains an open challenge. Despite their frequent use by the acoustic research community, publicly available datasets are often still tailored to specific use cases. Machine-learning models trained on such datasets may face difficulties when applied to unseen data. A diverse room-acoustic dataset must feature sufficient variability in many degrees of freedom. It should include a variety of source-receiver-configurations, room shapes, room sizes, acoustic materials, and other complex features like furniture and connected rooms. While the degrees of freedom are well understood, the required resolution along each dimension is less clear. A full factorial design of all dataset features is certainly not practical, and probably not required for satisfactory results, either. Therefore, establishing a set of guidelines towards creating flexible, general, and compact room-acoustic datasets will be an interesting future research topic.


References


References


signal processing,” in Proc. Int. Conf. Learn. Represent. (ICLR), (Online
conference), 2020.

visual learning of environment acoustics,” in Proc. 36th Conf. Adv. Neural

condition inference using room transfer functions,” in Proc. 10th Conv.
European Acoust. Assoc. (Forum Acusticum), (Turin, Italy), pp. 3149–3153,
2023.

networks: A deep learning framework for solving forward and inverse

informed neural networks (PINNs) for wave propagation and full waveform

neural networks for one-dimensional sound field predictions with parameter-
ized sources and impedance boundaries,” JASA Express Lett., vol. 1,
no. 12, article no. 122402, 2021.

[179] K. Shigemi, S. Koyama, T. Nakamura, and H. Saruwatari, “Physics-
informed convolutional neural network with bicubic spline interpolation
for sound field estimation,” in Proc. 17th Int. Workshop Acoust. Signal
Enhancement (IWAENC), (Bamberg, Germany), 2022.

[180] M. Olivieri, M. Pezzoli, F. Antonacci, and A. Sarti, “A physics-informed
neural network approach for nearfield acoustic holography,” Sensors, vol. 21,
no. 23, article no. 7834, 2021.

operators via DeepONet based on the universal approximation theorem of

[182] N. Borrel-Jensen, S. Goswami, A. P. Engsig-Karup, G. E. Karniadakis,
Sci. (PNAS), vol. 121, no. 2, article no. e2312159120, 2024.

neural operator networks for solving the 2D linear acoustic wave equation,”
in Proc. 10th Conv. European Acoust. Assoc. (Forum Acusticum), (Turin,


[185] T. Szandala, “Review and comparison of commonly used activation functions
for deep neural networks,” in Bio-inspired Neurocomputing (A. K. Bhoi,
P. K. Mallick, C.-M. Liu, and V. E. Balas, eds.), pp. 203–224, Singapore,

[186] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltz-
amann machines,” in Proc. 27th Int. Conf. Mach. Learn. (ICML), (Haifa,


