Realism in virtually supervised learning for acoustic room characterization and sound source localization

Prerak Srivastava

UNIVERSITÉ DE LORRAINE

Inria

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Thèse de Doctorat

PRERAK SRIVASTAVA

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Lorraine
Mention Informatique

École doctorale:
Informatique, Automatique, Electronique-Electrotechnique, Mathématiques et Sciences de L'architecture

Unité de recherche:
Laboratoire Lorraine de Recherche en Informatique et ses Applications
UMR 7503

Soutenue le 13 novembre 2023
Thèse N°:

Réalisme dans l'apprentissage virtuellement supervisé pour la caractérisation acoustique des salles et la localisation de sources

Composition Du Jury

Rapporteur :	**Rainer Martin**, Professeur, Ruhr-Universität Bochum, Allemagne
Rapporteur :	**Éric Bavu**, Maître de conférences, CNAM Paris, France
Président du jury :	**Marie-Odile Berger**, Directrice de recherche, Centre Inria de l'Univ. de Lorraine, France
Examinateur :	**Simon Leglaive**, Maître de conférences, CentraleSupelec Rennes, France
Directeur de thèse :	**Emmanuel Vincent**, Directeur de recherche, Centre Inria de l'Univ. de Lorraine, France
Co-directeur de thèse :	**Antoine Deleforge**, Chargé de recherche, Centre Inria de l'Univ. de Lorraine, France

Septembre 2023
Nancy, France

UNIVERSITÉ DE LORRAINE

Inria

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Doctoral Thesis

PRERAK SRIVASTAVA

Dissertation presented in order to obtain the
Doctoral Degree from the University of Lorraine
Computer Science

Doctoral School:
Computer Science, Automation, Electronics, Mathematics and Architectural Sciences (IAEM)

Research Unit:
Lorraine Laboratory for Research in Computer Science and its Applications
UMR 7503

Defended on November 13, 2023
Thesis N°:

Realism in Virtually Supervised Learning For Room Parameter Estimation And Sound Source Localization

Jury Members

Rapporteur :	**Rainer Martin**, Professor, Ruhr-Universitat Bochum, Germany
Rapporteur :	**Eric Bavu**, Associate Professor, CNAM Paris, France
Examiner :	**Marie-Odile Berger**, Senior Research Scientist, Centre Inria de l'Univ. de Lorraine, France
Examiner :	**Simon Leglaive**, Assistant Professor, CentraleSupelec Rennes, France
PhD supervisor :	**Emmanuel Vincent**, Senior Research Scientist, Centre Inria de l'Univ. de Lorraine, France
PhD co-supervisor:	**Antoine Deleforge**, Research Scientist, Centre Inria de l'Univ. de Lorraine, France

September 2023
Nancy, France

Dedicated to my loving family ...

Acknowledgements

Emmanuel and Antoine, I am grateful to have received a chance to work with you on this thesis. I learned a lot from both of you in terms of technical and non-technical aspects and this learning will have a great influence on my life and career. Throughout this work, I received a tremendous amount of support and warmth accompanied by a good level of understanding which helped me sustain this PhD thesis. The amount of appreciation that I can give will always be short for countless hours of wise discussion, whiteboard sessions, feedback, and guidance that were generously offered by you.

Emmanuel, I've always been in awe of your intellect and hard work. The way you manage multiple important assignments at the same time was a source of inspiration for me. Your observations were consistently insightful and accurate, and your coherence of ideas was admirable. I feel honored to have had the opportunity to work under your guidance and learn from you.

Antoine, your kindness and empathy have been a continual source of motivation and ease. The moment I got stuck on any problem related to signal processing, you were always enthusiastic to help with whiteboard sessions. I always waited for these sessions as they were full of knowledge packed with your unique style of teaching. I will always cherish those whiteboard sessions and your passion for solving math problems. You have supported me at every turn, giving me the freedom, trust, and encouragement I needed to finish this difficult academic project.

So I thank both of you for the lessons that I have learned in different aspects of research work, I hope to apply them throughout my research career.

My jury members: Rainer, Eric, Marie-Odile and Simon I thank you for the time and effort that you dedicated to reviewing my thesis, and for the thoughtful criticism and recommendations you made during the defense. Your insightful criticism has forced me to reflect carefully on my writing and strengthen my arguments. I find inspiration and motivation in your knowledge, discernment, and dedication to perfection. Having long appreciated your contributions to the field, I consider myself fortunate to have had the chance to have you on my jury.

I want to extend my gratitude to the Multispeech team. The team has the best work environment, with understanding and supportive peers. Everybody on the team is always ready to help and collaborate on any research problem, I have had great conversations with my colleagues and learned a lot from them on variety of research areas. The team was always evolving with young researchers and in between this ever-evolving team I found true-long lasting friendships. Thanks for all the shared memories, I will always cherish them and will

have a special place in my heart.

My sincere gratitude is extended to : Sewade Ogun, Marina Krémé, Can Cui, Sandipana Dowerah, Louis Abel, Michel Olvera, Marie-Anne Lacroix, Vinicius Ribeiro, Nicolas Zampieri, Nasser Monir, Sofiane Azzouz, Hélène Zganic, Nicolas Furnon, Joris Cosentino, Théo Biasutto-Lervat, Mickaella Verdon, Shakeel Sheikh, Manuel Pariente, Nicolas Turpault, Tulika Bose, Pierre Champion, Seyed Hosseini, Romain Serizel, Paul Magron, Louis Delebecque, Hugo Bergerat, Ashwin Geet D'Sa, Ali Golmakani, Soklay Heng, Soklong Him, Louis Lalay, Louis Bahrman, Colleen Beaumard, Berne Nortier, Antoine Bruez, Jean-Eudes Ayilo, Robin San Roman, Tom Sprunck, Stéphane Dilungana, Vincent Colotte, Georgios Zervakis, Ajinkya Kulkarni, Mostafa Sadeghi, Slim Ouni, Emmanuelle Deschamps and Denis Jouvet. I would also like to thank Hugo Bergerat, Sewade Ogun, and Michel Olvera for countless table tennis sessions, those sessions improved my skills in this sport and now it is one of my favorite hobbies.

Now I would like to thank my friends outside of the Multispeech team who deserve special mention, their support and unconditional warmth was the reason I was able to complete this thesis. They are Siyana Pavlova, Kelvin Han, Vishal Kumar Porwal, Hardik Sharma, Shan Akbaraly, Maulik Shah, Oussama Hamouli, Saad Tamazart, Salim Perchy, Gabriel, Annesh Rathi, Rahul Dhaked and Atal Singh.

Additionally, I would like to thank everyone at the Loria/Inria laboratory for cultivating an inclusive culture and for realizing the benefits of diversity in the workplace. Special thanks to Isabelle from the cafeteria, whose desserts always made my day even on the lowest of days that I encountered during the past 3 years.

Lastly, I would like to mention my family for their wholehearted support in pursuing my studies in the field of my interest. Staying miles apart they still managed to keep a check on my mental and physical well-being. On certain occasions, it was a bit hard to be away from them, but those occasions made me learn a lot about myself and improved my life-tackling skills. They always took care of themselves, which is something I appreciate because it freed me up to concentrate on my research.

I express my sincere gratitude to the ANR HAIKUS project and to Université de Lorraine for their financial support, which enabled me to maintain my living. Finally, I want to thank the Grid'5000 computing infrastructure for allowing me to use their GPUs to speed up my experiments and for granting me special access when I needed it most.

Résumé

La Réalité Augmentée Audio vise à intégrer un contenu audio virtuel dans l'environnement acoustique de l'utilisateur, créant ainsi une expérience audio immersive. La disponibilité commerciale de casques de réalité augmentée tels que l'Apple Vision Pro a encore renforcé l'intérêt pour ce domaine de recherche. Pour synthétiser un son spatial binaural capable de recréer la perception de la distance, de la direction et des indices acoustiques, la connaissance des paramètres acoustiques spécifiques de l'environnement de l'utilisateur est un prérequis. Les paramètres acoustiques se divisent en deux catégories : des paramètres globaux associés à la géométrie de la pièce, au temps de réverbération et aux matériaux des parois, et des paramètres locaux concernent la localisation de chaque source sonore. À l'aide de simulateurs acoustiques, ces paramètres sont utilisés pour simuler des réponses impulsionnelles des salle. Ces réponses impulsionnelles peuvent ensuite etre convoluées avec des signaux audio bruts pour synthétiser un son spatial binaural avec une perception de réalisme. Cependant, l'estimation des paramètres acoustiques est un défi. Des recherches antérieures ont tenté de résoudre ce problème grâce à des mesures in-situ laborieuses et chronophages, souvent peu pratiques. Dans cette thèse, nous relevons ce défi en utilisant des techniques d'apprentissage automatique supervisées utilisant des enregistrements de parole en entrée. Notre principal domaine d'application concerne les pièces cuboïdes avec des scénarios acoustiques statiques. Dans la première partie de notre travail, nous développons un réseau de neurones multi-tâches pour l'estimation des paramètres de la salle. Nous évaluons ensuite sa robustesse en utilisant des données réelles.

Dans la deuxième partie, nous déplaçons notre attention vers l'apprentissage virtuellement supervisé. Cette approche consiste à entraîner des modèles d'apprentissage automatique exclusivement sur des données simulées. La justification de cette stratégie repose sur la disponibilité limitée de jeux de données réels spécifiques à la tâche dans ce domaine. Pour assurer la généralisation des modèles ainsi appris, l'ensemble d'apprentissage doit ressembler de près aux scénarios rencontrés dans les ensembles de test. Afin de combler cette lacune, nous améliorons le réalisme du simulateur acoustique open-source Pyroomacoustics en y intégrant une extension de la méthode de source image. Nous utilisons, ce simulateur acoustique amélioré pour entraîner des réseaux neuronaux aux tâches d'estimation des paramètres de la salle et de localisation des sources sonores. Nous utilisons plusieurs ensembles de test réels pour évaluer l'impact positif de l'apprentissage à l'aide du simulateur amélioré. Nos expériences montrent que la généralisation est améliorée pour les deux tâches par rapport aux modéles appris pour la même tâche avec des données d'apprentissage moins réalistes. À notre connaissance, il s'agit de l'une des premières études à explorer

l'apprentissage virtuellement supervisé pour l'estimation des paramètres acoustiques de salle à la fois globaux et locaux.

Abstract

Audio Augmented Reality aims to integrate virtual audio content into the user's acoustic environment, creating an immersive audio experience. The commercial availability of augmented reality headsets such as Apple Vision Pro has further motivated interest in this research field. To synthesize binaural spatial audio that can recreate the perception of distance, direction, and acoustic cues, the knowledge of specific acoustic parameters of the user's environment is a pre-requisite. Acoustic parameters can be divided into two categories: global parameters associated with the room's geometry, reverberation time, and wall materials, and local parameters concerning the location of each sound source. With the help of room acoustic simulators, these parameters are used to simulate room impulse responses. These room impulse responses can then be convolved with dry speech signals to synthesize binaural spatial audio with a perception of realism. However, the estimation of these acoustic parameters is a challenge. Previous research has attempted to address this problem through cumbersome and time-consuming in-situ measurements, which are often impractical.In this thesis, we tackle this challenge by leveraging supervised machine-learning techniques using speech recordings as input. Our primary focus is on cuboid rooms with static acoustic scenarios. In the initial part of our work, we develop a multi-task neural network for room parameter estimation. We then assess its robustness using real-world data.In the second part, we shift our focus towards virtually supervised learning. This approach involves training machine learning models exclusively on simulated data. The rationale behind this strategy is rooted in the limited availability of task-specific real datasets within this domain. To ensure genralization, the training dataset should closely resemble the scenarios encountered in the test datasets. In order to bridge the gap, we improve realism in the open-source room acoustics simulator Pyroomacoustics by implementing an extended image source method. Further, this improved room acoustics simulator is used to train neural networks for the tasks of room parameter estimation and sound source localization. We employ several real test datasets to assess the positive impact brought by training the systems using the improved simulator. Our experiments show that the generalization of the system is improved across both tasks when compared to the systems trained for the same task with less realistic training data. To the best of our knowledge, this is one of the first studies to explore the field of virtually supervised learning for the task of global and local room acoustic parameter estimation.

Contents

List of figures

List of tables

List of acronyms

AAR Audio Augmented Reality
CNN Convolutional neural network
DFT Discrete Fourier Transform
DSHT Discrete Spherical Harmonics Transform
DRR Direct to Reverberant Ratio
DNN Deep neural network
DOA Direction of Arrival
FDTD Finite Difference Time Domain
FEM Finite Element Method
GMM Gaussian Mixture Model
GPU Graphic Processing Unit
GCC-PHAT Generalized Cross-Correlation with Phase Transform
GAN Generative adversarial network
HATS Head and Torso Simulator
ISM Image Source Method
IPD Interchannel Phase Difference
ILD Interchannel Level Difference
IDFT inverse discrete Fourier transform
MUSIC Multiple Signal Classification
MEMS Micro Electromechanical systems
RIR Room Impulse Response
STFT short time Fourier transform
SRT Stochastic Ray Tracing
SRP-PHAT Steered Response Power with Phase Transform
SOFA Spatially Oriented Format for Acoustics

1 Introduction

1.1 Motivation

Augmented Reality (AR) seeks to integrate computer-generated virtual content with the physical environment in a way that creates a seamless fusion, making the virtual content appear as if it were part of the real world (Azuma, 1997). AR has the potential to enhance people's perception of, and interaction with, their surroundings, making it easier for them to perform real-world tasks. One significant advantage of AR technology is its capacity to augment human senses, enabling people to interact with virtual objects and scenes just as effortlessly as they do with the physical world (Azuma et al., 2001). However, the majority of AR research has predominantly concentrated on visual augmentation (Kim et al., 2018).

Audio Augmented Reality (AAR) has received less attention compared to visual augmentation. In AAR, virtual auditory content is seamlessly integrated into the real acoustic environment, enhancing the user's experience. In order to create a realistic sense of direction, distance, and reverberation for the virtual sounds, the virtual sound sources are often binaurally spatialized. AAR technology is capable of creating immersive audio experiences for a variety of sound contents such as speech, music, beacons, and alerts (Li et al., 2018). Therefore, it can convey different types of information depending on the context. Readily available powerful computing devices such as mobile phones have the ability to simulate virtual sound sources and provide an immersive audio experience. The user can access these immersive audio experiences with many off-the-shelf headphones (Yang, 2021). Recently many AR headsets and earphones have been introduced in the market, that are able to perform real-time AAR computation. These include Apple Vision pro, Apple Airpod pro, Samsung Galaxy buds pro, and Sony WH-1000XM5.

This has spurred further interest in the field. Compared to audio simulation in virtual reality systems, AAR systems are more difficult to implement due to the complex technology required to augment audio in real-world conditions. More specifically, in virtual reality audio, the use of pre-designed virtual scenes can simplify the rendering of audio content, while in AAR, creating virtual sounds in the physical world and adapting them to the user in real-time is more complex. A functional AAR system comprises five technological components: head tracking, room acoustic modeling, spatial sound synthesis, interaction technology, and display technology (Yang et al., 2022). Interaction technology is required to get the user inputs to adjust the parameters for AAR, while display technology refers to the way the sound is played back for the user. To display the visual content accompanied by

immersive audio a screen is also required. In this thesis, we only focus on one of these five aspects of AAR, namely, room acoustic modeling.

1.2 Research context

The incorporation of room acoustic modeling in AAR arises from people's natural auditory perception of the real world, where the perception of a sound source can vary drastically in different environments (Yang et al., 2022) because the acoustic properties of the space such as the room geometry or wall materials influence the propagation of the sound. In particular, we consider acoustic scenes in cuboid rooms (rectangular floor map), commonly referred to as *shoebox* rooms in this thesis. To ensure that virtual sounds are perceived as if they belonged to the physical environment, an AAR system should model room acoustics when rendering virtual sounds. A common approach is to convolve the *dry* sound source that needs to be rendered with an RIR. An option is to perform an *in-situ* measurement of the RIR corresponding to the user's acoustic scene. However, this process is cumbersome, and for real-time AAR systems, it is impractical. Some current AAR systems such as those of Heller et al. (2014) and Mattheiss et al. (2020) overcome this issue by adding artificial reverberation to the rendered sound source, although the closeness of this artificial reverberation to the actual environment of the user has been barely studied. Room acoustics, in general, has been overlooked or inadequately modeled in many AAR systems (Valimaki et al., 2015; Yang et al., 2022).

The forward problem of estimating the RIR of an acoustic scene from a given set of acoustic parameters has been widely studied and there exists a variety of room acoustic simulators that are designed to solve this problem. The inverse problem of estimating acoustic parameters from one or more RIRs has been less studied, because for AR use cases the in-situ measurements of RIRs are impractical. Therefore, a more reasonable approach is to rely on unknown sound sources present in the acoustic scene to measure important acoustic parameters, and then simulate a RIR matching the user's environment based on these parameters.

Acoustic parameters that are relevant for the simulation of RIRs include the wall absorption coefficients, the geometry of the room, information about the reverberation time RT_{60}, and in some cases the location of the sound source. Methods based on signal processing exist for the estimation of these parameters (ISO.ASTM:E1050-9, 2006; Hald et al., 2019; Shlomo and Rafaely, 2021; Gaubitch et al., 2012; Brandão et al., 2015; Grumiaux et al., 2022), however, they all have limitations. These methods are known to struggle in noisy and reverberant conditions, some of them only work with RIR, or they require special apparatus and meticulous control over the user environment and some of them are not sufficiently accurate. Machine learning-based methods using Deep neural network (DNN), and in particular supervised learning, are a resort to this issue. Many applications in audio such as speech enhancement (Richter et al., 2023), speech separation (Hu et al., 2023), and

robust speech recognition (Vincent et al., 2017) have employed methods based on DNN and demonstrated better results in difficult real-world acoustic conditions. Also, DNN-based methods provide the ability to perform multi-task inference of various parameters, one such example is Yu and Kleijn (2020) that simultaneously estimates room geometry and reflection coefficients from a RIR.

Therefore, in all our contributions we will employ supervised DNN methods to address the respective tasks. One caveat with the supervised learning approach is the requirement of a large amount of labeled training data. For better generalization in real conditions, the training data should also be sufficiently diverse. Pre-recorded, annotated datasets consist of samples measured in real acoustic conditions. They are task-specific, for example, the dEchorate dataset (Di Carlo et al., 2021) is specifically designed for the task of room geometry estimation, and the LOCATA dataset (Evers et al., 2020) is commonly used for sound source localization. Also, most available datasets do not contain enough samples for the training of a supervised learning system and the dataset is specific to the recording microphone setup, the sound source, and the acoustic scene. For a DNN system which aims to generalize, the specificity of a training dataset is detrimental. Room acoustic simulators allow physical modeling of an RIRs and finer control over the type of microphone array, sound source, and acoustic scene in the virtual space. The simulation method can vary from wave-based methods which solve the wave equation to simulate an RIR (Svensson and Kristiansen, 2002), to the simpler Image Source Method (ISM) which is based on geometric room acoustics (Allen and Berkley, 1979). Depending on the simulation method, diverse and realistic training data can be generated, where the latter is usually associated with higher computational costs.

1.3 Objective and contributions

In this thesis, we focus on estimating global and local room acoustic parameters from unknown sound sources. Global acoustic parameters include the room's total surface area S, the volume V, the mean absorption coefficients $\overline{\alpha}$, and the reverberation time RT_{60}. As for the local parameters, we only aim to estimate the location of the sound source. We, therefore, contribute to two separate tasks: room parameter estimation and sound source localization. We cater the two tasks using a supervised DNN approach. These DNN systems are trained solely on simulated data, an approach commonly referred to as *virtually supervised learning*. We aim to answer the following questions: Is multi-task room parameter estimation (i.e., jointly estimating all room-acoustic parameters) beneficial compared to separately estimating each room acoustic parameter ? Does multichannel input help or are single-channel features sufficient? Do measurements from multiple positions inside the room help in the estimation of the global parameters? Does improving the simulation techniques help improving the generalization of the room parameter estimator and the sound source localizer on multiple real datasets? What are the specific parameters which help in

simulating realistic impulse responses, and how does the choice of each parameter affect the training and performance of the network on real datasets?

1.3.1 Contributions

1.3.1.1 Multichannel room parameter estimation using multiple viewpoints

Selecting the relevant acoustic parameters of the user environment poses a challenge. The concept of reverberation fingerprint proposed by Jot and Lee (2016) characterizes a room on the basis of its volume and its reverberation time per octave band. These two parameters are related to each other via Sabine's law under ideal diffuse sound field conditions, which also involves the room's total surface area and wall absorption coefficients. This motivated us to perform a joint estimation of these four parameters. Systems proposed in the literature typically estimate each of these parameters separately. Murgai et al. (2017) proposed an algorithm to blindly estimate the reverberation fingerprint based on the decay envelope of a single-channel clean speech signal, Kataria et al. (2017) trained a statistical model to blindly estimate the source position and mean wall absorption coefficient above 1 kHz using binaural signals and interchannel cues, and Genovese et al. (2019) proposesd a DNN for room volume estimation using unknown sound sources. Most of the methods presented in the literature operate on single-channel features and estimate broadband values for the frequency-dependent parameters. For learning-based methods an interesting approach is to learn a latent representation called *Room embedding*, that conditions an end-to-end network to convert an audio recorded in one environment to another. This approach falls short for systems that depend on modeling the RIRs, due to the lack of knowledge of room parameters.

Our first contribution is a multi-task multichannel framework for room acoustic parameter estimation. Specifically, we estimate simultaneously the four mentioned parameters in six octave bands. To do so, we propose a new multi-task DNN architecture that is capable of processing single-channel and multichannel features. This network is trained on simulated data with a maximum likelihood-based loss function that yields adaptive variances for each parameter. This allows us to fuse multiple independent observations from a room in a statistical motivated way. The results are evaluated on simulated and real reverberated speech recordings.

1.3.1.2 Improved simulation and its effect on room parameter estimation

A common bottleneck for any learning-based system is the need for a large annotated training dataset. For audio systems, a widely used technique to emulate sound scenes in a variety of acoustic spaces is to convolve dry source signals with RIRs. Acquiring thousands of real RIRs in various acoustic scenes is impractical. Therefore, in recent years, research has been directed towards the use of data augmentation techniques. Specifically, for the

task of room parameter estimation, Gamper and Tashev (2018), Genovese et al. (2019), and Götz et al. (2022) demonstrated that combining real RIRs with synthetic RIRs improves the generalization of the DNN for the task of room volume estimation and reverberation time estimation. While these studies show the effectiveness of using a few hundred annotated real RIRs for training, obtaining such real data is not always feasible for all tasks, and the acquisition process can be expensive even for a small amount.

An alternative is to train a DNN only on simulated data using the approach of *virtually supervised learning*. One approach could be to use large-scale simulated training datasets of high-quality RIRs. A dataset published by Tang et al. (2022) consists of approximately 2 million high quality RIRs that were simulated in various acoustic scenes using a highly accurate simulator combining a finite-difference time-domain wave equation solver at low frequencies with ray tracing at high-frequencies. To create an accurate acoustic environment they combined 3D realistic house models with an automatically matched database of material absorption profiles. The dataset was tested on a variety of single-channel speech processing tasks and showcased improved generalization compared to the model trained on less realistic simulators. However, generating such large datasets is computationally demanding, and also they cannot be generalize to, for example, multichannel settings with a specific microphone array geometry or to specific tasks.

Our second and third contributions address this issue. First, we improve the realism of the RIRs simulated by the open-source room acoustic simulator Pyroomacoustic (Scheibler et al., 2018) by implementing an extended version of ISM with frequency-dependent source and receiver directivity without significantly increasing its computational cost. Then, we train and evaluate our multi-task multichannel room acoustic parameter estimation model on data obtained with this simulator, and conduct an ablation study to evaluate the effects of different source and/or receiver directivity profiles and wall absorption coefficient distributions. The model is then tested on real speech signals from the dEchorate dataset, comprising four different room configurations.

1.3.1.3 Improved simulation and its effect on speaker localization

Our fourth contribution diverts from global to local room acoustic parameter estimation. Specifically, we focus on localizing a single active speaker in the environment. In the context of virtually supervised learning for source localization, the literature has a plethora of work on training DNN models with a simple ISM based simulator (Chakrabarty and Habets, 2019; Adavanne et al., 2018a; Diaz-Guerra et al., 2020; Nguyen et al., 2020). These studies are using the simplest version of shoebox ISM with frequency-independent omnidirectional sources and microphones combined with frequency-independent wall absorption coefficients, which makes the simulation process computationally inexpensive. While ISM simulators employ acoustic simplification compared to advanced simulators, these systems have still shown decent performance on real test sets. A recent study on source localization by Gelderblom et al. (2021) used an advanced ISM which models source directivity and

diffuse late reverberation. Their results show that source directivity has a positive impact while diffusion shows no added improvement.

Despite the widespread use of shoebox ISM-based simulators for training speaker localization systems, the impact of incorporating more realistic source and receiver directivities and surface absorption profiles at training time, and its resulting effect on generalization ability at test time, has been scarcely studied. This contribution aims to address this gap by using the extended ISM introduced in our second contribution and performing an ablation study to understand the influence of each added layer of realism on the training set. We present results on three real datasets, comprising speech excerpts from human speakers in different rooms and microphone arrays.

1.4 List of published papers

The research conducted during this Ph.D. program resulted in the following publications.

- **P. Srivastava**, A. Deleforge, and E. Vincent. *Blind room parameter estimation using multiple multichannel speech recordings.* In Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2021. (pp. 226-230)
- **P. Srivastava**, A. Deleforge, and E. Vincent. *Realistic sources, receivers and walls improve the generalisability of virtually-supervised blind acoustic parameter estimators.* In International Workshop on Acoustic Signal Enhancement (IWAENC), 2022. (pp. 1-5)
- **P. Srivastava**, A.Politis, A. Deleforge, and E. Vincent. *How to (virtually) train your speaker localizer.* In Interspeech, 2023.

One secondary contribution, not presented here, was also made during the same period.

- A. Politis, S. Adavanne, D. Krause, A. Deleforge, **P. Srivastava**, T. Virtanen. *A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection.* In Proc. DCASE, 2021, (pp. 125–129).

1.5 Structure of the thesis

Chapter 2 provides the necessary background for a better understanding of the work. It introduces many fundamental equations and their notations, which will be used throughout the thesis.

Chapter 3 is an overview of the state of the art for room parameter estimation, source localization, and virtually supervised learning. It also provides details on various room acoustic simulators and existing audio datasets, that are integral parts of a virtually supervised learning approach.

Chapter 4 presents a novel DNN architecture for multi-task, multichannel room parameter estimation using multiple view-points.

Chapter 5 introduces an advanced room acoustic simulator based on an extension of

the ISM. This work is a contribution to the open-source simulator Pyroomacoustics [1]

Chapter 6 presents an ablation study on the effect of using this advanced ISM simulator for training DNN models. The study is presented for the tasks of room parameter estimation and source localization.

Chapter 7 summarizes the thesis and provides future research directions.

[1] https://github.com/LCAV/pyroomacoustics/tree/dev/dirpat

2 Background

This chapter aims to briefly describe the concepts related to audio and acoustics that will be used throughout this thesis. Section 2.1 describes devices that are used to capture and produce sound in an environment. It is followed by Section 2.2 that lays the fundamentals of the digital signal model. After this, Section 2.3 presents the different techniques that are used for the representation of the discrete sound signals in the frequency and time-frequency domains. Section 2.4 begins with wave propagation theory and goes further to explain sound reflection, diffraction, and scattering, which help defining room acoustics and room impulse responses. Section 2.5 finishes this chapter with a brief introduction to the deep learning techniques that have been used in this work.

Most of the material in Section 2.2 and 2.3 is inspired from the books of Vincent et al. (2018) and Kuttruff (2016), except the subsection on spherical harmonics that is taken from Zotter (2009). Other sections are motivated by the PhD theses of Schröder (2011), Pariente (2021) and Di Carlo (2020).

2.1 Microphones and loudspeakers

Microphones and loudspeakers are meant to emit and record sound signals. In this work, different types of microphones and loudspeakers are used in acoustic simulations to increase the diversity of the data used to train our models. Additionally, a variety of microphones and loudspeakers are used in the testing of our models. In the following, we briefly describe the functioning of different loudspeakers and microphones, their frequency response, and their directivity, which are repeatedly used at various points in this work.

2.1.1 Microphones

Microphones are transducers that convert the changes in acoustic pressure to equivalent changes in electric current. The amplitude and frequency fluctuation in the pressure field is effectively transferred to the AC voltage produced by the microphone. Microphones can be defined on the basis of their internal working structure (Lee and Lee, 2008).

1. Dynamic microphones: A diaphragm that detects the change in pressure field is attached to a voice coil. Changes in pressure field lead to movement of the voice coil across magnets generating electric current. Dynamic microphones are known for being rugged and do not require an external power source.

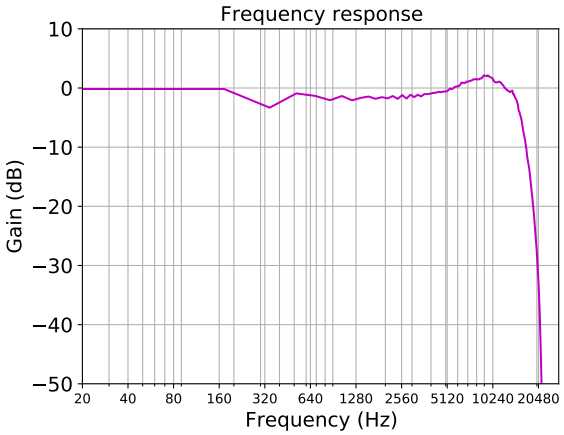

Figure 2.1: Frequency response of the AKG c480 microphone. The data is from the DIRPAT dataset Brandner et al. (2018).

2. Condenser microphone: Instead of a voice coil, condenser microphones use charged plates that move back and forth according to the air pressure detected by the diaphragm. Their use is combined with a high impedance amplifier. Condenser microphones are used in studios and are known for detecting low frequencies and delivering crisp audio. Condenser microphones have been extensively used in this study. Simulation and recording of the datasets were mostly done using one of these microphones: AKG C414 Series, AKG CK32, SHURE MX 391/0.
3. Electret microphones: Shares similarity to Electrostatic capacitor microphone. The diaphragm is attached to a capacitor, and the capacitor is permanently charged evicting the need of a constant source of electrical charge. The voltage varies with change in pressure field resulting in current flow.Micro Electromechanical systems (MEMS) use the same working principle but are constructed differently. Components and systems in MEMS are miniaturized and they sit on a single die taking advantage of the semiconductor manufacturing process. Data recorded by MEMS microphones have been used in this work as part of the VoiceHome2 dataset (Bertin et al., 2019).

2.1.1.1 Frequency response

Sound waves consist of various rapidly changing frequencies and amplitudes. When converting a sound wave into electric current, a microphone may enhance or attenuate the amplitude at each frequency. The frequency response of a microphone provides information about the range of frequencies to which the microphone is sensitive and the degree of enhancement or attenuation of each frequency. Some microphones are capable of recreating sound in the full frequency range, while most of them are only sensitive to a limited frequency. Figure 2.1 shows the frequency response of an AKG c480 microphone measured

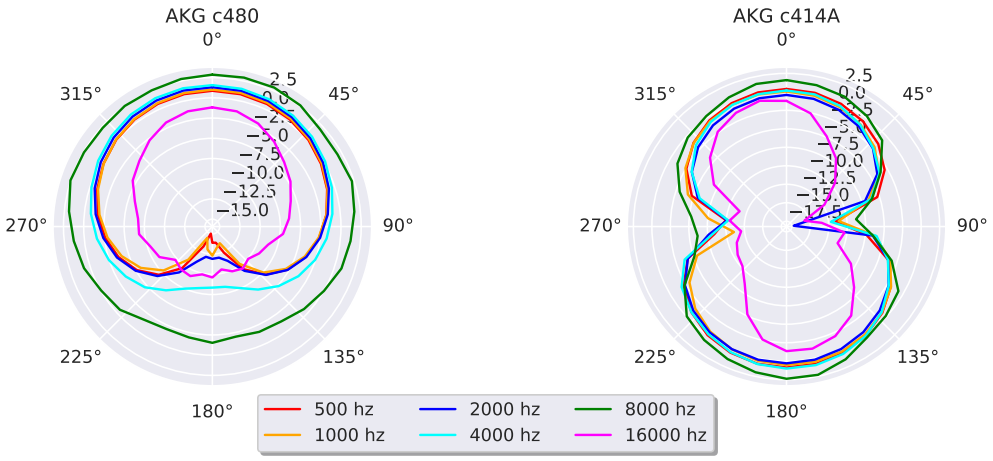

Figure 2.2: Directivity pattern of the AKG c480 (cardioid) microphone and AKG c414A (figure-of-eight) microphone for 6 octave bands shown at the equator elevation. The data is taken from the DIRPAT dataset (Brandner et al., 2018).

in decibels (dB). 0 dB acts as a reference line: frequencies with a gain equal to 0 dB are perfectly reproduced without attenuation, while frequencies with a gain greater than 0 dB are enhanced and frequencies with a gain below 0 dB get attenuated.

2.1.1.2 Directivity

Directional microphones are sensitive to the sound arriving from a particular angle or direction. Their frequency response varies in relation to the direction of the sound pressure field toward the microphone. Usually, the capsules in directional microphones are designed so as to let in sound pressure fields from only the specific paths from the front of the diaphragm to the back. This affects the high-frequency roll-off. There are a number of directional patterns available for microphones.

1. Omnidirectional microphones have a constant sensitivity irrespective of the direction of arrival.
2. Unidirectional microphones are more sensitive to the sounds arriving from the front direction, fixed as 0° in Figure 2.2, and are less sensitive to the other directions. There exist 3 common types of directional patterns: cardioid, super-cardioid and hyper-cardioid.
3. Bidirectional microphone have maximum sensitivity in two opposite directions and minimum sensitivity in the other two directions. These directivity patterns are often described as figure-of-eight.

The directionality of the microphone often varies with frequency. Omnidirectional microphones tend to behave directive above 500 Hz, similarly, directive microphones are less directive below 500 hz. Figure 2.2 illustrates the directivity pattern of a cardioid microphone across 6 octave bands. The source-to-microphone distance also directly affects the frequency response of a directional microphone: a much closer source results in a boosting of the lower frequencies, a phenomena known as the *proximity effect*. To avoid this proximity effect, most of our experiments have been conducted, with a minimum source-to-microphone distance of 30 cm.

2.1.1.3 Ambisonics

Some microphone arrays record sound in specific formats referred to as *Ambisonics*. It is a method used to encode a sound field based on its directional properties.

1. Zero-order Ambisonic consists of only one channel that has pressure field information at the origin W.
2. First-order ambisonic (FOA), also known as B-format (Gerzon, 1975), contains 4 channels where the first channel is identical to zero order Ambisonics and the other 3 channels encode the acoustic velocity at the origin along 3 perpendicular axes X,Y,Z.
3. Higher-order Ambisonics (HOA) contains $(l+1)^2$ channels for a given order $l \geq 2$, which encode the spherical harmonics decomposition of the sound field (see Section 2.3.4), resulting in a much finer spatial resolution.

As a result, any 3D sound fields can be represented, up to a spatial resolution depending on the chosen order l (Zotter and Frank, 2019). FOA signals can be recorded with soundfield microphones such as OKTAVA MK 4012, Soundfield ST450, while HOA signals are often recorded with Eigenmike em32 consisting of 32 channels, which are then encoded into 25 Ambisonic channels ($l = 4$). The Eigenmike is used in Chapter 6 as a part of STARS22 dataset (Politis et al., 2022).

2.1.2 Loudspeakers

Loudspeakers are also a type of transducer and convert electric signals into sound signals. A loudspeaker consists of a movable coil adjoined by magnets. The coil moves back and forth with respect to the varying electric current, which causes a cone attached to the moving coil to convert the electric current to a pressure field. As microphone and loudspeaker both work on the same principle, they can be switched to perform opposing functions, albeit usually with poor performance. It is mainly because of physical constraints and the design construction of both devices.

2.1.2.1 Frequency response

The range of frequencies over which a loudspeaker can transmit sound signals is represented by its frequency response. Loudspeakers can be divided into 4 categories based on

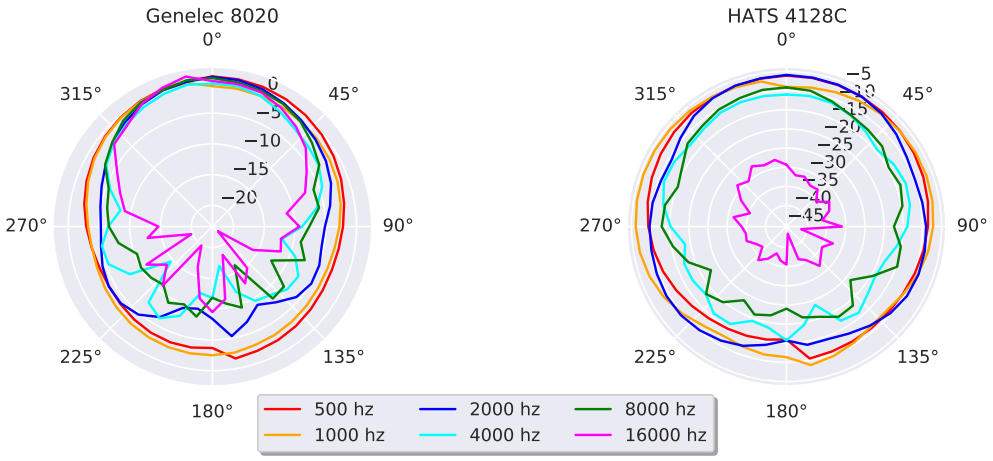

Figure 2.3: Directivity pattern of the Genelec 8020 loudspeaker and B&K head and torso mannequin (HATS 4128C) for 6 octave bands shown at the equator elevation. The data is taken from the DIRPAT dataset (Brandner et al., 2018)

their frequency range. Woofers, midrange and tweeters transmit low (< 200 Hz), medium ($500 - 3,000$ Hz) and high ($> 3,000$ Hz) frequencies, respectively, while full-range loudspeakers work throughout the frequency range $100 - 15,000$ Hz. The specified loudspeakers can be used together in a speaker system by dividing the sound spectrum into parts, and sending the sound for a certain frequency range to the corresponding speaker system component. This can be achieved using an electric circuit known as a crossover. Multiple crossovers can build a crossover network, also known as a band-pass filter.

2.1.2.2 Directivity

Two types of sources are considered in this thesis: loudspeaker, and human speaker. Human speakers are directive toward the front face while the intensity of directivity depends on the speaker's orientation. Similarly, the directivity of a loudspeaker depends on its orientation, although there exists certain types of speakers that are directive in multiple directions or even close to omnidirectional, such as dodecahedral loudspeaker arrays. Similar to directional directivity in the microphones, loudspeakers radiate in a wider angular range at low frequency and vice-versa at high frequency. The area of this radiation can be defined by calculating the effective radiation angle for a speaker in the horizontal plane. Figure 2.3 shows the directivity pattern of a loudspeaker whose radiation angle is $120°$ degrees, i.e., a listener standing at $60°$ in both directions of the circle will perceive the volume of sound at similar level as the zero-degree direction. This thesis involves the simulation

of both omnidirectional and directive loudspeakers, including loudspeaker models such as the YAMAHA DXR8 or the Genelec 8020. In an attempt to simulate a human speaker, a B&K head and torso mannequin is also used. Apart from their utilization in simulated data, real data is also used employing the aforementioned directive loudspeakers and real human speakers.

2.2 Digital signal model

2.2.1 Analog-to-digital conversion

Sound pressure field at a microphone position is a function of continous time. The continuous pressure field is approximated as a continuous raw analog signal captured by a microphone. For any given time t, the continous function outputs a value and the signal does not break. Continuous functions are represented as $\tilde{x}(t)$ where $t \in \mathbb{R}$. As most of the processing takes place in digital systems, this continuous signal is converted to a discrete signal $x[n]$ for $n \in \mathbb{Z}$ by sampling periodically at rate F_s [Hz] :

$$x[n] = (\kappa_{LP} \star \tilde{x})(n \cdot T)\,. \tag{2.1}$$

Here, κ_{LP} is an analog low pass filter with frequency support in $]-F_s/2, F_s/2]$, $T = 1/F_s$ is the sampling period for $n \in \mathbb{Z}$ and $\star$ is the convolution operator in continuous time. A formal definition of convolution in time domain for two signals $\tilde{v}$ and $\tilde{a}$ is given by,

$$(\tilde{v} \star \tilde{a})(t) = \int_{-\infty}^{\infty} \tilde{v}(\tau)\tilde{a}(t-\tau)d\tau\,. \tag{2.2}$$

2.2.2 Signal model and terminologies

We define the signal model in discrete time for a static point source and a point microphone in a homogeneous medium (air) as,

$$c_{mj}[n] = (h_{mj} * s_j)[n] + \eta_m[n]\,, \tag{2.3}$$

where $*$ denotes discrete convolution and the source signal s_j is modeled as a band-limited discrete-time signal. h_{mj} is the RIR, that encapsulates acoustic information of the room. It is a system response (in this case the room) captured by microphone m, when excited with an impulse from source j. The positions and directivities of the source and microphone influence the RIR. The discrete-time convolution of h_{mj} with s_j generates a reverberated signal. Here $m \in \{1, \ldots, M\}$, where M denotes the number of *channels* present in the microphone array. $M = 1$ is called a *single-channel* system while $M > 1$ is a multi-channel system. The resulting reverberated signal for each source j and microphone m is denoted c_{mj}. The combined signal for source j at all M channels, can be written as a vector-valued signal in $\mathbb{R}^M$:

$$\mathbf{c}_j[n] = [c_{1j}[n], ..., c_{Mj}[n]]\,. \tag{2.4}$$

In the signal model, η_m is the background noise. In our experiments, we conceptualize undesired speech or nonspeech sources as point source *interferers* or *diffuse noise*, which are considered as background noise for the mixtures depending on the use-case scenario. Diffuse noise refers to the noise that is present in the background of the signal. Examples include noise from air conditioning vents, car noise, wind noise etc. Such noise are spatially diffuse, i.e., they cannot be associated with a single point in space. Two types of diffuse noise are simulated in this thesis: additive white Gaussian noise (AWGN) and speech-shaped noise (SSN) whose average power spectrum matches that of babble noise. Both are provided with spatial characteristics of real diffuse noise (Habets and Gannot, 2007).We assume J different point sources in the room and they are indexed by $j \in \{1, \ldots, J\}$. Point sources are defined as distinct points in space which emit sound. *Far-field* is a notion that is generally relative to the size of the antenna/size of the source and to the wavelength. In a far-field scenario, point sources are a good approximation of human speakers or loudspeakers. This theory is described in detail using the wave equation in Section 2.4.1. If all the J point sources are active at the same time, then the respective sounds fields reaching the microphone array are additively combined into a single sound. The resulting superimposition of sound signals into one is defined as the *mixture* $\mathbf{c}[n]$.

$$\mathbf{c}[n] = \sum_{j=1}^{J} \mathbf{c}_j[n] \tag{2.5}$$

Interferers are unwanted signals that have similar properties as point sources. Likewise, in a mixture of speech signals where J speech signals are active at the same time, each speech source acts as an interferer for the other active sources.

2.3 Sound representations

2.3.1 Discrete Fourier transform

Sound signals are usually presented in discrete time domain representations known as waveforms. Another way of describing the content of the signal is its representation in the frequency domain. The Fourier transform is one such tool that analyzes the frequency content of the signal. One specific flavour of Fourier transform called the Discrete Fourier Transform (DFT) is especially applicable to finite length discrete signals. So far we considered infinite length discrete signals that help in performing linear convolutions. In the following, we therefore consider finite-length discrete signals $x[n]$ with $n \in \{0, ..., N-1\}$. Equivalently, infinite-length discrete periodic signals maybe considered. The DFT converts the signal from the time domain to the frequency domain by using a set of basis vectors made up of complex exponentials that are linearly spaced in frequency (see Figure 2.4). DFT analysis can be written as :

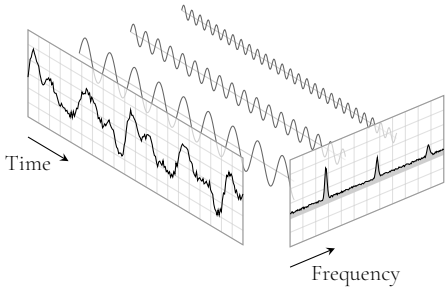

Figure 2.4: A signal decomposed as a Fourier series, which is a sum of sines and cosines that are represented as peaks in the frequency domain.

$$X[f] = \sum_{n=0}^{N-1} x[n] e^{-2i\pi fn/F}, f \in \{0, ..., F-1\}. \tag{2.6}$$

The frequency coefficients $X[f]$ are complex numbers, the magnitudes $|X[f]|$ of these coefficients form the *magnitude spectrum*, while the arguments $\angle X[f]$ constitute the phase spectrum. When $F \geq N$, the DFT is invertible, hence, the coefficients $X[f]$ can be used to synthesize back the original time-domain signal using the inverse discrete Fourier transform (IDFT). The DFT exhibits many important properties shown by Lyons (2004). We state two properties that are used frequently in this thesis.

- *Linearity* indicates that the DFT of the sum of two signals is equal to the sum of the individual transform of each signal :

$$\mathbf{c}[n] = \sum_{j=1}^{J} \mathbf{c}_j[n] \overset{\mathrm{F}}{\rightarrow} \hat{\mathbf{c}}[f] = \sum_{j=1}^{J} \hat{\mathbf{c}}_j[f]. \tag{2.7}$$

- The *Convolution Theorm* for the DFT states that for two finite length discrete time domain signals v and a, the product of their DFTs is the circular convolution $\circledast$ between the individual discrete sequences :

$$x[n] = (v \circledast a)[n] \overset{\mathrm{F}}{\rightarrow} X[f] = V[f]A[f]. \tag{2.8}$$

  With proper zero padding of both signals in the time domain, this equation also become valid for linear convolution.

2.3.2 Short-time Fourier transform

The frequency components of signals such as music and speech exhibit variations over time. To simultaneously track the variations in both time and frequency, we represent the signal in the *time-frequency domain*. The short time Fourier transform (STFT) is one such

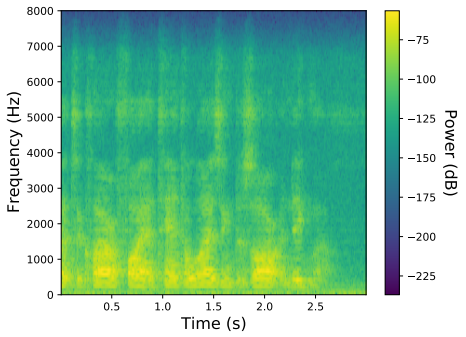

Figure 2.5: Speech signal represented as a STFT spectrogram

time-frequency representation, which is illustrated in Figure 2.5. The main idea is to divide the signal into small chunks of similar length and compute the DFT over each of these chunks. The STFT is formulated as,

$$X[f,a] = \sum_{n=0}^{L^{\text{win}}-1} w[n]x[n + aL^{\text{hop}}]e^{-i2\pi fn/F} \in \mathbb{C}\,. \tag{2.9}$$

The window function $w[n]$ of length L^{win} is repeatedly used over the signal to create *frames*. Each frame is analyzed using the DFT and then $w[n]$ hops to the next part of the signal, where L^{hop} is the hop size. The resulting time-frequency coefficients $X[f,a]$ form a complex-valued matrix $X \in \mathbb{C}^F$, where (f,a) are the frequency and time indices. Similar to the DFT, the STFT also has magnitude and phase spectrograms. The power spectrogram $|X[f,a]|^2$ represents the spread of power across frequency and time. Power spectrograms are best visualized in log-power scale rather than in linear scale, due to the wide dynamic range of natural sounds. This results in a log-power spectrum in dB scale: $10\log_{10}|X[f,a]|^2$.

Apart from these conversion of spectogram scales, both magnitude and phase spectograms encodes crucial information. Magnitude spectograms present information on how much frequency content is present per frame, while the latter encodes change of phase as a function of frequency and time. The choice of window length becomes a crucial parameter: long window lengths represent the frequency content accurately but lead to poor time resolution, while short window lengths results in the opposite trade-off. This is in line with Gabor's uncertainty principle, which states that it is infeasible to locate a signal in both time and frequency accurately (Benedetto, 2021).

In this thesis we will experiment with magnitude and phase spectrograms and their scaled variants such as power and log-power spectrograms.

2.3.3 Spherical harmonics

The directivities of sources and microphones can be represented as functions of azimuth θ and elevation ϕ on the sphere, $g(\theta, \phi)$, also called *spherical functions*. More specifically, the output of these spherical functions for a particular value of (θ, ϕ) is an impulse response $g(\theta, \phi, t)$ or a frequency response $G(\theta, \phi, f)$, where G is the Fourier transform of g. The spatial resolution of the directivity pattern depends on the discretization of the sphere, i.e., the grid of (θ, ϕ) values at which impulse responses have been measured. More points on the grid lead to higher resolution and vice-versa. Usually, the directivity patterns of sources and microphones are provided by their respective manufacturers, and, the corresponding grid varies from one to the another. For some applications, including the work in Chapter 5 of this thesis, there arises a need of high-resolution directivity patterns. A work-around to this is to extrapolate or interpolate the responses to a finer target grid with increased resolution. The spherical harmonics transform is one such tool used to solve this problem.

Spherical harmonics provide a set of orthogonal basis functions on the sphere. These functions originate from solving Laplace's equation in spherical coordinates. The spherical harmonic basis function is given by:

$$Y_l^w(\theta, \phi) = N_l^w P_l^w(\cos\theta) e^{iw\phi} \,. \tag{2.10}$$

$Y_l^w(\theta, \phi)$ is the spherical harmonic fucntion of degree l and order w , where $l \in \mathbb{N}$ and $-l \leq w \leq l$. This gives $2l + 1$ basis functions for degree l. The right handside of (2.10) is made up of 3 different parts. N_l^w denotes the normalization constant,

$$N_l^w = \sqrt{\frac{(2l+1)}{4\pi} \frac{(l-w)!}{(l-w)!}} \,. \tag{2.11}$$

$P_l^{|w|}(\cos\theta)$ unties the function on the θ axis, where $P_l^{|w|}$ is the associated Legendre polynomial and $\cos\theta$ maps θ to the [-1,1] interval. $e^{iw\phi}$ can be expanded to sinusoids using Euler's formula and it shows the dependence over the ϕ axis.

2.3.4 Discrete spherical harmonic transform

As explained previously, the set of spherical harmonics forms an orthonormal basis for functions over the sphere. Using them, any real valued spherical function can be expanded as a sum of an infinite set of scaled spherical harmonics:

$$g(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{w=-l}^{l} Z_{l,w} Y_l^w(\theta, \phi) \,, \tag{2.12}$$

where $Z_{l,w}$ denotes the coefficients on spherical harmonics. In most of the practical cases, the spherical function $g(\theta, \phi)$ is band-limited and the function is sampled over a discrete

set of spherical angles. Therefore we change (2.12) to an expansion on a discrete grid (θ_i, ϕ_i), where i is the discrete index.

$$g(\theta_i, \phi_u) = \sum_{l=0}^{L-1} \sum_{w=-l}^{l} Z_{l,w} Y_l^w(\theta_i, \phi_u)\,. \tag{2.13}$$

This discrete expansion is known as the inverse discrete spherical harmonics transform (DSHT). The exact expansion of the function g might require an infinite basis set with $l \to \infty$, but an approximate representation can be calculated using a degree l up to some threshold L. Low- spatial-frequency components in g can be well represented with few degrees, while high-spatial-frequency components need a higher degree for good approximation.

The coefficients of the spherical harmonics expansion for a discrete function which is sampled on a strictly sampled grid (i.e quadrature) (Zotter, 2009) can be found using the analysis equation,

$$Z_{l,w} = \sum_{i=0}^{I-1} \sum_{u=0}^{U-1} q_i g(\theta_i, \phi_u) \bar{Y}_l^w(\theta_i, \phi_u)\,, \tag{2.14}$$

also called forward Discrete Spherical Harmonics Transform (DSHT), where $\bar{Y}_l^w$ denotes the complex conjugate of the spherical basis function and q_i denotes quadrature weights. According to Zotter (2009), (2.14) falls under the category of forward DSHT using quadratures. A discussion about different types of quadrature weights is found in studies by Sneeuw (1994) and Driscoll and Healy (1994). Additionally, Zotter (2009) also points to the strict requirement on the order of the spherical harmonics and quadrature. In many cases, depending on the nature of the spherical function and of the discrete grid on which the function is sampled, the inverse problem of computing the coefficients $Z_{l,w}$ could be ill-defined. Therefore, a variant of forward DSHT transform using least squares is proposed by Zotter (2009). It does not pose a strict requirement on the sampled grid, although it requires a least square pseudo-inverse of the spherical harmonic basis. For uniform distribution of least-square errors on the surface of the sphere, a *weighted-least square* approach is employed, whose vectorized form is,

$$\mathbf{z}_L = (\mathbf{Y}_L)^{+} \mathbf{Q} \mathbf{g}\,, \tag{2.15}$$

where $\mathbf{z}_L = [Z_{0,0}, \ldots, Z_{L,L}]^\intercal$ is a column vector and $\mathbf{Y}_L = [\mathbf{y}_L(\theta_0, \phi_0), \ldots, \mathbf{y}_L(\theta_I, \phi_U)]$ is a matrix of column vectors $\mathbf{y}_L = [Y_0^0(\theta_i, \phi_u), \ldots, Y_L^L(\theta_i, \phi_u)]^\intercal$. Here, $+$ denotes the pseudo-inverse, $\mathbf{Q}$ is a diagonal matrix whose entries are the surfaces of the Voronoi cells associated with the points on the grid [1], and $\mathbf{g}$ is a vector of spherical function $g(\theta, \phi)$. Throughout this thesis, a forward DSHT should be associated with Equation (2.15).

[1] Equation (2.15), is inspired from the vectorized notation provided in Zotter (2009). For more details, the reader is encouraged to look at Chapter 4 of Zotter (2009)

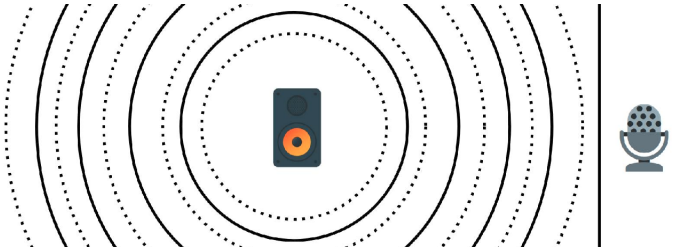

Figure 2.6: Sound wave propagation in the far-field scenario. Incoming waves towards the microphone can be approximated as plane waves.

2.4 Acoustics

2.4.1 Sound wave propagation

Sound can be defined as a vibration that carries energy through a medium such as air or water. Throughout this work, we consider the medium of propagation as air and the corresponding speed of sound $c = 343$ m/s. The vibrations produced by the excited sound source cause the air molecules in the medium to expand and contract. This produces an oscillatory motion, leading to a fluctuation of air pressure. This fluctuation as a function of space (position $\boldsymbol{r}$) and time t defines the *sound pressure field* p. The acoustic wave equation describes the evolution of p as a function of $(\boldsymbol{r},t)$ in *free-field* conditions, i.e., open air without any obstacles (no-boundaries) :

$$\nabla^2 p(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{r},t)}{\partial t^2} = 0\,, \tag{2.16}$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ is the Laplacian operator in 3D space. Due to the absence of a sound source it is also referred to as the homogeneous wave equation.

The Fourier transform of Equation (2.16) is known as the *Helmholtz equation* :

$$\nabla^2 P(\boldsymbol{r},f) + k^2 P(\boldsymbol{r},f) = 0\,, \tag{2.17}$$

where k is the *wave number* that relates to the frequency f via $k = \frac{2\pi f}{c}$, and P denotes the Fourier transform of p. Adding a source signal $s_0^{\text{src}}(t)$ located at $\boldsymbol{r}_0^{\text{src}}$ to the right side of Equation (2.16) makes it inhomogeneous,

$$\nabla^2 p(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{r},t)}{\partial t^2} = s_0^{\text{src}}(t)\delta(\boldsymbol{r}-\boldsymbol{r}_0^{\text{src}})\,. \tag{2.18}$$

When the source signal is a Dirac signal $\delta(t)$ in time, the above equation becomes :

$$\nabla^2 h(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2 h(\boldsymbol{r},t)}{\partial t^2} = \delta(t)\delta(\boldsymbol{r}-\boldsymbol{r}_0^{\text{src}})\,, \tag{2.19}$$

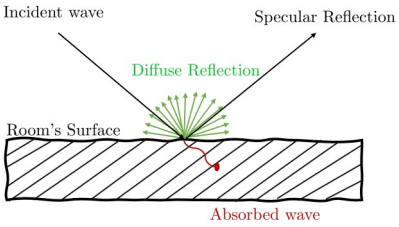

Figure 2.7: Reflection, scattering, and absorption of an incident sound wave

where $h(\boldsymbol{r}, t)$ denotes the response of the system to a Dirac. A standard solution of this equation for a receiver placed at $\boldsymbol{r}^{\text{mic}}$ is mentioned by Kuttruff (2016) and is written as :

$$h(\boldsymbol{r}^{\text{mic}}, t) = \frac{1}{4\pi \|\boldsymbol{r}^{\text{mic}} - \boldsymbol{r}_0^{\text{src}}\|} \delta\left(t - \frac{\|\boldsymbol{r}^{\text{mic}} - \boldsymbol{r}_0^{\text{src}}\|}{c}\right). \tag{2.20}$$

This result can be interpreted as an outward spherical wave propagating at the speed of sound in free-field conditions. The signal emitted by the source is attenuated by a factor of $\frac{1}{4\pi \|\boldsymbol{r}^{\text{mic}} - \boldsymbol{r}_0^{\text{src}}\|}$ and delayed by $\frac{\|\boldsymbol{r}^{\text{mic}} - \boldsymbol{r}_0^{\text{src}}\|}{c}$ when reaching the receiver. Figure 2.6 shows the propagation of an omnidirectional point source in free-field conditions, where the pressure field is moving away from the source in a spherical pattern. As shown, the microphone is placed far away from the source, hence the curvature of the waves can be ignored due to waves becoming close to plane waves. This scenario is called *far-field*. Conversely, the curvature of waves should be taken into account when the receiver is very close to the source, which describes the *near-field* scenario.

2.4.2 Room acoustics

Until now, we have presented propagation of sound under the free-field scenario, although in this thesis we work with real or simulated rooms. A room is an enclosed space that is bounded by sound reflecting materials on the walls, ceiling and floor. The characteristics of the sound wave gets changed when it encounters the boundaries of the room. The acoustic and geometric properties of each encountered surface determine how much the sound wave is reflected, diffracted or absorbed by the surface. A combination of these effects is responsible for *room acoustics* and complex sound fields. In the rest of this section, we assume that the sound source is not too close to a wall, so that the sound waves are treated as incident and undisturbed plane waves based on the far-field scenario.

2.4.2.1 Reflection and scattering

Reflection occurs when a sound wave hits a solid surface. It often changes the amplitude and the phase of the impinging sound wave. Primarily, there are two types of acoustic

reflections.

1. Specular reflections : follow the Snell-Descartes law, i.e., the angle of incidence and the angle of reflection are equal.
2. Diffuse reflections : refer to sound waves getting scattered in every direction.

These reflections depend on the surface irregularities and the wavelength of the sound wave. A perfectly rigid surface reflects sound rays in a specular manner, while on an irregular surface, the behavior depends on the size a of the irregularities and the wavelength λ of the sound wave.

- When $\lambda >> a$ surface irregularities are non-existent, thus sound rays are reflected in a specular manner.
- When $\lambda \approx a$ each sound ray bounces in a different direction, resulting in a *scattering effect*.
- $\lambda << a$ leads to specular reflection on each individual irregularity present on the surface.

Specular reflections either reflect or absorb a certain fraction of energy, depending on the surface material. This is modeled by the following quantities:

- The acoustic impedance $\hat{Z}_v(f)$ characterizes the behaviour of surface v in terms of its rigidity or penetrability. It is defined as the ratio between sound pressure and particle velocity on the surface. It takes values in $\mathbb{C}$ with real and imaginary parts called acoustic resistance and reactance. The former describes the lost energy and the latter the stored energy.
- The reflection coefficient β is the proportion of incident wave magnitude absorbed by the surface. It depends on θ^{in} through the following formula :

$$\beta(f, \theta^{\text{in}}) = \frac{\hat{Z}_v(f)\cos\theta^{\text{in}} - \xi^{\text{air}}(f)}{\hat{Z}_v(f)\cos\theta^{\text{in}} + \xi^{\text{air}}(f)}, \tag{2.21}$$

  where $\xi^{\text{air}}(f)$ denotes the intrinsic impedence of the propagation medium (air).
- The absorption coefficient $\alpha(f)$ represents the ratio of energy that is not reflected by the surface. It is defined via the following approximations, primarily due to difficulties that arise while measuring acoustic impedance (Di Carlo, 2020) :

$$\alpha(f) = 1 - |\overline{\beta}(f)|^2, \tag{2.22}$$

  where $\overline{\beta}(f)$ denotes the average reflection coefficient averaged over all incidence angles.
- The scattering coefficient Υ of surface material is defined as the proportion of impinging energy that gets scattered.

2.4.2.2 Room impulse response

The inhomogenous free-field wave equation (2.19) for a point source emitting $\delta(t)$ is defined for a space Ω, where $\Omega \subseteq \mathbb{R}^3$. In order to complement the wave equation with

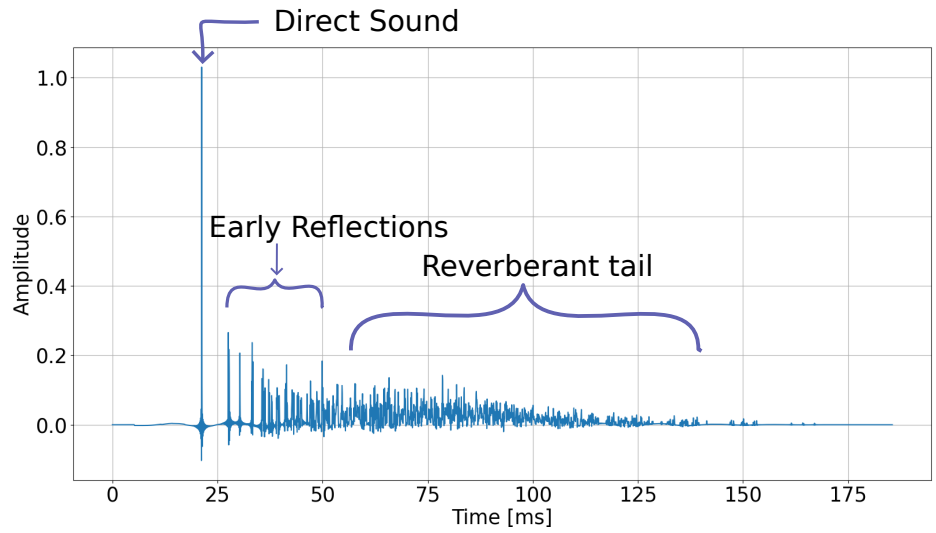

Figure 2.8: Schematic illustration of the components of RIR.

boundary conditions, another equation needs to be defined on the boundary

$$\boldsymbol{n}(\boldsymbol{r}) \,.\, \nabla h(\boldsymbol{r}, t) + \frac{\partial}{\partial t}[\varepsilon \star h](\boldsymbol{r}, t) = 0 \,, \boldsymbol{r} \in \partial\Omega \,, \tag{2.23}$$

where $\boldsymbol{n}(\boldsymbol{r})$ is the normal vector of $\partial\Omega$ at $\boldsymbol{r}$ and $\varepsilon(\boldsymbol{r}, t)$ is the admittance that defines the acoustic properties of the room boundary at $\boldsymbol{r}$. Solving for Equations (2.19) and (2.23) provides us with the RIR of the room denoted as $h(\boldsymbol{r}, t)$. This solution is also called a *Green's function*. The RIR establishes a straightforward relation between the sound pressure field $p(\boldsymbol{r}, t)$ and the source signal $s^{\text{src}}(t)$ at $\boldsymbol{r}_0^{\text{src}}$ which is linear and time invariant, according to equation (2.19). Hence, this relation can be expressed as,

$$p(\boldsymbol{r}, t) = [h(., t) \star s](t) \,. \tag{2.24}$$

The RIR is specific to the room geometry and the acoustic properties of its boundary surfaces along with source and microphone positions. A generalized form of Equation (2.24) for a source signal $s(t)$ positioned at $\boldsymbol{r}_0^{src}$ for M different microphones present in the scene is :

$$\mathbf{p}(t) = [\mathbf{h} \star s](t) \,, \tag{2.25}$$

where $\mathbf{p}(t) = [p_1(t), \ldots, p_M(t)]$ and $\mathbf{h}(t) = [h_1(t), \ldots, h_M(t)]$, while $h_m(t) = h(\boldsymbol{r}_m^{\text{mic}}, t)$ and $p_m(t) = p(\boldsymbol{r}_m^{\text{mic}}, t)$ for $m \in \{1, \ldots, M\}$. $\mathbf{h}(t)$ can be regarded as a multichannel RIR for a single source signal. We make extensive use of multichannel RIRs in this thesis, they provide better spatial information that is crucial to infer room acoustic parameters, as we shall see in Chapter 4.

Reverting back to Equation (2.3) that specifies the signal model in the discrete time domain, we can observe that with the RIR for a particular room and source-microphone pair, we can simulate a sound signal as if the sound had been emitted in that room. The RIR can hence be seen as a fingerprint of the room and it generally exhibits a defined pattern (Jot and Lee, 2016). As shown in Figure 2.8 a RIR can be divided into 3 parts :

$$h(t) = h^{\mathrm{d}}(t) + h^{\mathrm{e}}(t) + h^{\mathrm{l}}(t) \tag{2.26}$$

- The direct response $h^{\mathrm{d}}(t)$ arrives first and corresponds to the line-of-sight path between the source and the receiver. For omnidirectional source and receivers, it has the strongest energy since it is attenuated only by the medium and delayed due to the distance between the source and the receiver.
- Early reflections $h^{\mathrm{e}}(t)$ refer to the first reflections from obstacles and surfaces in the room. They are sparse in the time domain but carry more energy than later reflections. Still, with a combination of direct response and early reflections, one can derive crucial information about source position, distance, loudness, etc. Early reflections are less in number and sparse in nature. They can be perfectly modeled with the ISM, presented in Section 2.4.2.3.
- Late reverberation $h^{\mathrm{l}}(t)$ consists of later reflections which are often diffuse in nature, producing a diffuse sound field. This results in a random-like response with an exponentially decreasing energy. There exists a point in time (usually defined as 50 ms), where it becomes hard to discern the late reverberation from the early reflections. This is called the *mixing time* of an RIR and divides the RIR into two distinct parts. Reverberation provides the room its individual acoustical attribute giving a very individual sound, but it can also be used to deduce room characteristics such as volume and shape.

RIR simulators have been extensively used in this thesis and there are three main simulation approaches :

- Wave-based methods accurately model the wave equation by relying on space and/or time discretization, but become computationally intractable at high frequencies.
- Geometric methods, based upon deterministic and stochastic approaches, are typically faster, but rely on assumptions that are valid at high frequencies and less valid at low frequencies.
- Hybrid methods: Wave-based and geometric methods are jointly used to simulate RIRs at low and high frequencies, respectively. This switch is based on the mixing time that acts as a cross-over point between the two approaches (Savioja and Svensson, 2015).

A more detailed literature review on acoustic simulators is given in Section 3.4 of chapter 3. In the next section we describe the ISM proposed by Allen and Berkley (1979). The ISM is extensively utilized in numerous acoustic simulators and has been the primary option for simulating large datasets in various applications (see Section 3.5). This thesis revolves

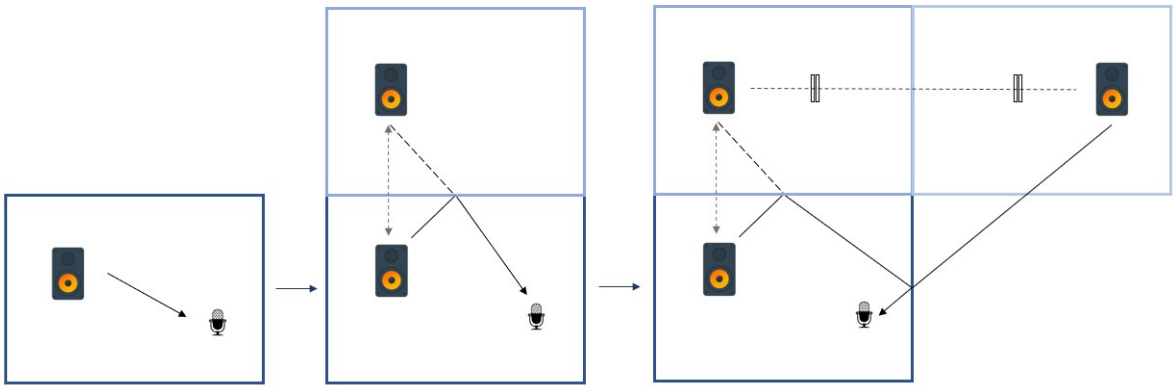

Figure 2.9: Illustration of ISM

around the ISM, which, besides its role in simulation, has also been extended and implemented as part of the *pyroomacoustics* simulator (Scheibler et al., 2018). Therefore, it is essential to delve into the details of the ISM

2.4.2.3 Image source method

The wave theory has been extensively used until now for the description of room acoustics, due to its correctness from a physical point of view. However, the wave theory is not ideal in practical situations, such as simulating a concert hall or for the analysis of room acoustics, particularly due to the high computational complexity involved in solving wave equations at high frequencies and for complex room geometries. A simple and compelling approach is to model sound waves as acoustic rays. Acoustic rays are normal to the wavefront and consist of a direction that follows the propagation law of a light ray, albeit with a different propagation velocity. Each acoustic ray consists of energy without the information of phase, therefore during the presence of multiple sound fields the energies of the rays are added while the phase relation is not taken into account.

Acoustic rays are a fundamental concept of geometric room acoustics. This simple notions allows the practical computation of sound fields. Moreover, diffraction and transmission phenomena are usually ignored in geometric room acoustics as the propagation of rays in a straight line is the major postulate.

The ISM proposed by Allen and Berkley (1979) is a geometric method used to generate RIRs. It is based on the concept of geometric room acoustics. Thus it models specular reflections of acoustic rays in a deterministic approach. This method is highly efficient and accurate for modeling direct responses and early reflections in an RIR. In ISM, every reflection of a ray on the boundary of the enclosed space is modeled by a virtual sound source that is located behind the boundary surface. This virtual source is equidistant from the reflected surface as the original source is from the wall (see Figure 2.9). The direct path

distance between each of the virtual sound source and the microphone is equivalent to the total distance traveled by the acoustic ray inside the enclosed space until it reaches the microphone.

Allen and Berkley (1979) show that the solution to the free-field wave equation with boundary conditions (2.19) and (2.23) for a shoebox room with perfectly rigid walls, so-called *Neumann* boundary conditions, is equivalent to the one given by the ISM method. Additionally, they also prove that in that case, Equations (2.19) and (2.23) can be rewritten as

$$\nabla^2 h(\boldsymbol{r}, t) - \frac{1}{c^2}\frac{\partial^2 h(\boldsymbol{r}, t)}{\partial t^2} = \sum_{k=0}^{\infty} \delta(t)\delta(\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}) \,, \tag{2.27}$$

where the source term in the free-field wave equation becomes an infinite constellation of synchronized point impulse image sources k that replace the boundary of the room, for higher order reflection $k \to \infty$. The solution to the above equation provides the general expression of RIR construction using the ISM

$$h(r, t) = \sum_{k=1}^{\infty} \frac{\delta(t - \|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|_2/c)}{4\pi\|\boldsymbol{r} - \boldsymbol{r}_k^{\text{src}}\|_2} \,, \tag{2.28}$$

and for a particular microphone position $\boldsymbol{r}_m^{mic}$, it can be rewritten as:

$$h(\boldsymbol{r}_m^{\text{mic}}, t) = \sum_{k=0}^{K} \frac{d_{k,m}}{4\pi\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\|_2} \delta\left(t - \frac{\|\boldsymbol{r}_m^{\text{mic}} - \boldsymbol{r}_k^{\text{src}}\|_2}{c}\right) \,, \tag{2.29}$$

where $d_{k,m} \in [0, 1]$ is the attenuation of the image source k at microphone m. It is the product of the absorption coefficients of the boundaries traversed by the path. Equation 2.29 yields an idealized RIR due to many different naive assumptions such as frequency-independent specular walls, omnidirectional point source, and microphones with frequency-independent responses. An extended version of the ISM is presented in Chapter 5, which yields a more realistic RIR.

2.4.2.4 RIR perception and reverberation time

Direct path, early reflection and late reverberation all play different roles in the perception of the sound waves. The direct path helps to reveal the incoming angle of the source. Early reflections provide a sense of geometry of the scene. Most often direct sound and early reflections are correlated. Discerning the delay between them is a challenging task when it is between 5 ms and 40 ms, making them perceived as a single auditory event, as highlighted by Wallach et al. (1973). This effect is also known as the *precedence effect*, and due to this, even in conditions with strong reflections humans can localize a source source (Huang et al., 1997). Early and late reflections are also responsible for the perception of distance and depth, which provide cues for the 3D localization of the source. In the context of virtual or augmented reality systems, accurate simulation of distance and depth is important for the

correct impression of depth (Kearney et al., 2012). At last, late reverberation is responsible for the sound field immersion in the acoustic scene and is mainly characterized by diffuse effects. These effects are related to the size of the room and the material applied to the walls. An acoustic parameter that is associated with it and is often used in this thesis is *Reverberation time*.

The **Reverberation time** (RT_{60}) is defined as the time taken by the reverberant tail of the impulse response to decay by 60 dB. To calculate the RT_{60} of an enclosed space, one needs to measure its RIR. A method to calculate the RT_{60} from a RIR was devised by Schroeder (1979) and is still commonly used today, where the square of the RIR is backward integrated resulting in an energy decay curve, and a linear regressor is fitted on the log of the decay curve to obtain the RT_{60}. Sometimes, due to the unavailability of the full decay up to −60 dB, $RT_{[10,20,30]}$ are used and extrapolated to RT_{60}. While measuring RIRs is a cumbersome process and is highly sensitive to noise, there exists other empirical methods that can approximately estimate the RT_{60} based on the room dimension and absorption properties. One such method is *Sabine's Law*.

$$\mathrm{RT}_{60} \approx 0.161 \frac{V}{\overline{\alpha} S} \ , \ [s] \tag{2.30}$$

where $S = \sum_{v=1}^{v=6} S_v$ is the total area of room's 6 surfaces in $[m^2]$ and $\overline{\alpha} = \sum_{v=1}^{v=6} \frac{\alpha_v S_v}{S}$ is area weighted mean absorption coefficient, V is the total volume of the room $[m^3]$ This formula suggests that reverberation is directly related to the mean absorption, volume, and surface of a room. This intertwined relation motivates the development of method that estimates all these quantities purely based on audio recordings, which is the focus of Chapter 4. However, this formula is mostly valid under diffuse sound field conditions, e.g., in rooms not to far from a cubic shape and whose surfaces are made of similar materials. Larger rooms and experimental reverberation chambers have high RT_{60} up to 10 s, while for typical shoebox rooms such as offices and classrooms the RT_{60} ranges between 0.3 s and 2.0 s. There exists experimental chambers known as anechoics chambers where the RT_{60} is close to zero. Sounds that are recorded in a such environment are referred to as *"dry"* sounds.

2.5 Deep learning

Deep learning refers to a range of machine learning techniques loosely inspired by the human brain (Rosenblatt, 1958; Goodfellow et al., 2016). Deep learning techniques offer a solution to tasks that humans perform intuitively, such as recognizing spoken words (Tjandra et al., 2017) or focusing on a specific speaker in a noisy environment (Malek et al., 2017). These tasks pose a challenge for computers because they are difficult to formally describe. Deep learning offers a way for computers to learn through experience. It involves the acquisition of complex concepts through a hierarchy of simpler ones. This hierarchical representation of concepts gives rise to the term 'deep learning' due to the depth of the concept graph. These algorithms have become popular in recent years, primarily because of two

reasons: the vast availability of different types of datasets and the increased computational power that has enabled the community to solve various problems that weren't possible to solve with pre-deep learning methods. A deep learning system (a.k.a model) is characterized by its set of *parameters*. A *dataset*, is used to determine the best values for the parameters to address the targeted task, e.g., classification, regression, etc. This learning process is called training. In supervised learning techniques, for each example in the dataset, the model aims to estimate the *target* or *label* of that example given a set of attributes called *features* or *inputs*. During the training process, the model improvement can be gauged a *loss-function*, that is a function of the model parameters which quantifies the gap between the model output and the desired output by treating the training dataset as constant. The gradient of the loss function with respect to a given set of input data is calculated using a process known as back-propagation. Various *optimization algorithms* are used to search for the best parameter values which minimizes the loss function. Two such algorithms are stochastic gradient descent and Adam (Kingma and Ba, 2014). In this thesis, we focus on deep supervised learning techniques, where the deep learning models consist of many successive nonlinear transformation layers, which gradually extract higher-level information from the data in a successive manner. Such, deep neural networks (DNNs) are able to learn a complex nonlinear representation of the data.

Throughout this work, we use a combination of different DNN mechanisms such as Convolutional neural network (CNN), multi-layer perceptrons (MLP), and recurrent neural networks (RNN), in addition to different regularization and normalization schemes, learning rate schedulers, skip or residual connections, and many different types of activation functions and many variations of the loss function. This thesis does not provide a detailed description of their working terminologies as it out of scope of this work and throughout this work, we use DNNs as a modeling tool. The interested reader should refer to the textbook by Zhang et al. (2021).

2.6 Conclusion

This chapter provided us with a brief introduction to the digital signal model, sound representations, room acoustics and along with it the functioning and characteristics of different microphones and speakers. The introduced terminologies will be used multiple times during the thesis and are the main protagonists of this work. In the following chapter, we will describe the literature survey on state-of-the-art approaches that are appropriate for this thesis.

3 State of the art

This chapter describes state-of-the-art techniques in the fields of room parameter estimation and sound source localization that are related to the contributions that are described in the thesis. In addition to that, state-of-the-art techniques in the field of virtually supervised learning and a discussion on various room acoustic simulators are also presented. We end this chapter by mentioning a list of existing RIR datasets that have been used to train and evaluate DNN-based methods.

3.1 Room parameter estimation

The relevance of dynamically parameterizing the local acoustic space of listeners by estimating acoustic parameters, such as room geometry, reverberation time, and other related parameters has seen increased interest in the audio processing community (Ick et al., 2023). The information deduced from these parameters can be used to improve the performance of speech processing applications such as dereverberation (Lebart et al., 2001; Habets, 2007) and speech recognition (Couvreur and Couvreur, 2000). Additionally, the acoustic parameters have also been shown to be related to the perceived sound quality (Del Vallado et al., 2013) and intelligibility (Kuttruff, 2016) of the reverberant recordings, and proven to be useful in audio forensics (Moore et al., 2014; Mascia et al., 2015). The estimation of these parameters becomes increasingly crucial in order to generate highly convincing spatial audio in the realm of AAR, which has recently gained ground in the works of Geronazzo et al. (2016), and Yang et al. (2022). Augmented reality systems aim to merge the virtual and real worlds, creating an interactive and improved audio-visual experience by adding artificial elements from the virtual world to the real-world environment of the user. AAR, as a part of augumented reality systems, involves creating perceptually pleasing virtual sound sources in the user's environment. To achieve this, an accurate simulation of the acoustic environment is necessary, thus requiring the estimation of room parameters in the user's environment. One of the major challenges in improving the plausibility of immersive audio is to select the relevant parameters of the acoustic environment in which the users move, from a variety of possible room acoustic parameters.

The term Reverberation Fingerprint coined by Jot et al. (2018) and Ick et al. (2023) refers to a condensed form used to characterize a room for realistic binaural rendering in headphones that implement AAR. The volume of the room V in m^3 and frequency-

dependent reverberation time $\mathrm{RT}_{60}(b)$ together form a reverberation fingerprint (RF). In diffuse-field situations, these parameters are related through Sabine's law. Chapter 4 of this thesis focuses on the estimation of room acoustic parameters. RIRs serve as a room signature because many room parameters can be determined based on the decay curve, early echoes, and direct path from a clean RIR measurement. Furthermore, the availability of accurate and efficient room acoustics simulators (Habets, 2007; Schimmel et al., 2009; Scheibler et al., 2018) allows for a well-defined forward physical problem, where RIRs are modeled by specifying acoustic parameters. Conversely, the inverse problem of estimating acoustic parameters from RIRs is challenging and has recently been the subject of research as is surveyed in the following subsections (Shabtai et al., 2009; Foy et al., 2021). This inverse problem becomes particularly difficult in the context of AAR due to the unavailability of RIRs for the user's specific environment. Additionally, measuring RIRs in a new acoustic environment is a cumbersome and time-consuming task. Therefore, an attractive direction is to blindly estimate acoustic parameters solely on the basis of received microphone signals from unknown sound sources present in the room.

3.1.1 Pre-deep learning methods

One of the most common techniques to measure the absorption coefficients of a specific material in a test is done by using the impedance tube (ISO.ASTM:E1050-9, 2006). This method typically requires a cumbersome experimental apparatus. Reverberation room methods outlined in ISO.354:2003 (2003) provide an alternative approach. It uses reverberation theory and is based on a diffuse sound field assumption. The method requires a controlled reflective environment, which can be achieved using specially designed reverberation rooms. In contrast to these methods, *in situ* measurements are attractive because they do not require complex environments or impedance tubes. In situ measurements can be conducted in any existing environment, such as classrooms or office spaces. Typically, each material is tested separately and the dedicated apparatus is kept at a specific position inside the room to minimize the interference from other walls. Brandão et al. (2015) provide an extensive review of *in situ* approaches used to measure acoustic impedence. However, many of the *in situ* approaches mentioned by Brandão et al. (2015) and Hald et al. (2019) require meticulous control over the environment, excitation signal, experimental setup, and post-processing. This level of control can be challenging to achieve during practical experiments, highlighting the difficulty of this task.

One of the main objectives of the works conducted by Nava et al. (2009), Antonello et al. (2015), Bertin et al. (2016b) and Okawa et al. (2021) is to simplify the acoustic diagnosis process. These methods aim to estimate the absorption profile of all the walls within a room using a small number of measurements. However, these approaches are based on discretization of the wave equation, which is known for its high computational cost and limited effectiveness at high frequencies above 1 kHz. In an effort to address this problem, Dilungana et al. (2022) propose a probabilistic approach. They use the approximate geom-

etry of the room as an input parameter to the algorithm and select the relevant parts of a set of RIRs. The selected parts are used to optimize an objective function in the short-term Fourier magnitude domain.

Estimating RIRs from a pre-defined room geometry has long been a point of interest for the purpose of simulation. The method provided by Allen and Berkley (1979) served as a cornerstone for simulating RIRs in shoebox room with perfectly reflective walls. Building on this, Borish (1984) extended the method to simulate RIRs in polyhedral rooms. However, the inverse problem of estimating room geometry from RIRs remains a challenge. Recovering information about the image sources that contribute to an RIR, such as the image source location, time of arrival, and intensity is deemed to be helpful for room geometry inference (Mabande et al., 2013; Crocco et al., 2016). All these tasks are considered as different parts of a larger problem known as the image source recovery problem (Sprunck et al., 2022). In the literature these problems have been solved using RIRs (Kowalczyk et al., 2013; Crocco and Del Bue, 2016) as well as *blindly*, i.e., using audio recordings of unknown sound sources instead of RIRs (Shlomo and Rafaely, 2021; Tervo and Korhonen, 2010). They are also referred to as *echo-aware* methods (Di Carlo et al., 2021). Additionally, the knowledge of the room geometry can facilitate the recovery of relevant information about the image sources, including the 3D position, using the ISM. Efforts have been made to reconstruct rooms both in 2D (Dokmanić et al., 2011; Antonacci et al., 2012; El Baba et al., 2016) and 3D (Jager et al., 2016; Remaggi et al., 2016; El Baba et al., 2017; Lovedee-Turner and Murphy, 2019; Park and Choi, 2021) from RIRs. As presented by Shih and Rowe (2019), these approaches typically require prior information about multiple source-microphone pairs along with their locations, and the majority of these algorithms estimate the wall locations based on the detection of first order or second order reflections. Alternatively, some work, e.g., by Peng et al. (2015) and Zhou et al. (2017), addresses the problem by using mobile devices and smartphones instead of multiple microphones. Similarly, the work by Shih and Rowe (2019) goes a step further and estimates room geometry without the need to detect all the first-order echoes. With the correct estimation of room geometry calculating surface area and volume of the room is a straightforward task. In our survey, we could not find a method that directly estimates surface area and volume from measured RIRs.

One ongoing research topic is the blind estimation of a room's RT_{60}. Ratnam et al. (2003) were among the first to explore blind estimation techniques using single-channel noisy signals and exploiting the maximum likelihood method. Since then, a plethora of works have been published on blind estimation of RT_{60}. Among these, Gaubitch et al. (2012) investigated three state-of-the-art methods employing different calculation techniques for RT_{60} estimation. They tested these methods on noisy and noise-free datasets and concluded that most of the methods exhibit significant bias in the presence of additive noise in the signal. Several works were subsequently published to address this shortcoming, some of which utilizing machine learning techniques. Thus, to assess and determine the state-of-the-art methods for blind RT_{60} and Direct to Reverberant Ratio (DRR) esti-

mation in single and multichannel scenarios with additive noise, the ACE Challenge was introduced (Eaton et al., 2016).

The ACE Challenge dataset includes recorded RIRs obtained using multi-microphone setups with five different microphone arrays, capturing the acoustic characteristics of seven different rooms. Additionally, the dataset contains anechoic speech and multiple noise conditions recorded in the same acoustic environment to create reverberant and noisy speech signals. The algorithms submitted to the ACE Challenge can be divided into 3 categories: maximum likelihood, spectral decay distribution (Wen et al., 2008), and machine learning with multiple features. The results showed that single-channel estimation of RT_{60} in noisy conditions is a well-established field, with the lowest error rates observed for spectral decay or maximum likelihood methods. Conversely, methods based on machine learning outperformed the others for DRR estimation, indicating that parametric methods could benefit from improved features. The results also highlighted that joint estimators of RT_{60} and DRR were not able to outperform the algorithms that focused on estimating a single parameter. Furthermore, the absence of multichannel RT60 estimators suggests that there is room for further research to exploit the spatial information of audio signals. These results provided an inflection point for future work in machine learning approaches that can alleviate the listed shortcoming.

3.1.2 Deep learning methods

After the ACE challenge, there has been a push to estimate acoustic parameters using machine learning approaches. This is partly due to the success of machine learning methods in other fields such as computer vision and natural language processing. Adding to that, the results on estimating the DRR using machine learning methods showed promising results in the ACE challange, motivating more work in this area. The key issues that persisted after the ACE challenge include the robustness of the model in noisy reverberant conditions, joint estimation of the acoustic parameters, the model performance in a variety of real datasets, and the lack of multichannel approaches to exploit spatial features.

One of the earliest works on blind estimation of RT_{60} using artificial neural networks was published by Cox et al. (2001). Their approach is considered semi-blind and requires re-training the system when there is a change in the acoustic environment. Following the ACE challenge the authors of Lee and Chang (2016, 2018) presented single-channel and multichannel methods for blind RT_{60} estimation using the mel-spectrogram or a combination of mel-spectogram and inter-channel cross-correlation as input features. However, their work was limited to testing on signals convolved with simulated RIRs. Another approach by Gamper and Tashev (2018) showed the advantage of using a CNN and its ability to exploit local and temporal features for this particular task. The model is trained on synthetic RIRs convolved with speech signals and employed a gammatone filterbank feature extractor as the front end. Joint estimation of RT_{60} and DRR has also received some attention in the work of Xiong et al. (2018). However, the performance of these models was

found to be limited due to small and imbalanced training sets. To address the limitations of small datasets, Bryan (2020) proposed data augmentation techniques to generate a balanced dataset from a small collection of real RIRs. The proposed method was tested on a task similar to that of Xiong et al. (2018) and the results demonstrated its superiority over previous state-of-the-art methods for this task on the test set of the ACE challenge. This highlights the importance of training DNN systems with more data for better results. Subsequently, many methods have been developed for single channel RT_{60} estimation, with recent approaches targeting online RT_{60} estimation (Deng et al., 2020) and estimation in dynamic acoustic conditions (Götz et al., 2022).

Apart from the time-energy parameters of an RIR, rendering an acoustic space also requires the estimation of other spatial parameters and boundary information. The efforts for the estimation of these parameters have been minimal in comparison to the abundance of DNN-based approaches for RT_{60} and DRR estimation.

Information about the boundary of an acoustic space can be gathered in the form of estimating the room geometry or reflection/absorption coefficients. One of the first works involved a DNN to jointly estimate 3D room geometry, including the room length, width, height, and reflection coefficients using random RIRs taken in a room (Yu and Kleijn, 2020). The suggested methodology uses a CNN-based architecture, is trained on simulated data, and employs a transfer learning technique on a small dataset of real RIRs to improve generalization. The joint estimation task is formulated as a regression problem with a mean square error loss function. In addition, a blind method has been proposed that reconstructs the room geometry by leveraging the relationship between the direct signal and first-order reflections (Gao et al., 2022). This method captures the signal of arbitrary sound sources using a high-order spherical array. Volume is another important parameter of acoustic spaces and is related to the RT_{60} according to Sabine's law. The estimation of volume from a single-channel noisy speech signal has recently been proposed by Genovese et al. (2019). The proposed DNN is based on multiple convolution layers, and is trained on both simulated and real data. As of our survey, there haven't been any other published works on this task. Due to its ingenuity, this work has been used as a baseline to contrast the contribution on room parameter estimation given in this thesis.

The material used in a wall determines its nature of absorbance towards a sound wave. Absorption coefficients are a metric that explains the level of absorbance of the walls. Absorption coefficients are frequency dependent, which highlights the challenge of estimating them for all the walls of an enclosed surface based solely on a single RIR or a sound source. Limited work has been proposed to tackle this issue using DNNs. To simplify the complexity of this task, authors such as Kataria et al. (2017) and Foy et al. (2021) estimate the surface-weighted mean absorption coefficients

$$\overline{\alpha}(b) = \frac{\sum_v \alpha_v(b) S_v}{\sum_v S_v} \in [0, 1], \tag{3.1}$$

of a room. It is an analytical quantity that effectively summarizes the acoustical charac-

teristics of all six surfaces. According to Sabine's formula (2.30), $\overline{\alpha}(b)$ can be linked to $\mathrm{RT}_{60}(b)$ for the specific frequency band b. Kataria et al. (2017) estimate $\overline{\alpha}$ for walls along with the direction of the source, using binaural audio signals. The system was trained on a training set with little acoustic variabilities, therefore the results were only reported on simulated data. Foy et al. (2021) used a single RIR to estimate $\overline{\alpha}$ in six frequency bands. They obtained errors in a similar range on real data, without the need for a complicated apparatus, controlled conditions, or a prior requirement on the setup of room geometry. However, the system's performance on real data at lower frequencies (< 1000 Hz) was found to be poor. Yu and Kleijn (2020) tackled the different task of estimating an absorption coefficient for each wall in a fixed frequency band. The estimates were obtained by a model trained on single-channel RIRs, while simulated datasets were used for testing the system. Limited representations on the range of absorption coefficients in the training set questions the potential performance of the system on real datasets. Dilungana et al. (2021) estimated the absorption coefficient for all six walls and across six-octave bands using multiple single-channel RIRs in the room. Nevertheless, the lack of testing on real datasets raises the question on the generalization ability of the system.

Interesting future directions towards the task of DNN-based room parameter estimation include audio-visual and end-to-end approaches. Su et al. (2020) portray an end-to-end method to convert a signal recorded in one environment to another. The system takes a waveform as input and generates a latent representation of the environment called room embedding. The room embedding is then used to condition the output of the network in a new environment. Furthermore, some audio-visual methods incorporate vision and echo information to infer the depth and floor map of a 3D scene (Christensen et al., 2020; Purushwalkam et al., 2021). These directions open up exciting possibilities for advancing DNN-based room parameter estimation by integrating different modalities and exploring the potential of end-to-end learning.

3.2 Sound source localization

Apart from room parameter estimation, this thesis also addresses the task of localizing a sound sources in an acoustic scene using the audio captured by an array of microphones. This task is commonly referred to as *sound source localization*. Sound sources can be localized in 1D, 2D, or 3D space. Most commonly, sound source localization provides the Direction of Arrival (DOA), i.e., the angle at which the sound arrives with respect to the microphone array. In the case of a linear microphone array a 1D angle of arrival $\theta \in [0°, 180°]$ is estimated, while for non-planar microphone array configurations a 2D angle comprising an azimuth $\theta \in (-180°, 180°)$ and an elevation $\phi \in (-90°, 90°)$ can be estimated. The DOA can be estimated for single or multiple active sources in the acoustic scene. In addition to DOA estimation, it is also possible to identify the 3D position of a source with respect to a specific point in space using Cartesian coordinates (Shimada et al., 2021; Emmanuel et al.,

2021). However, estimating the distance to a source is a more challenging problem that has received less attention (Yiwere and Rhee, 2017; Bologni et al., 2021).

Although sound source localization and room parameter estimation are considered as independent tasks in this thesis, there is some overlap between them. One example of such overlap is in *Simultaneous localization and mapping* (SLAM) (Evers and Naylor, 2018), where the system simultaneously maps the 3D positions of objects of interest in the surrounding area while localizing an unknown moving observer. While echo-aware methods have often been used independently for estimating room geometry (Dokmanić et al., 2011) or localizing sound sources (Di Carlo et al., 2019), the work of Kreković et al. (2016) simultaneously estimates both for the purpose of SLAM.

Apart from SLAM, sound source localization as an independent technique has numerous applications. Several signal processing tasks, including power spectral density estimation, beamforming, and spatial coding, require accurate estimates of the DOA of active sources as a pre-requisite. These tasks play a pivotal role in various practical applications such as human-robot interaction (Kagami et al., 2008), source separation (Gannot et al., 2017), speech recognition (Busso et al., 2005), bio-diversity monitoring (Chu et al., 2009), smart home applications (Crocco et al., 2016) and search and rescue robots (Nakadai et al., 2017).

Algorithms for estimating the DOA of unknown sound sources can be categorized into two groups. The first category is signal-processing based methods, which were developed prior to the advent of machine learning methods and rely on conventional signal-processing techniques. Signal-processing based methods offer physical interpretability and can work online without requiring training data. Consequently, they are frequently utilized as a feature extraction layer in many DNN-based DOA estimation pipelines. However, it has been noted that they often struggle in noisy and reverberant conditions and fail to generalize in real-life test scenarios. Some examples of signal-processing based algorithms include Generalized Cross-Correlation with Phase Transform (GCC-PHAT) and Multiple Signal Classification (MUSIC). A detailed review of signal-processing based sound source localization algorithms was presented by DiBiase et al. (2001). The second category consists of machine learning based algorithms, which differ from SP-based algorithms in that they require a lot of training data. DNN-based algorithms have demonstrated better performance in challenging conditions, as mentioned by a plethora of published research (Grumiaux et al., 2022). However, generalization to unseen conditions is still a challenge. The following parts describe different signal-processing based methods and then state-of-the-art DNN-based approaches for DOA estimation.

3.2.1 Signal-processing-based methods

Signal-processing-based methods for sound source localization can be grouped into four categories: Time difference of arrival estimation (TDOA), beamforming, subspace methods, Gaussian Mixture Model (GMM) and compressive sensing.

The TDOA between two signals can be estimated by GCC-PHAT. The IDFT of the weighted cross power spectrum (CPS) between the signals from two microphones can be expressed as

$$\ell_{1,2}(\tau) = \sum_{f=0}^{F-1} \frac{X_1(f)X_2(f)^*}{|X_1(f)X_2(f)^*|} e^{j\,2\pi f\tau/N}\,, \tag{3.2}$$

where the CPS is defined by $X_1(f)X_2(f)^*$ ($\star$ denotes complex conjugate). The phase of the CPS is perfectly linear as a function of frequency in free-field conditions that is without noise and reverberation. The TDOA

$$\hat{\tau} = \operatorname{argmax} \ell_{1,2}(\tau)\,, \tag{3.3}$$

is estimated as the peak of the GCC-PHAT angular spectrum $r_{1,2}$. The SRP-PHAT algorithm extends GCC-PHAT to arrays of more than two microphones by taking advantage of multiple microphone pairs. SRP-PHAT uses beamforming to create an acoustic power map $A(\tilde{\mathbf{r}})$ over all possible directions on a regular grid with spatial coordinates $\tilde{\mathbf{r}}$. The local maxima of the power map indicate the presence of a sound source.The power map $A(\tilde{\mathbf{r}})$ is obtained by averaging GCC-PHAT across all microphone pairs in the array :

$$A(\tilde{\mathbf{r}}) = \sum_{m_1=1}^{M} \sum_{m_2=m_1+1}^{M} \ell_{m_1,m_2}(\tau_{m_1,m_2}(\tilde{\mathbf{r}}))\,, \tag{3.4}$$

where the CPS between each pair of microphones m_1, m_2 is calculated for the delay $\tau_{m_1,m_2}(\tilde{\mathbf{r}})$ corresponding to each spatial position $\tilde{\mathbf{r}}$ on the grid. However, computing power maps $A(\tilde{\mathbf{r}})$ is a computationally expensive process due to the involvement of a grid search.

Another class of estimation techniques is subspace methods. These methods operate on multichannel signals with the assumption of correlated source signals and uncorrelated noise signals between the channels. One algorithm in this category is MUSIC, which leverages this assumption by performing eigenvalue decomposition on the multichannel signal matrix to separate the signal and noise subspaces. Based on these subspaces, an angular spectrum function is constructed, which is then examined in all directions to detect the presence of an active source. Another algorithm called ESPRIT directly estimates the source DOA by exploiting the structure of the source subspace. It provides a faster alternative to MUSIC but may result in slightly less accurate DOA estimates. While DOA estimations based on subspace algorithms are relatively robust to noise, reverberant conditions are a challenge (Moore et al., 2016).

Many estimation algorithms also employ GMM. Authors such as Madhu and Martin (2008), Schwartz and Gannot (2013) have made consistent efforts in advancing this approach. In most of these models, the algorithm's parameters are estimated "online" during the test phase while localizing sound sources. These methods exploit the sparsity of sources in the time-frequency domain by maximizing the data likelihood function using expectation maximization algorithms. Finally, methods based on compressive sensing convert a

high-dimensional signal to a low-dimensional representation using either sparse synthesis (Candes et al., 2006) or sparse analysis (Nam et al., 2013). In the context of sound source localization, this translates to a sparsity assumption in the spatial domain, which can be solved using Bayesian or greedy methods (Kitić et al., 2014). This results in a remarkable performance for localization tasks.

In the realm of multisource DOA estimation systems, Bechler and Kroschel (2003) attempted to utilize a peak-picking method based on certain threshold levels applied to the spatial spectrum generated by signal-processing-based algorithms. In the same study, it has been noted that the spatial spectrum gets corrupted in the presence of noise and reverberation in the signal, which results in errors and degrades the accuracy of the system. Nevertheless, Tho et al. (2014) made progress in estimating the angular spectrum in the presence of noise and demonstrated the ability to mitigate the effects of reverberation for multisource DOA estimation algorithms.

3.2.2 Machine learning methods

Methods based on GMM such as Madhu and Martin (2008), and Schwartz and Gannot (2013) presented in the previous section do not heavily rely on extensive training data, as they estimate the GMM parameters online from the observed data. By contrast, some sound source localization algorithms, such as the one of Woodruff and Wang (2012), based on a GMM, use training data for pre-estimating the parameters. A similar approach was taken by Deleforge and Horaud (2012), who presented a variant of GMM known as Gaussian mixture regression for single-source localization. This approach was further extended to multi-source localization by Deleforge et al. (2015). Both models were trained on a dataset made of synthetic RIRs convolved with noise signals, demonstrating a close connection to DNN-based methods. Kim and Ling (2011) were among the first to use a DNN for sound source localization. They introduced a multi-task network based on an multi-layer perceptron architecture capable of performing multi-source counting and localization. The evaluation of the system was done on reverberant data, while training was performed on anechoic data. This work served as an inspiration for numerous subsequent studies, leading to a plethora of DNN-based localization systems (Grumiaux et al., 2022). These systems can be further categorized according to their input features, architecture and outputs.

Input features The majority of DNN-based sound source localization methods are designed for speech sources since there exists a strong connection between speech enhancement, speaker diarization, and separation (Gannot et al., 2017; Vincent et al., 2018). Systems such as those of Chakrabarty and Habets (2019),Hao et al. (2020), Grumiaux et al. (2021b) and He et al. (2021) are specialized in localizing speech sources. However, with the introduction of the SELD task in the DCASE challange (Politis et al., 2020), a range of systems have emerged that can successfully localize non-speech sources such as barking dogs, door-knocking, alarms, and more. When it comes to feature extraction from the input, some networks such as those of Suvorov et al. (2018) and Jenrungrot et al. (2020) directly work

on raw waveforms using an *end-to-end* approach. However, this approach requires more computational power, leading to larger network sizes and the need for extensive training datasets. The majority of the DNN-based sound source localization systems are smaller in size compared to the networks seen in other fields such as NLP. Therefore, in most systems, handcrafted interchannel features, spectrograms (Chakrabarty and Habets, 2019; Bohlender et al., 2021), or signal-processing-based features like GCC-PHAT and MUSIC (He et al., 2019; Vesperini et al., 2016) are used as input. Common interchannel features include relative transfer functions (Chazan et al., 2019; Bianco et al., 2019) and binaural features such as Interchannel Phase Difference (IPD) and Interchannel Level Difference (ILD) (Ma et al., 2015). Features based on Ambisonics signal have also been recently exploited, as they effectively represent the spatial properties of the sound field. Systems such as those of Adavanne et al. (2018b, 2019) use FOA spectrograms as input features, while Poschadel et al. (2021) and Varanasi et al. (2020) employ third-order ambisonic spectrograms. Perotin et al. (2018) progressed with a similar idea and used intensity-based ambisonic features referred to as *ambisonics pseudo-intensity vectors*, which showed superior performance compared to a system based on raw ambisonic waveforms.

Architectures Numerous DNN architectures have been proposed for sound source localization. Efficient architectures and complex models have been proposed that have drawn inspiration from other domains, leveraging their demonstrated performance on various types of signals. Architectures based on convolutional neural networks (Chakrabarty and Habets, 2019; He et al., 2019), recurrent neural networks (Wang et al., 2018) and a combination of convolutional and recurrent networks (Adavanne et al., 2018b,a) are the most common to be found in the literature. Recent additions to the field include networks with residual layers (Pujol et al., 2019), attention mechanisms (Grumiaux et al., 2021b), and encoder-decoder networks (Chazan et al., 2019; Bianco et al., 2019).

Output Sound source localization systems can be classified into *single-source* or *multi-source* localizers, capable of localizing single or overlapping speech or sound events. There is a substantial body of literature focused on single-source estimation, as it represents the simplest scenario. Significant progress has been made in developing systems that can localize sources in reverberant environments while simultaneously performing source activity detection (Yalta et al., 2017; Perotin et al., 2018; Bologni et al., 2021). While multi-source localization presents a more challenging problem current systems have demonstrated the ability to localize sources even in adverse acoustic conditions (Ma et al., 2015; Perotin et al., 2019; He et al., 2021). In multi-source systems, the detection of activity from multiple sources becomes a source-counting problem (Grumiaux et al., 2021a), which leads these systems to be termed as multi-tasked. The SELD task of the DCASE challenge is based on this idea, where the submitted systems must address sound event detection and localization jointly. Another example of multi-task sound source localization system that also performs dereverberation is that of Wu et al. (2021). Most of the sound source localization systems mentioned so far can be classified as either classification-or regression-based systems, which

localize active sources using Cartesian or spherical coordinate systems. In order to have a robust system the training data need to be sufficiently diverse to include most directions in the space. There also exist DNN-based methods that do not directly estimate the DOA but instead pass the high-level processed features to another conventional algorithm for DOA estimation (Pertilä and Cakir, 2017).

The variety and the amount of work done in the field of DNN-based sound source localization systems is enormous, making it impossible to cover all of the work within the scope of this thesis. Interested readers are recommended to refer to the comprehensive study published by Grumiaux et al. (2022) for a more extensive review.

3.3 Virtually supervised learning

Developing a robust DNN-based system that generalizes to unknown conditions requires a large amount of labeled data. However, acquiring labeled data in various conditions for domain-specific applications is costly and time-consuming, often resulting in a shortage of high-quality real data across different domains. As a solution, the concept of training DNN systems with simulated data has emerged, as annotations can be automatically generated. Transfer learning from simulated data to real data has been employed in various fields, leveraging the availability of reliable simulators. This technique is used to increase the size of training data in applications such as computational biology (Behboodi and Rivaz, 2019), financial market predictions (Maeda et al., 2020), geosciences and remote sensing (Malmgren-Hansen et al., 2017) to computer vision applications such as scene classification (Bird et al., 2020) and autonomous driving (Pan et al., 2017). In the field of audio, the term *virtually supervised learning* was first introduced by Gaultier et al. (2017) who applied this concept to sound source localization, demonstrating effective mapping using a virtually learned model on real sound signals. The same study also proposed a large simulated dataset for training sound source localization models. Simulated training data has been widely used in audio applications to learn mappings between audio features and properties. Examples include automatic speech recognition (Huang and Bocklet, 2019), sound source localization (Grumiaux et al., 2022), speech enhancement (Gannot et al., 2017), and diarization (Horiguchi et al., 2020) systems. Simulated acoustic datasets have also found applications in navigation systems (Chen et al., 2020), floorplan reconstruction (Purushwalkam et al., 2021), and many other areas. All the presented applications use a dataset of simulated RIRs or simulate RIRs on the fly using acoustics simulators.

However, there still exists a gap in realism between synthetic and real data distributions. Improving simulators to produce more realistic examples is a way to close this gap, which will help the DNN generalize to real-world conditions. However, improving simulators often comes with added computational overhead. One research direction is to leverage DNNs to improve simulation and achieve more realistic results. Richter et al. (2022) address similar objectives by employing Generative Adversarial Networks (GANs) (Creswell

et al., 2018) to enhance the realism of synthetic photos generated by graphic engines.

Improving realism in acoustics simulators will directly affect all DNN-based audio applications. Currently, a wide variety of acoustics simulators are available, offering different levels of realism at varying computational expenses. Given the substantial amount of simulated data required for DNN-based systems, there is a need for a fast and realistic acoustics simulator. While some studies, such as Diaz-Guerra et al. (2021) and Ratnarajah et al. (2021), demonstrate the efficacy of utilizing Graphic Processing Unit (GPU) accelerators and Generative adversarial network (GAN)s to generate a substantial number of RIRs, they may not preserve all acoustic characteristics and may lack realism. One of our contributions in this thesis improves the realism of a renowned fast acoustics simulator (Scheibler et al., 2018) with minimal computational overhead. To validate its effectiveness in improving the generalizability of trained models, we focus on tasks such as room parameter estimation (see Chapter 4) and sound source localization (see Chapter 5).

3.4 Room acoustics simulators

Modeling room acoustics computationally requires physics-based simulation of the propagation of sound in a digital acoustic scene. This type of simulation is particularly useful for acoustic designers as it enables them to evaluate the acoustical design of a room using auralization and numerical acoustic data (Siltanen et al., 2010). Room acoustics simulator (RAS) comprise two main kinds of approaches: Wave based methods and geometrical acoustics.

3.4.1 Wave-based simulation

Wave-based methods aim to numerically solve the wave equation in order to determine the room impulse response (RIR). This approach provides an accurate simulation of room acoustics but comes with a high computational cost, which increases with frequency. One common strategy is to discretize the bounding surface or the space into small components while simulating the interaction between these components. The interaction between components is calculated by solving the wave equation with appropriate boundary conditions. Two commonly used methods for solving the wave equation are the Finite Element Method (FEM) and the Finite Difference Time Domain (FDTD) method. Fast calculations of the solution are possible for low frequencies of sound waves in smaller rooms. However, the opposite is true for intricate room shapes and high-frequency sound waves. Consequently, the practical use of wave-based methods is limited to modeling low-frequency RIRs up to the Schröder frequency. Software such as ODEON and a recent GPU-based implementation of FDTD (Hamilton, 2021) are examples of simulators capable of performing wave-based simulations. Mentioning all the wave-based software and simulators is out of the scope of this thesis. Interested readers are referred to the surveys by Svensson and

Kristiansen (2002) and Siltanen et al. (2010).

3.4.2 Geometric-acoustics based simulation

Geometric acoustics works on the basis of considering sound as rays instead of waves, disregarding the wavelength of sound. In this approach, rays or particles are used to represent sound reflections off room surfaces. These methods analyze sound energy rather than relying on sound pressure or particle velocity. An RIR is divided into three parts (see Chapter 2), where simulation of the direct path and early reflections should be precise with timing and phase information, which is crucial for sound source localization. The ISM is often employed to do so due to its ability to efficiently determine all specular reflections using a cluster of image sources. While this method is straightforward, as each reflected ray arriving at the receiver can be regarded as a delayed attenuated function of the Dirac deltas, it also has some caveats. The ISM struggles in modeling higher-order reflections, due to cubical increase in the required number of image sources. Additionally, Bork (2000) notes that the ISM cannot effectively simulate obstacle scattering. Schroeder (1987) suggests a solution to this issue by using statistical methods to model late reverberant tails, particularly for higher frequencies and larger rooms. Stochastic Ray Tracing (SRT) methods such as diffuse rain (Schröder, 2011) and other radiance transfer methods (Siltanen et al., 2010) can be used for this purpose.

The ISM and SRT can be used independently to simulate an RIR, but each method has its own disadvantages. The ISM requires careful selection of the order of image sources, while SRT lacks phase information as it focuses solely on energy. By combining them, a hybrid geometric acoustics simulator can be created. These simulators offer the advantages of both the ISM and SRT: the ISM can model phase effects in the early part of the RIR, while SRT can handle diffuse scattering. However, these simulators have limitations in modeling diffraction and other low-frequency effects around surfaces and objects in space. Despite this, the hybrid geometric acoustics approach provides sufficient detail for specific applications while maintaining computational efficiency.

3.4.3 Room acoustics simulation libraries

Many simulators based on geometric acoustics have been proposed, and among them, simulators following the ISM have gained popularity in various applications (Gannot et al., 2017; Arberet et al., 2010), as they provide sufficient realism related to the requirement of targeted applications while maintaining computational efficiency. Furthermore, a large number of simulators are available as open-source libraries. One such library is the Roomsim toolbox by Campbell et al. (2005), implemented in MATLAB. It has undergone improvements, such as in Roomsimove [1], which incorporated support for moving sources, and

[1]Developed by Emmanuel Vincent https://members.loria.fr/EVincent/software-and-data/

by Schimmel et al. (2009), where SRT computation was added, transforming it into a hybrid geometric acoustics simulator. Subsequently, Barumerli et al. (2021) created a Python binding and extended its functionality to load HRTFs for generating Binaural Room Impulse Responses. Other widely used MATLAB-based libraries include the RIR generator by Habets (2006), the spherical microphone impulse response generator (Jarrett et al., 2012), and the multichannel simulator for arbitrary microphone arrays (MCRoomSim) by Wabnitz et al. (2010). Schissler et al. (2014) proposed an HGA simulator capable of handling high-order diffraction, and its Python binding is available as PYGSOUND [2].

With the advancement of DNN based systems whose libraries are mostly found in Python, there is a demand for Python-based room acoustics simulators. Scheibler et al. (2018) addressed this need by introducing pyroomacoustics, an open-source library with highly modular components. In this thesis, we used it extensively and improved its functionalities by adding support for source and receiver directivity with the help of an extended ISM model presented in Chapter 5.

Here we describe simulators that exploits GPUs for efficient, large-scale simulation of RIRs. Due to the presence of CUDA cores, GPUs are known for performing fast complex calculations at a large scale. Several wave-based simulators have utilized GPUs to provide solutions for FEM and FDTD problems (Raghuvanshi et al., 2009; Röber et al., 2006). Fu and Li (2016) and Diaz-Guerra et al. (2021) suggested leveraging GPUs to accelerate ISM RIR simulations, where the latter also provides an open-source library.

In addition, DNN-based techniques such as the one of Ratnarajah et al. (2021) use GANs for learning to approximate the distribution of real RIRs with the aim of producing realistic RIRs on-the-fly. Similarly, Luo and Yu (2022) generates RIRs on-the-fly and make use of approximate physical modeling of the reflection process and sound propagation. Notably, their approach does not require specific computational equipment such as GPUs. Although all these methods are good for generating huge RIR datasets on the fly, they lack flexibility and realism, especially in terms of their inability to simulate RIRs in complex acoustic environments, simulating different types of microphone arrays, and modeling directivities for sources and microphones. Hence, the community is in need of a simulator that is fast and generates realistic RIRs. The work shown in Chapter 5 is an attempt to bridge this gap.

3.5 RIR and audio datasets

Datasets of real labeled data measured in a variety of situations are important for the assessment of signal-processing/DNN-based room parameter estimation and sound source localization systems. Collecting real RIRs on a large scale is not trivial. Fortunately, there is a wide range of freely available RIR datasets that cater to specific application requirements.

[2] https://github.com/GAMMA-UMD/pygsound

3.5.1 RIR datasets

Several datasets were specifically collected to capture RIRs in various realistic conditions. One of these datasets is the DTU three-channel dataset (Fernandez-Grande et al., 2021), which consists of 152 RIRs measured using a three-channel array in a single room. This dataset is intended for DOA estimation. Another multichannel dataset is the BIU impulse response database (Hadad et al., 2014), which can be used for applications such as speech separation, DOA estimation, and speech enhancement. The BIU dataset includes 1,800 RIRs measured at three different reverberation levels using uniform linear arrays (ULA). The protocol and modular acoustics room established for the BIU impulse response dataset were also used by Di Carlo et al. (2021) for the dEchorate dataset. This multichannel dataset was created using six microphone arrays, each consisting of five microphones. The scenes were excited with six loudspeakers, and the dataset was measured in 11 different acoustic conditions created within the same shoebox room, resulting in 1,800 RIRs. The primary purpose of the dEchorate dataset is to test algorithms for acoustic echo retrieval, room parameter estimation, and echo-aware signal-processing methods. The three multi-channel datasets of Dokmanić et al. (2011),Crocco et al. (2016), and Remaggi et al. (2016) were gathered only for the purpose of room geometry estimation. These datasets lack proper annotation of source positions and scattered microphones throughout the rooms, making them unsuitable for DOA estimation and speech enhancement tasks. Arni (Prawda et al., 2021) is a recently measured RIR dataset that also employed a modular acoustic room with 5,312-panel combinations. It captures single-channel RIRs and aims to investigate the change in acoustic parameters such as RT_{60} and speech clarity in relation to variations in wall absorption coefficients. The MIT IR survey (Traer and McDermott, 2016) is also a single-channel dataset consisting of 271 RIRs measured in 10 different rooms.

The BUT Reverb DB (Szöke et al., 2019) and the RWCP sound scene database (Nakamura et al., 1999) are datasets specially designed to test multiple audio processing systems such as sound source localization, speech enhancement, sound retrieval, and automatic speaker recognition. The BUT reverb DB consists of RIRs, environmental noise, and recorded excerpts from the Librispeech corpus played by loudspeakers. The recordings were made in 8 rooms of varying sizes with 31 microphones divided into various microphone arrays placed in each room. The sources were positioned randomly at 5 different positions. The RWCP database includes similar contents recorded in 9 different rooms with phonetically balanced words instead of speech excerpts. From all the described datasets we specifically use the dEchorate dataset Di Carlo et al. (2021) for our expirements in Chapter 4 and 6 of this thesis.

3.5.2 Binaural room impulse response (BRIR) datasets

BRIR datasets aim to capture room acoustics with binaural effects. Datasets such as the Aachen room impulse response and the Surrey BRIR dataset were measured using

a binaural microphone array where microphones are positioned in the ears of a Head and Torso Simulator (HATS) to incorporate the filtering effect of a human head. The former was measured in 6 different rooms with random source and HATS positions; it was aimed to test speech enhancement algorithms. The latter also used a HATS microphone array but measured RIRs in four different rooms. Both datasets have been used for sound source localization (Venkatesan and Ganesh, 2017; Ma and Liu, 2019).

3.5.3 Smart-home datasets

Another category of datasets focuses on smart home applications. We describe three datasets namely DIRHA (Cristoforetti et al., 2014), VOICEHOME 1 and 2 (Bertin et al., 2016a, 2019) and SWEET-HOME (Vacher et al., 2014). The main objective of these datasets is to provide multi-channel speech recordings, primarily spoken by humans. All three datasets were recorded in a domestic environment. The DIRHA corpus consists of recordings made in the kitchen and living room of an apartment. Microphones are placed in various configurations throughout the space, and speech is delivered by six native English speakers. VOICEHOME 1 and 2 present a series of two corpora that focus on short commands and dialog scenes. The recordings involve 12 native French speakers in 12 rooms distributed across four houses. Similarly, the SWEET-HOME dataset records 26 hours of speech utterances in French, captured in a flat with four rooms. All of these datasets are annotated with positions of source and microphone arrays combined with the geometry of the scene. However, the publicly available versions of DIRHA and SWEET-HOME do not include measured RIRs. The applications of these datasets are targeted towards sound source localization, automatic speech recognition, and speech enhancement. The DIRHA and VOICEHOME2 datasets are used in the experiments of Chapter 6 of this thesis.

3.5.4 Audio challenge datasets

Over the years, many challenges have been organized to serve different applications, and this has led to the emergence of real datasets that are specifically generated to evaluate the submitted systems. The ACE Challenge (Eaton et al., 2016) aims to find the best system for blind acoustic parameter estimation. The evaluation dataset consists of real noisy reverberant speech signals. RIRs are measured in 7 rooms using devices such as mobile phones, notebooks, and an Eigenmike. The LOCATA (Evers et al., 2020) and DCASE speech event localization detection (SELD) challenges (Politis et al., 2020) were designed for multi-task single and multiple sound source localization tasks. The LOCATA challenge includes source tracking with sound localization detection. The dataset includes source/microphone scenarios in static or moving conditions. The signals have been captured using various types of microphone arrays, namely planar arrays, Eigenmikes, and a microphone array positioned on a robot head and recordings are done in only one room. The DCASE SELD challenge includes an extra task of sound event detection with source localization.

The evaluation dataset of the latest DCASE SELD challenge is named as STARSS22 and consists of a naturally acted scene where the speakers are allowed to move and orient themselves naturally (Politis et al., 2022). Apart from the dominant speech signals, there are 13 different events, (i.e., non-speech sources) present in the scene that are active at different times. Furthermore, all these active events are spatio-temporally annotated. A considerable amount of diffuse and directional noise sources are also present in the recorded signals making this dataset challenging. The signals in the STARSS22 datasets are captured using the Eigenmike and presented in two formats: a four-channel signal from a tetrahedral sub-array or FOA, while the recording is done in two different reverberant conditions. Both datasets use optical tracking systems (OPTI-TRACK) for accurate annotations of the positions. The STARSS22 dataset is used in the experiment of Chapter 6 of this thesis.

3.5.5 Audio-visual datasets

There are several audio-visual datasets available that include video footage of acoustic scenes along with recorded audio. Examples of such datasets include AV 16.3 (Lathoud et al., 2005) and CHIL (Stiefelhagen et al., 2008) which have been utilized in audio-visual localization and tracking systems. Additionally, there are audio-visual experiments that focus on estimating RIRs through the use of acoustic scene images (Singh et al., 2021). Real datasets such as Matterport 3D (Chang et al., 2017) and Open-Air (Murphy and Shelley, 2010) offer RIRs for multiple acoustic scenes in addition to images.

3.5.6 Synthetic datasets

The requirement to close the gap between simulated data and real data has led to many developments in advanced simulator systems as seen in Section 3.4. Although the generation of large simulated datasets consumes power and requires ample time, due to ample demand for simulated RIRs in many DNN-based audio applications, many simulated datasets have been introduced.

GWA (Tang et al., 2022) and Soundspace (Chen et al., 2020) are two highly realistic synthetic RIR datasets. GWA (Tang et al., 2022) provides 2 million RIRs, which are simulated using a complex pipeline that incorporates realistic 3D house models and a matched database of absorption profiles. The authors employed a combination of FDTD and SRT methods to generate high-quality RIRs, requiring approximately 1,300 CPU/GPU hours for the simulations. Models trained on this dataset performed well on many single-channel tasks and outperformed models trained on less realistic simulated datasets. The Soundspace dataset (Chen et al., 2020) also utilizes 3D models of indoor spaces with wall profiles obtained from a matched database. However, the simulation of RIRs in Soundspace is achieved through SRT techniques, resulting in RIRs that are less realistic compared to GWA. BIRD (Grondin et al., 2020) and VAST (Gaultier et al., 2017) are other such datasets made along the lines of Virtually Supervised Learning. Both are less realistic than GWA

and SOUNDSPACE. The BIRD dataset provides 100,000 multichannel RIRs that are simulated using the ISM technique in rooms with random-profile wall absorption coefficients. The VAST dataset simulates 110,000 binaural RIRs using the hybrid geometric acoustic method in a realistic indoor acoustic environment, with carefully selected absorption coefficients that represent commonly encountered surfaces in indoor settings. All these datasets fail to generalize to multi-channels tasks with the demand of a specific microphone array.

3.6 Summary

This chapter begins by conducting a literature review on room parameter estimation and sound source localization which are the two primary tasks focused in the following chapters. Subsequently, we introduced virtually supervised learning for audio and then gave an overview of its two main components room acoustic simulators and diverse audio datasets. The contributions outlined in Chapter 4,5 and 6 draw inspiration from and often reference the research presented in various sections of this chapter.

4 Multichannel room parameter estimation using multiple viewpoints

This chapter provides a detailed description of our contribution to the task of blind room parameter estimation. We investigate the joint estimation of room acoustic parameters that are crucial to reverberation fingerprint and are related to each other using Sabine's law in diffuse field conditions. These parameters are the room's volume $V[\mathrm{m}^3]$, surface area $S[\mathrm{m}^2]$, reverberation time $\mathrm{RT}_{60}(b)[\mathrm{s}]$ and absorption coefficient $\alpha_v(b)$ for all surfaces v. Six octave bands were used in this study, namely, $b \in [125, 250, 500, 1000, 2000, 4000]$ kHz. The estimation of $\alpha_v(b)$ for all surfaces and octave bands increases the complexity of this task by manyfold. Instead, we estimate the room's mean absorption coefficients $\overline{\alpha}(b)$ calculated with Equation (3.1). With the aim to solve an inverse problem, we estimate these parameters blindly using a multi-microphone setup with randomly positioned speech sources in the acoustic scene. Further, we present a novel DNN to efficiently use the data captured by multiple microphones, and carve out techniques to fuse multiple observations (source-receiver pairs) in the same room. The network is trained with virtually supervised learning on a training dataset that is simulated using a hybrid geometric acoustics simulator. We propose to study the problem of room parameter estimation in two parts. This chapter focuses on the first part, i.e., establishing a state-of-the-art method for room parameter estimation, while Chapter 6 focuses on the effect of the simulation strategy on the room parameter estimation performance. This chapter is divided into three parts. Section 4.1 explains the simulation of training data, Section 4.2 introduces the novel DNN network and the training procedures while Section 4.3 presents different experiments and their results on simulated and real data, the chapter concludes with Section 4.4.

4.1 Training data

4.1.1 RIR simulation

To train a blind room parameter estimation system, a large dataset of noisy, reverberated speech signals properly annotated with room acoustic parameters is required. We first describe how diverse and realistic RIRs were generated. In this chapter, we use a room

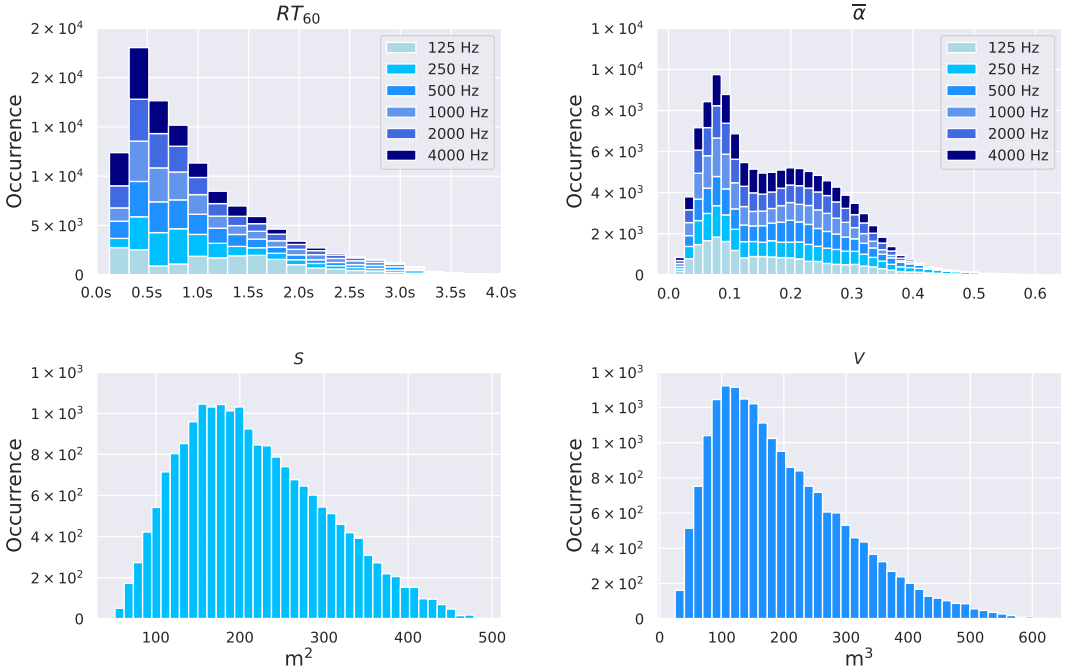

Figure 4.1: Distribution of RT_{60}, $\overline{\alpha}$, S and V in the simulated training set. A stacked histogram is used for RT_{60} and $\overline{\alpha}$ to show data in six-octave bands.

acoustic simulator called *Roomsim*, developed by Schimmel et al. (2009). Roomsim is a hybrid geometric acoustics simulator that uses ISM and a stochastic ray tracing method to simulate specular and diffuse reflections. The Roomsim simulator uses the diffuse rain technique for the realization of stochastic ray tracing computation (Schröder, 2011). Simulations are run using a sampling frequency of 48 kHz, an image source order of 10 and 2,000 rays for stochastic ray tracing computations. Simulations were conducted for 20,000 different shoebox rooms whose length, width, and height are uniformly sampled from a range of $[3, 10]$ m, $[3, 10]$ m, and $[2.5, 6]$ m respectively. This resulted in $S \in [48, 440]$ m^2 and $V \in [18, 600]$ m^3. Five random source-microphone array positions were simulated in each room (a.k.a observations). The microphone array considered in our study consists of two microphones at a distance of 22.5 cm.This distance resembles closely to a head or a headset width. The source and microphones were simulated as omnidirectional. The source and microphone array are kept at least 30 cm apart from each other and from each wall of the room to avoid near-field artifacts.

To have realism in the acoustic scene of the sampled rooms, the walls of the shoebox rooms need to have frequency-dependent absorption coefficients $\alpha_v(f)$ with $v \in [1, \ldots, 6]$. However, the vanilla ISM described in Equation (2.29) is restricted to work with frequency-independent absorption and reflection coefficients. To incorporate frequency-dependent

	Walls	Floor	Ceiling
125 Hz	[0.01 - 0.50]	[0.01 - 0.20]	[0.01 - 0.70]
250 Hz	[0.01 - 0.50]	[0.01 - 0.30]	[0.15 - 1.00]
500 Hz	[0.01 - 0.30]	[0.05 - 0.50]	[0.40 - 1.00]
1000 Hz	[0.01 - 0.12]	[0.15 - 0.60]	[0.40 - 1.00]
2000 Hz	[0.01 - 0.12]	[0.25 - 0.75]	[0.40 - 1.00]
4000 Hz	[0.01 - 0.12]	[0.30 - 0.80]	[0.30 - 1.00]

Table 4.1: Ranges of $\alpha_v(b)$ used for the reflectivity-biased strategy (Foy et al., 2021).

coefficients, a generalized Fourier-domain formulation of Equation (2.29) is written as

$$H(\mathbf{r}_m^{\text{mic}}, f) = \sum_{k=0}^{K} \frac{D_{k,m}(f)}{4\pi \|\mathbf{r}_m^{\text{mic}} - \mathbf{r}_k^{\text{src}}\|_2} e^{-i2\pi f \frac{\|\mathbf{r}_m^{\text{mic}} - \mathbf{r}_k^{\text{src}}\|_2}{c}}, \tag{4.1}$$

where $D_{k,m}(f)$ is the total wall attenuation of the k^{th} image source defined in the frequency domain. The absorption properties of the materials used in buildings are generally provided by the manufacturers in the form of absorption coefficients given as values in $[0, 1]$ per octave band. Hence, there is a need to interpolate these coefficients in the Fourier domain and to add a phase response to walls. Roomsim achieves this by linear interpolation of the coefficients and a minimum-phase wall-response design.

To generate our training set, the absorption coefficients for six surfaces in six octave bands are sampled using the *reflectivity-biased* sampling strategy described in Foy et al. (2021). This strategy provides each wall with a 50% chance of either being a frequency-independent reflective wall or a frequency-dependent absorbent wall, where values for the reflective wall profile are uniformly drawn at random in the range $\alpha_v(b) \in [0, 0.12]$, while the values for the absorbent wall profiles depend on the surface type and are uniformly drawn at random from the ranges shown in Table 4.1. These ranges are derived from measured databases of surface materials that are frequently encountered in typical buildings. Instead of sampling $\alpha_v(b)$ at random in the range $[0, 1]$, the advanced sampling strategy leads to a realistic distribution of $\text{RT}_{60}(b) \in [0.2, 3.2]$ s and $\overline{\alpha}(b) \in [0.02, 0.6]$ in the simulated dataset, as can be observed in Figure 4.1. The scattering coefficient used for the diffuse-rain stochiastic ray tracing method is drawn randomly in the range of $[0.2, 1]$.

4.1.2 Mixture generation

To generate noisy reverberated mixtures, the simulated RIRs generated for many room configurations are convolved with speech signals and augmented by noise. The computation is as follows,

$$c_m[n] = (h_m * s)[n] + \rho(h_m^{\text{tail}} * \eta_m^{\text{SSN}})[n] + \nu\eta_m^{\text{static}}[n], \tag{4.2}$$

where $m \in [1, 2]$ and h_m are simulated RIRs that are downsampled from 48 kHz to 16 kHz and convolved with randomly taken anechoic speech signals s from the Librispeech corpus (Panayotov et al., 2015). Reverberated speech signals are cropped after first 20 ms to avoid silence to a length of 3 seconds. They are altered by adding diffuse babble noise ($h_m^{\text{tail}} * \eta_m^{\text{SSN}}[n]$) and static microphone noise (η_m^{static}). Diffuse babble noise is generated by convolving speech-shaped noise $\eta_m^{\text{SSN}}[n]$ with a late reverberant tail (after 50 ms) of a random RIR simulated in the same room, denoted by h_m^{tail}. The static microphone noise is an independent white standard Gaussian noise acting on each channel. These noise signals are weighted by coefficients ρ and η whose values are explained in what follows.

Importantly, realistic noise levels exhibit a behavior where the signal-to-noise ratio (SNR) tends to be lower if the source is kept far from the receiver, and vice-versa for the source closer to the receiver. To replicate this we use a *reference signal*. The reference signal is simulated in a typical condition, in a shoebox room of dimensions (5,5,3) with frequency-independent absorption coefficients of 0.2 given to all 6 walls. The same microphone array used in the training dataset is placed at the center of the room, and the source is positioned at a distance of 1 M. The resulting RIR is convolved with a fixed white noise signal. Let the variance, i.e., the power, of the reference signal be σ^2_{ref}, which is calculated by concatenating its channels. For each noisy mixture generated with Equation (4.2), the variances σ^2_{diff} and σ^2_{static} of the diffuse and static noises are then calculated with respect to the reference signal variance according to the following formula :

$$10 \log_{10}\left(\frac{\sigma^2_{\text{ref}}}{\sigma^2_{\text{diff}}}\right) = \text{SNR}^{\text{diff}}, \tag{4.3}$$

$$10 \log_{10}\left(\frac{\sigma^2_{\text{ref}}}{\sigma^2_{\text{static}}}\right) = \text{SNR}^{\text{static}}, \tag{4.4}$$

where SNR^{diff} and $\text{SNR}^{\text{static}}$ are randomly drawn in the ranges $[30, 60]$ dB and $[70, 90]$ dB for every observation inside the room. On the other hand, the power of diffuse babble noise and static noise

$$\sigma^2_{\text{diff}} = \rho^2 \, \text{Var}\left((h_m^{\text{tail}} * \eta_m^{\text{SSN}})[n]\right), \tag{4.5}$$

$$\sigma^2_{\text{static}} = \nu^2 \, \text{Var}\left(\eta_m^{\text{static}}[n]\right), \tag{4.6}$$

where the variance of the standard Gaussian distribution is equal to 1, hence $\sigma^2_{\text{static}} = \nu^2$. Solving for ρ and ν we obtain :

$$\rho = \frac{\sigma_{\text{ref}}}{\text{Std}\left((h_m^{\text{tail}} * \eta_m^{\text{SSN}})[n]\right) \times 10^{\text{SNR}^{\text{diff}}/20}}, \tag{4.7}$$

$$\nu = \frac{\sigma_{\text{ref}}}{10^{\text{SNR}^{\text{static}}/20}}. \tag{4.8}$$

Scalars ρ and ν are used to scale the respective noise components in Equation (4.2). In practice, this procedure led to an overall SNR of the mixtures in the dataset lying in the

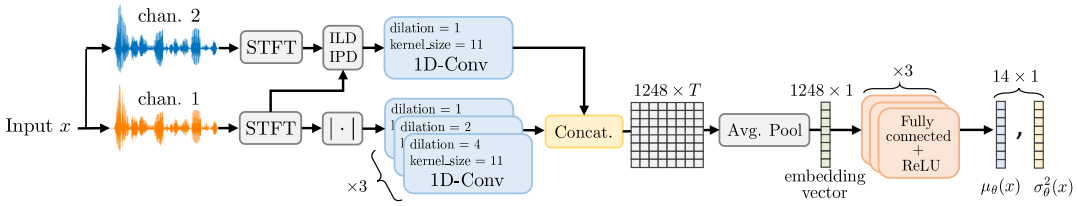

Figure 4.2: Proposed neural network architecture.

range $[-10, 65]$ dB. Multiple observations are collected in each room and are concatenated as follows :

$$\mathbf{x} = [\mathbf{c}_{d=1}, \ldots, \mathbf{c}_D], \tag{4.9}$$

where D denotes the number of observations and $\mathbf{c}$ is a multichannel signal for each observation that contains a single source. Using $D = 5$ observations in each of the 20,000 different rooms led to a dataset of 100,000 multichannel noisy speech signals.

This dataset is divided into non-overlapping training, validation, and test sets of sizes 80k, 10k, and 10k, respectively. $Q = 14$ parameters are associated with each room, namely, S, V, and 6 mean absorptions, 6 reverberation times for all octave bands. To obtain a unique, reference value of the reverberation time for each room from five different observations, the RT_{60} of the five corresponding RIRs are calculated in each octave band, and their median is used as the labeled RT_{60} for this room.

4.2 Neural network model

To perform the blind room acoustic parameter estimation task, we propose a new DNN architecture shown in Figure 4.2. The basic component of this architecture uses 1D convolutional blocks, which are inspired by the Conv-Tas Net architecture initially introduced for the task of speech separation in Luo and Mesgarani (2019). Each 1D convolutional block is made up of different parts, namely, a separable convolution layer followed by a ReLU and layer normalization (Ba et al., 2016). Separable convolutions decouple a convolution operation into a depthwise and a point-wise convolution. It has been found useful to reduce the number of parameters, hence allowing for faster and deeper networks. Scale-invariant representations are created by layer normalizations which were found crucially important for this multi-task DNN network. The network is divided into 2 pipelines. The bottom part has three 1-D convolution blocks with a progressive dilation rate of [1,2,4] along the frequency axis and a constant kernel size of 11. Only one 1-D convolutional blocks gave satisfactory results when applied to the upper part. Both parts are meant to process different feature sets. The bottom part is meant for single-channel features (SC) and the other for inter-channel features (IC). Refined features obtained from both pipelines are concatenated along the frequency axis and average pooled across time to form an embed-

ding vector of size (1248 x 1). This is followed by 3 fully connected layers of respective dimensions of 96, 48, 28, yielding 28 outputs.

4.2.1 Input features

Although the proposed architecture works on any input signal length, the length of input signals was fixed to 3 seconds for all training samples. The multichannel input signal $\mathbf{c}$ from a single observation is first processed by the STFT with a 96 ms Hann window and 50% overlap, resulting in a spectrogram $C_m(f,t)_{f,t=1}^{F,T}$ with F =769 positive frequency bins and T =63 time frames for each channel m. This representation is then used to calculate specialized single-channel and inter-channel features. Four different single-channel features were considered : $|C_1(f,t)|^2$, $\sqrt{|C_1(f,t)|}$, $\log|C_1(f,t)|$ and $|C_1(f,t)|$. The latter performed the best out of the four in our preliminary experiments, so we opted for this choice Meanwhile, inter-channel features are obtained by concatenating the Interaural level differences (ILD) and the cosine and sine of Interaural phase differences (IPD):

$$\mathrm{ILD}(f,t) = \log|C_1(f,t)| - \log|C_2(f,t)|\,, \tag{4.10}$$

$$\mathrm{IPD}(f,t) = \left[\mathrm{Re}, \mathrm{Im}\left(\frac{C_1(f,t)C_2^*(f,t)}{|C_1(f,t)C_2^*(f,t)|}\right)\right]. \tag{4.11}$$

4.2.2 Loss function

To account for the different magnitudes of uncertainty in jointly estimating the Q = 14 target parameters, the network outputs are modeled as independent Gaussians and a maximum likelihood approach is employed. The neural network parameters is denoted by Θ. Hence, the model estimates 28 values of which 14 are the mean $\mu_\Theta(\mathbf{c})$ i.e., the estimates of the regressor. The rest are variances $\sigma^2_\Theta(\mathbf{c})$,i.e., uncertainties of the estimated parameters. Maximizing the likelihood of this model yields the following loss function against which the network parameters are optimized :

$$\mathcal{L}_\Theta(\mathbf{c},\mathbf{y}) = -\log\mathcal{N}(\mathbf{y};\mu_\Theta(\mathbf{c}),\sigma^2_\Theta(\mathbf{c})) = \frac{1}{2}\sum_{q=1}^{Q}\log\sigma^2_{q,\Theta}(\mathbf{c}) + \frac{(y_q-\mu_{q,\Theta}(\mathbf{c}))^2}{\sigma^2_{q,\Theta}(\mathbf{c})}\,, \tag{4.12}$$

where the ground truth labels are denoted by $\mathbf{y} \in \mathbb{R}^Q$.

4.2.3 Fusion of the estimates

This approach has two advantages. First, it allows for the adaptive weighing of errors on individual parameters. Second, the dataset includes multiple observations in the same room, and the network outputs multiple estimations based on these observations. These

estimations can be fused together using the variance that the network outputs for all parameters, based on this formula derived from Bayes' theorem :

$$p_\Theta(y_q|\mathbf{x}) = \mathcal{N}(y_q; \overline{\mu}_{q,\Theta}(\mathbf{x}), 1/\overline{\gamma}^2_{q,\Theta}(\mathbf{x}))\,, \tag{4.13}$$

where $\overline{\mu}_{q,\Theta}(\mathbf{x})$ is the fused estimate whose formula is given by :

$$\overline{\mu}_{q,\Theta}(\mathbf{x}) = \sum_{d=1}^{D} \frac{\gamma^2_{q,\Theta}(\mathbf{c}_d)}{\overline{\gamma}^2_{q,\Theta}(\mathbf{x})} \mu_{q,\Theta}(\mathbf{c}_d)\,, \tag{4.14}$$

where $\overline{\gamma}^2_{q,\Theta}(\mathbf{x}) = \sum_{d=1}^{D} \gamma^2_{q,\Theta}(\mathbf{c}_d)$, and $\gamma^2_{q,\Theta}(\mathbf{c}_d)$ is the estimated precision that is the inverse of the estimated variance $\gamma^2_{q,\Theta}(\mathbf{c}_d) = 1/\sigma^2_{q,\Theta}(\mathbf{c}_d)$.

4.2.4 Hyperparameters and training

The network was trained for 120 epochs with the ADAM optimizer (Kingma and Ba, 2014) using a learning rate of 10^{-4}. To avoid overfitting, dropout layers are used after each convolutional block with a dropout probability of 0.2, and after each fully connected layer with a dropout rate of 0.4. The network, trained on a set of 80k binaural examples, typically converges within 100-150 epochs, with a patience of 15 epochs on the validation set as a pre-condition for early stopping.

The parameters estimated by the network exhibit different scales. In order to avoid issues due to scale differences, we normalized the ground truth values y at training time, by dividing them by their respective standard deviations calculated over the training set. At inference time these standard deviations are reused and multiplied with the network output.

4.2.5 Alternative DNN architectures

On the way to designing the above-mentioned DNN architecture, we experimented at different levels of the architecture, namely, input features, the layers, the loss function, and the learning approach. These experiments gave us important clues to proceed further. Hence, mentioning them may provide readers with an insight on certain errors to be avoided while experimenting with this task.

Learnable input filter banks were proposed by the author of Conv-TasNet (Luo and Mesgarani, 2019). We tried to replace the STFT by advanced learnable filter banks for single-channel and inter-channel feature calculation, but this change yielded a 25% increase in the average error and doubled the variability of the system. With the current STFT implementation, we experimented with different window sizes of 32, 64, 96 and 128 ms. Best results were obtained with 96 ms.

Alternatively, X-vectors (Snyder et al., 2018) are a state-of-the-art architecture for speaker recognition. Inspired by this, we experimented with modified versions of this architecture adapted to our task. However, these models did not perform well on our data

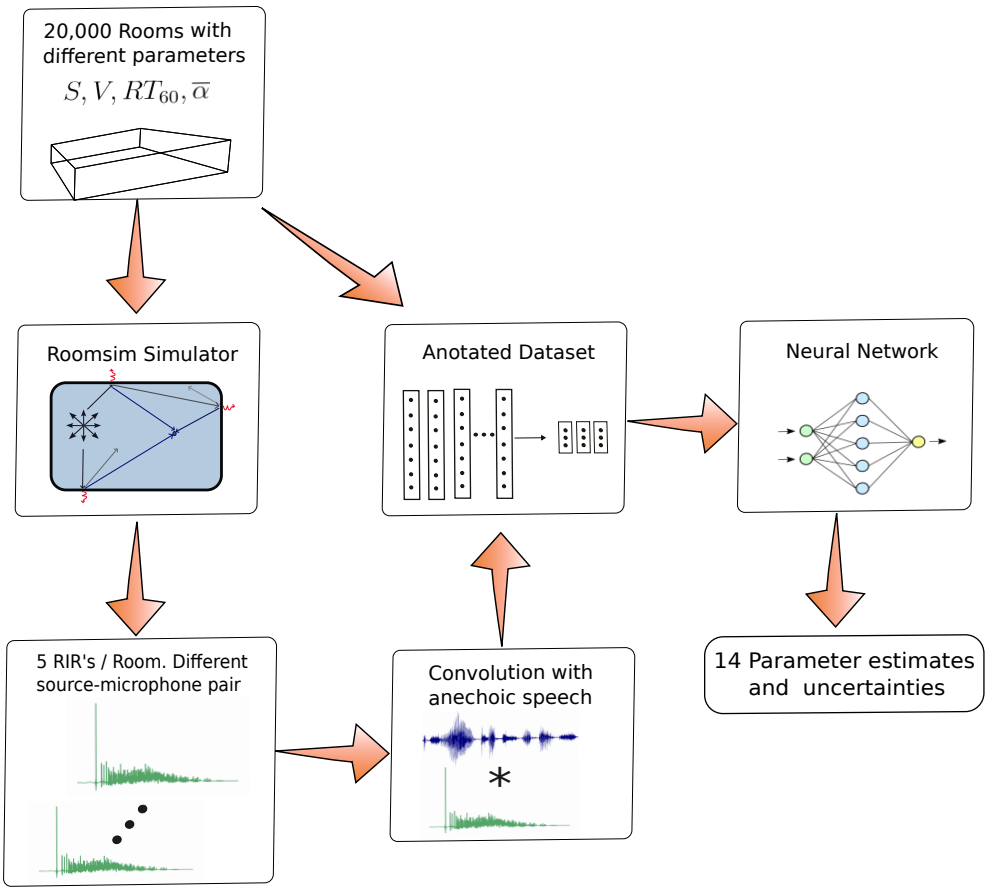

Figure 4.3: Training workflow for the experiments

and we observed a 1.25 times increase in variance on the errors. Continuing further, we experimented with the loss function. We replaced the Gaussian negative log-likelihood loss function with a mean squared error loss. This change increased the average error by three times and also took away the functionality of fusing multiple observations in a room. As suggested by Genovese et al. (2019), we experimented with using log-scale labels for the surface and volume at the output, i.e., $\log(S)$ and $\log(V)$. This resulted in much worse results, probably due to the range difference, as we considered a narrower range of room sizes in our dataset. At last, we gave a chance to single-task learning. We trained four copies of the proposed architecture to estimate the $(\overline{\alpha}, \mathrm{RT}_{60}, S, V)$ parameters separately. Identical results were obtained, at the cost of a four-fold increase in the number of parameters and training time.

4.3 Experiments and results

4.3.1 Baseline system and evaluation metric

The literature survey on room parameter estimation conducted in Chapter 3 revealed a dearth of DNN-based methods capable of performing blind multi-task room parameter estimation. Furthermore, to the best of our knowledge, no existing methods are capable of estimating a room's surface area, frequency-dependent reverberation time, and mean absorption coefficient solely based on noisy speech signals. However, a method does exist for blind one-channel single-position room volume estimation (Genovese et al., 2019). We use this work as a baseline to compare the task of volume estimation for different tests on both simulated and real data. Due to the non-availability of code for this work, we re-implemented their architecture from scratch and trained it on the same dataset as the proposed model. Throughout the experiments, the mean absolute error on each parameter is used as a metric.

4.3.2 Simulated data

The network is trained according to the workflow provided in Figure 4.3. After training, we perform 2 different experiments with a simulated test set consisting of 2,000 unseen rooms. The aim of our first experiment is to verify whether the fusion of multiple observations taken in the same room using the approach proposed in Section 4.2.2 does in fact improve the results both for one-channel and two-channel input signals.

Therefore, for this experiment, we compared two approaches:

1. The full architecture depicted in Figure 4.2, which takes into account two-channel input .
2. The same architecture is used but without the upper inter-channel processing part of the network, thus processing only the single channel.

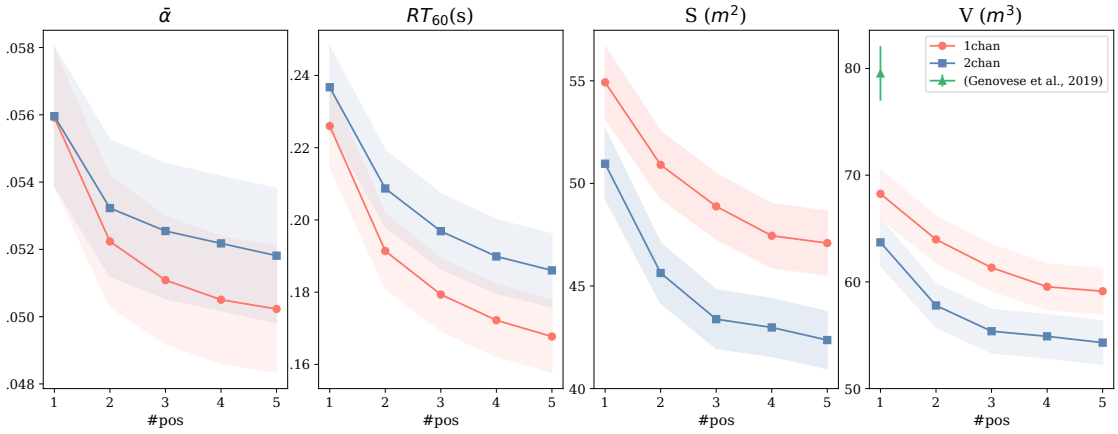

Figure 4.4: Mean absolute error achieved on simulated data by Genovese et al. (2019) vs. the proposed model with one- or two-channel inputs, as a function of the number of source-receiver positions fused in each room. Shaded areas indicate 95% confidence intervals

The obtained results are compared between the two approaches for the estimation of S, RT_{60} and $\overline{\alpha}$, while the comparison extends to the baseline method for the estimation of V.

The results in Figure 4.4 for $\overline{\alpha}(b)$ and $\mathrm{RT}_{60}(b)$ are averaged across all octave bands to show the average errors on $\overline{\alpha}$ and RT_{60}. The figure clearly shows that the fusion of multiple observations per room lowers the error in a progressive manner on all parameters. Fusing 5 observations of a two-channel input signal provides the best performance with a mean absolute error of 0.052 for $\overline{\alpha}$, 0.18 s for RT_{60}, 42 m^2 on S and 54 m^3 on V, while the training set ranges for these quantities are as follows: $[0.02, 0.6]$ for $\overline{\alpha}$, $[0.2, 3.6]$ s for RT_{60}, $[48, 360]$ m^2 for S and $[18, 400]$ m^3 for V. The results for the estimation of V reveal that our proposed method outperforms the baseline on both approaches with only a single observation. One-channel with one observation reduces errors by 13%, while two-channel with five observations reduces errors by 31 % with respect to the baseline. The provided 95% confidence intervals in Figure 4.4 show significance in the mean absolute error between the compared approaches. The two-channel approach achieves a significantly lower mean absolute error for the estimation of S and V compared to the one-channel approach. This result can be associated with inter-channel features that efficiently capture the spatial characteristics of the room, which helps in estimating S and V which are spatial quantities that depend on early reflections. However, given the conflicting confidence intervals, using two channels instead of one does not seem to help the estimation of $\overline{\alpha}$ and RT_{60}.

We conduct a second experiment to examine the impact of combining single-channel and inter-channel representations at the embedding layer. This investigation aims to determine whether the improvement observed in the previous experiment comes from combining single-channel and inter-channel features or if it is associated with the doubling of signal length emitted by the source. This experiment was realized with 1 source-microphone

Input	Feature	$\bar{\alpha}$	RT_{60} (s)	S (m^2)	V (m^3)
1 microphone, 1 signal	SC	0.055	0.225	55.0	68.8
1 microphone, 2 signals	SC	0.058	0.222	55.4	69.7
2 microphones, 1 signal	SC	0.057	0.221	54.0	67.7
2 microphones, 1 signal	SC+IC	0.055	0.236	**49.9**	**63.7**

Table 4.2: Mean absolute errors achieved on simulated data for one source-receiver position using different inputs and features. Where SC, IC denote single-channel and inter-channel features. Bold numbers indicate the best statistically significant result per column based on 95% confidence intervals on the differences, when there is one.

observation with a signal length of 3 seconds. Four types of inputs and features are compared:

1. One speech signal and its single-channel representation (1 microphone, 1 signal)
2. Two speech signals and their average single-channel representations (1 microphone, 2 signals)
3. One speech signal from 2 individual mics and their concatenated single-channel representations (2 microphones, 1 signal)
4. One speech signal and its concatenated (single-channel + inter-channel) representations (2 microphones, 1 signal), i.e., our proposed method.

The results shown in Table 4.2 suggest that the inclusion of inter-channel features significantly improves the error for S and V, despite the fact that the dimension of the input features is the same as in experiments 2 and 3.

The results presented until now show values averaged across all octave bands for $\overline{\alpha}$ and RT_{60}. Detailed errors on a per-octave-band basis are shown in Table 4.3 for two-channel with one and five observations. These results reveal that the mean absolute error for $RT_{60}(b)$ decreases progressively from 125 Hz to 4 kHz by a factor of 3, while $\alpha(b)$ are well estimated and do not depict significant error differences across octave bands. These results on the RT_{60} might be due to less information present in narrower lower octave bands.

4.3.3 Real data

To test the trained model's generalization ability to unseen real acoustic conditions, wet speech recordings from the dEchorate dataset (Di Carlo et al., 2021) were used as a real test set. The dEchorate dataset was described briefly in Section 3.5. Apart from recorded RIRs the dataset also contains wet speech measurements taken in a modular acoustic shoebox room whose wall, floor, and ceiling can switch between reflective and absorbent characteristics with $S = 125$ m^2 and $V = 82$ m^3. As for our simulated data, ground truth for RT_{60} is estimated using Schroeder's integration method, and the median value is taken for each room from the calculated set of corresponding RT_{60}. Moreover, the ground truth val-

Octave bands	1 position		5 positions	
	$\bar{\alpha}$	RT_{60} (s)	$\bar{\alpha}$	RT_{60} (s)
125 Hz	0.056	0.392	0.051	0.320
250 Hz	0.060	0.305	0.055	0.249
500 Hz	0.057	0.228	0.051	0.170
1 kHz	0.053	0.188	0.050	0.146
2 kHz	0.054	0.165	0.051	0.122
4 kHz	0.052	0.139	0.049	0.106

Table 4.3: Mean absolute errors on $\bar{\alpha}(b)$ and $\mathrm{RT}_{60}(b)$ in the 6-octave bands achieved by the proposed model with multi-channel and 1 or 5 source-receiver positions per room on simulated data.

Method	Features	# pos	$\bar{\alpha}$	RT_{60}	S	V
Genovese et al. (2019)	SC	1	-	-	-	137.8
Ours	SC	1	**0.061**	0.134	129.6	154.5
Ours	SC	5	**0.060**	0.097	125.8	149.1
Ours	SC+IC	1	0.084	0.101	89.4	107.6
Ours	SC+IC	5	0.094	**0.062**	**50.2**	**68.8**

Table 4.4: Mean absolute error achieved over 3 rooms from the real dEchorate dataset. Where SC, IC denote single-channel and inter-channel features. Bold numbers indicate the best statistically significant result per column, based on 95% confidence intervals.

ues for $\overline{\alpha}$ is estimated using the Eyring model with a technique based on the aggregation of multiple RIR measurements (Foy et al., 2021). For both the parameters, the values in lower octave bands (125 kHz & 250 kHz) were omitted as the image source method used to train our model is known to be inaccurate in this regime (below the Schroeder frequency). Out of 11 room configurations in the dataset we choose 3 room configurations that have 3 or more reflective surfaces with $\overline{\alpha}(b) \in [0.16 - 0.35]$ and $\mathrm{RT}_{60}(b) \in [0.25 - 0.66]$ sec. These rooms resemble real acoustic rooms and fall under the scenario chosen for the training set. We chose a combination of 30 microphone pairs for each of the three rooms, resulting in $3 \times 30 = 90$ 3-second speech signals.

Results on the real test set are shown in Table 4.4. The mean absolute errors are reported in the table using the proposed method for both one-channel and two-channel approaches. It also provides a comparison with the baseline method on V estimation. They were calculated on 1 or 5 observations corresponding to a fixed microphone array and different source positions that were randomly chosen from the 6 sources present in the room. An important point seen from the table is that the errors generated by our approach fall within the same range as those achieved on simulated data. Also, errors for the estimation of V using the one-channel approach with one observation are comparable to the base-

Method	Features	# pos	$\bar{\alpha}$	RT_{60}	S	V
Genovese et al. (2019)	SC	1	-	-	-	10.0
Ours	SC	1	0.030	0.161	27.2	31.8
Ours	SC	5	0.024	0.090	19.6	23.0
Ours	SC+IC	1	0.031	0.100	34.7	39.7
Ours	SC+IC	5	0.015	0.054	16.5	18.9

Table 4.5: Standard deviation of parameter estimates for room "011100" of the real dEchorate dataset. Where SC, IC denote single-channel and inter-channel features.

line method. The results using the two-channel approach with inter-channel features also bolster the hypothesis that inter-channel features are important for the estimation of S and V as their inclusion significantly decreases errors. Also, the numbers again show that increasing the number of observations decreases the error on RT_{60}, S, and V. However, errors obtained on $\bar{\alpha}$ do not show a pattern and are more varied than the ones obtained on the simulated dataset. This could be explained by two reasons: First, the difficult problem of annotating absorption coefficients in a real room. Second, the modeling of absorption coefficients in the ISM used for training the model is less valid at lower frequencies.

At last, Table 4.5, provides standard deviations across all parameters for 30 speech signals from a specific room of the dEchorate dataset with 3 reflective surfaces. Encouragingly, the relatively small standard deviations indicate that the models are capable of producing stable parameter estimates within a room, and are not heavily influenced by the position of the source and receiver. Moreover, it can be observed that using multiple observations decreases the standard deviation of estimates across all the parameters for a given room.

4.4 Summary

The findings of this study demonstrate that incorporating interchannel cues can greatly enhance the blind estimation of room volume and surface from noisy speech. However, when estimating reverberation and absorption parameters, a single channel is found to be adequate. It should be noted that the estimation of absorption coefficients is not very reliable, due to challenges in annotating real data and the limitations of the absorption model for walls used in the ISM. Furthermore, the study emphasizes that combining multiple measurements reduces estimation errors and variances for all parameters. The results also demonstrate that a system trained on a meticulously simulated training set exhibits satisfactory generalization capabilities when applied to real-world data. However, it should be noted that a single real room was used, with a fixed surface and volume, limiting the ability to draw conclusions for a wider variety of room types. The lack of representation is also seen in the training set as it was limited to a specific range of room acoustics and is unlikely to generalize to acoustics beyond this range, e.g., cathedrals, recording studios, and non-

rectangular rooms. Lastly, source and receiver directivities and frequency responses were not taken into account while training the system.

5 Extended image source method and implementation under Pyroomacoustics

This chapter focuses on improving the realism of the open-source room acoustic simulator *Pyroomacoustics* by augmenting the simulator with the implementation of a more realistic version of the ISM without compromising its computational time. While there exists a large range of room acoustic simulators that can simulate RIRs at various degrees of realism (see Section 3.4), most of them are either not open-source or not written in Python. The Pyroomacoustics library satisfies both requirements. We view this contribution as a helpful step to improve the generalizability of virtually-supervised audio signal processing methods. The chapter starts with a description of the functioning of the Pyroomacoustics simulator in Section 5.1, then we describe the extended ISM in Section 5.2. The implementation of the extended ISM in the Pyroomacoustics simulator is presented in Section 5.3. We provide a qualitative comparison of the old and new versions of the simulator in Section 5.4. We conclude the chapter with Section 5.5, giving a brief discussion on further work in this direction.

5.1 Functioning of Pyroomacoustics

Pyroomacoustics is an open-source Python library that provides quick routines for acoustic simulation and various multi-channel processing algorithms that are easily accessible by Python object-oriented methods. RIR simulation can be performed in 2D and 3D polyhedral rooms with multiple sources and receivers via hybrid geometric acoustics approach. Additionally, the toolbox also provides algorithms for the STFT, source separation, single channel denoising, adaptive filtering, and beamforming, making it a versatile library in order to test and develop audio systems (Scheibler et al., 2018). This library provides high-level Python APIs that are built upon a Python binding known as *Cython* (Behnel et al., 2010). Parts of various algorithms that are associated with high computational costs are efficiently performed by code written in C++, which is wrapped by Cython. In this thesis, the usage of the Pyroomacoustics simulator is focused on simulating RIRs in 3D shoebox rooms using the ISM.

Before detailing our changes to implement our extended ISM, we describe how Py-

roomacoustics simulates RIRs using the ISM. The ISM in the simulator follows the principles of Allen and Berkley (1979), but its implementation follows a different paradigm: it creates RIRs via time domain processing of image source filters. This implementation is shown as pseudo-code in Implementation 1. The notations follow Python-style scripting and use array broadcasting for multiplication purposes. Also, the symbol $\odot$ denotes element-wise multiplication. For an easier interpretation of the algorithm, we consider an example of simulating an RIR using a single source-microphone pair. For the sake of legibility, the microphone index m is therefore omitted in the following. Most of the variables used in the implementation are a result of some external computations and they are described in the rest of this section in no particular order.

Simulating a RIR in a shoebox room requires the user to define high-level objects. The room is defined by its properties (room geometry, frequency-dependent $\alpha(b)$ for walls, sampling frequency, and maximum order for ISM). Similarly, the source and microphone need to be defined by their properties (position and directivity). Following the room, source, and microphone description, the function `get_pos_attn` calls the C++ engine and returns the position of all the K image sources in 3D coordinates, which will be referred to as $\mathbf{I}^{\text{pos}} \in \mathbb{R}^{K\times 3}$, as well as the reflection order of each image source and their associated wall attenuation. The returned wall attenuations will be referred to as $\tilde{\mathbf{D}} \in \mathbb{R}^{K\times B}$. B denotes the number of octave bands. The absorption coefficient α of wall materials is often provided in 6 octave bands $b \in \mathbf{b} = [125, 250, 500, 1000, 2000, 4000]$ Hz. For each image source k the wall attenuation in octave bands is calculated as the product of the attenuations due to the walls encountered along the path from the image source to the microphone :

$$\tilde{D}_k(b) = \prod_v \overline{\beta}_{i_{v,k}}(b), \tag{5.1}$$

where $\overline{\beta}_i(b)$ is the reflection coefficient and is calculated from the absorption coefficient $\alpha_i(b)$ of surface i according to Equation (2.22), and $i_{v,k}$ is the index of the v-th surface encountered. Python functions such as `get_time_delay` and `get_angles` use $\mathbf{I}^{\text{pos}}$ to calculate the following quantities for each image source k :

- The time of arrival of each image source k, which is denoted as a vector $\mathbf{t}^{\text{delay}} = [t_k^{\text{delay}}]_{k=0}^{K-1}$ of size K. Further, t_k^{delay} is converted from time in seconds to time in samples and every time is split into an integer sampled point $t_k^{\text{int}} \in \mathbb{Z}$ and a fractional sampled point $t_k^{\text{frac}} \in \mathbb{R}$.
- The incoming angle of arrival (θ_k, ϕ_k) at the microphone and the outgoing angle of departure $(-\theta_k, -\phi_k)$ from the source, whose calculations are based on Brinkmann et al. (2019). In the pseudo-code, they are referred to in matrix form with the angle of arrival denoted by $\mathbf{AOA} \in \mathbb{R}^{K\times 2}$ and the angle of departure by $\mathbf{AOD} = -\mathbf{AOA}$.

The sensitivity of the source and the microphone for angles $\mathbf{AOD}$ and $\mathbf{AOA}$ is retrieved using a `get_response` function. It is represented as $\tilde{\mathbf{g}}^{\text{src}} \in \mathbb{R}^K$ and $\tilde{\mathbf{g}}^{\text{mic}} \in \mathbb{R}^K$. With no specified directivity for the microphone and source we have $\tilde{\mathbf{g}}^{\text{src}} = \tilde{\mathbf{g}}^{\text{mic}} = [1]_{k=0}^{K-1}$. This simulates the condition of omnidirectional directivity patterns.

Implementation 1 Time domain RIR construction in *Pyroomacoustics* for a single source-microphone pair

$\mathbf{I}^{\text{pos}}, \tilde{\mathbf{D}} \leftarrow$ `get_pos_attn()` *# retrieve image source positions and attenuations*
$\mathbf{AOD}, \mathbf{AOA} \leftarrow$ `get_angles`$(\mathbf{I}^{\text{pos}})$ *# retrieve incoming and outgoing angles of image sources*
$\mathbf{t}^{\text{delay}} \leftarrow$ `get_time_delay`$(\mathbf{I}^{\text{pos}})$ *# retrieve time of arrival of each image source*
$\tilde{\mathbf{g}}^{\text{src}} \leftarrow$ `get_response`$(\mathbf{AOD})$ *# retrieve frequency-independent gain for source*
$\tilde{\mathbf{g}}^{\text{mic}} \leftarrow$ `get_response`$(\mathbf{AOA})$
for $b \leftarrow 1$ to B **do** *# loop over octave bands*
  $\tilde{\mathbf{D}}[:, b] \leftarrow \tilde{\mathbf{g}}^{\text{src}} \odot \tilde{\mathbf{g}}^{\text{mic}} \odot \tilde{\mathbf{D}}[:, b]$
  $\mathbf{rir}_b \leftarrow$ `fast_sinc_interp`$(\tilde{\mathbf{D}}[:, b], \mathbf{t}^{\text{delay}})$ *# builds RIR*
  $\mathbf{rir}_b \leftarrow$ `octave_band_analysis`$(\mathbf{rir}_b, \mathbf{b}[b])$ *# filter RIR with octave band filter*
end for

A recent update to the Pyroomacoustics simulator added the ability to load frequency-independent cardioid patterns to individual sources and microphones. To devise such directivity patterns an analytical formula is used

$$\tilde{g}(\theta) = \psi + (1 - \psi)\cos(\theta), \tag{5.2}$$

where $\psi \in [0, 1]$. Directivity patterns such as the figure-of-eight, hyper-cardioid, subcardioid, and omnidirectional are obtained by setting $\psi = [0, 0.25, 0.5, 0.75, 1]$. The pointing direction of the directive pattern is controlled by θ, while the dependency towards elevation is not considered. According to the incoming and outgoing angles of an image source k, the analytical pattern is queried resulting in the direction-sensitive values that are constant throughout the frequency bands, hence the function `get_response` returns sensitivity values as a vector of size K, ignoring the frequency-dependency by repeating the same value across $\mathbf{b}$.With attenuations and time delay per image source, two functions `fast_sinc_interp` and `octave_band_analysis` create multiple RIRs in the time domain for octave bands in $\mathbf{b}$. The `fast_sinc_interp` function takes 2 parameters: the attenuation $\tilde{\mathbf{D}}[:, b]$ and the source senstivity for octave band b and $\mathbf{t}^{\text{delay}}$. The function fills up the vector $\mathbf{rir}_b$ by going through the attenuation of each image source k, windowing it using a Hann window, and adding a fractional delay at t_k^{frac} using a convolution with a fractional delay filter, i.e., a sinc function. This results in an image source filter that is added to $\mathbf{rir}_b$ at time t_k^{int}. Although $\mathbf{rir}_b$ is filled by attenuated and delayed image source filters, these RIRs lack information over frequency. The `octave_band_analysis` function convolves $\mathbf{rir}_b$ with a raised cosine filter $\zeta_b(f)$ whose frequency response (shown in Figure 5.1) matches the requested octave band in $\mathbf{b}$.Eventually, the resulting $\mathbf{rir}_b$ vectors are summed together to obtain the RIR :

$$h = \sum_{b=0}^{B-1} \mathbf{rir}_b. \tag{5.3}$$

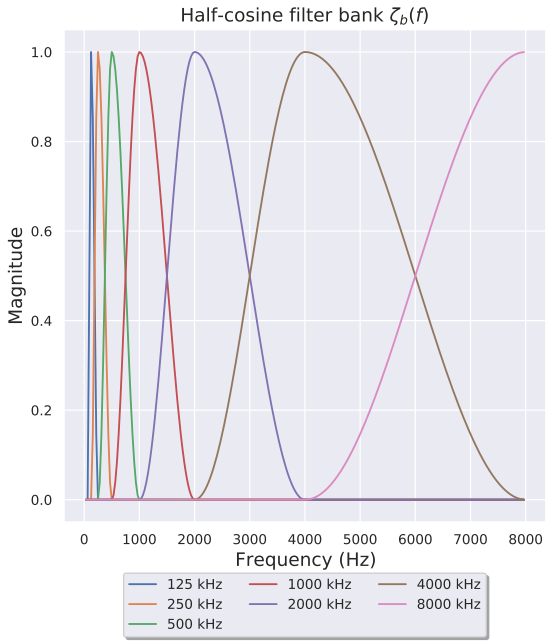

Figure 5.1: Frequency response of half cosine filter bank

5.2 Extended ISM

In Section 2.1.1.2 we discuss the directivity patterns of real sound sources and microphones which tend to be frequency dependent, approximately omnidirectional at lower frequencies, and more directive at higher frequencies. In the previous section, it was shown that the construction of RIR in Pyroomacoustics takes place in the time domain with the help of image source attenuations that are only defined and calculated in octave bands. This not only disregards the full frequency scale but also makes it difficult to include frequency-dependent directivity patterns in the construction of RIRs. To make the simulator and RIRs more realistic for the purpose of virtually supervised learning (see Section 3.3), we propose to add the ability to load measured directivity patterns of real sources and microphones into the simulator. In order to take into account the entirety of the frequency scale, our contribution towards the simulator changes the core construction model of RIRs in Pyroomacoustics, leading to RIRs being constructed in the frequency domain. Most room acoustic simulators do not implement the ISM in its extended form, although it can be done with little computational overhead.

In the previous chapter, Equation (4.1), represents an ISM model that includes frequency-dependent wall attenuation. Wall attenuation is modeled as a filter $D_{k,m}(f)$ for every microphone m and every image source k, and a RIR is represented as a sum of delayed filters. Now, to account for the source and receiver directivity patterns in the RIR simulation, we

extend Equation (4.1) as

$$H(\mathbf{r}_m^{\text{mic}}, f) = \sum_{k=0}^{K} \frac{D_{k,m}(f)}{4\pi \|\mathbf{r}_m^{\text{mic}} - \mathbf{r}_k^{\text{src}}\|_2} e^{-i2\pi f t_{k,m}^{\text{delay}}} D^{\text{air}}(f) G^{\text{src}}(-\theta_{k,m}, -\phi_{k,m}, f) \\ G^{\text{mic}}(\theta_{k,m}, \phi_{k,m}, f), \quad (5.4)$$

where atmospheric attenuation in the frequency domain is denoted as $D^{\text{air}}(f)$. The time of arrival at a microphone m for an image source k is denoted by $t_{k,m}^{\text{delay}}$. Microphone and source directivity patterns are denoted by G^{mic} and G^{src}, the incoming angles in azimuth and elevation of an image source k towards a microphone m are denoted by $(\theta_{k,m}, \phi_{k,m})$. The outgoing angle from a given source is the opposite of the incoming angle at the microphone. This equation will be referred to as the extended ISM model. A similar formulation was given by Schröder (2011). It also acts as the main protagonist of this chapter, as most of the following work is based on our implementation of this equation in the Pyroomacoustics simulator.

5.3 Extended ISM implementation in Pyroomacoustics

This section details the implementation of the extended ISM in the Pyroomacoustics simulator. We start with a description of the directivity datasets for sources and receivers that are currently supported by the simulator. Then we describe interpolation of the directivity patterns using the DSHT which is vital for our extended ISM implementation. In the following subsection, we present RIR construction in the frequency domain and end this section by presenting extra features that are added to the modified Pyroomacoustics simulator.

5.3.1 Directivity datasets

While surveying for measured directivity patterns, we found that equipment manufacturers often only provide the frequency response of their products, with the full measured directivity pattern being absent. There are a few reasons that contribute to the absence of open-source directivity patterns: the difficulties in accessing an anechoic chamber or acquiring related equipment, and a relative lack of interest in the exact directivity patterns of equipment from most users. However, we found two open-source directivity datasets (see Section 5.3.1.2), and on the basis of these sources of data we develop the extended version of the Pyroomacoustics simulator.

5.3.1.1 SOFA format

A special data storage format known as the Spatially Oriented Format for Acoustics (SOFA) was proposed by Majdak et al. (2013). The SOFA format is designed to store acoustic information pertaining to a particular geometric configuration. The SOFA format was initially used to store information on HRTF measurements, but was later extended to the storage of binaural room impulse responses (BRIRs) and spatial room impulse responses (SRIRs). Prior to this standardized method, custom file formats were used to store HRTF measurements, which made it challenging to exchange and distribute datasets (Geronazzo et al., 2013). Both datasets used in our implementation use the SOFA format.

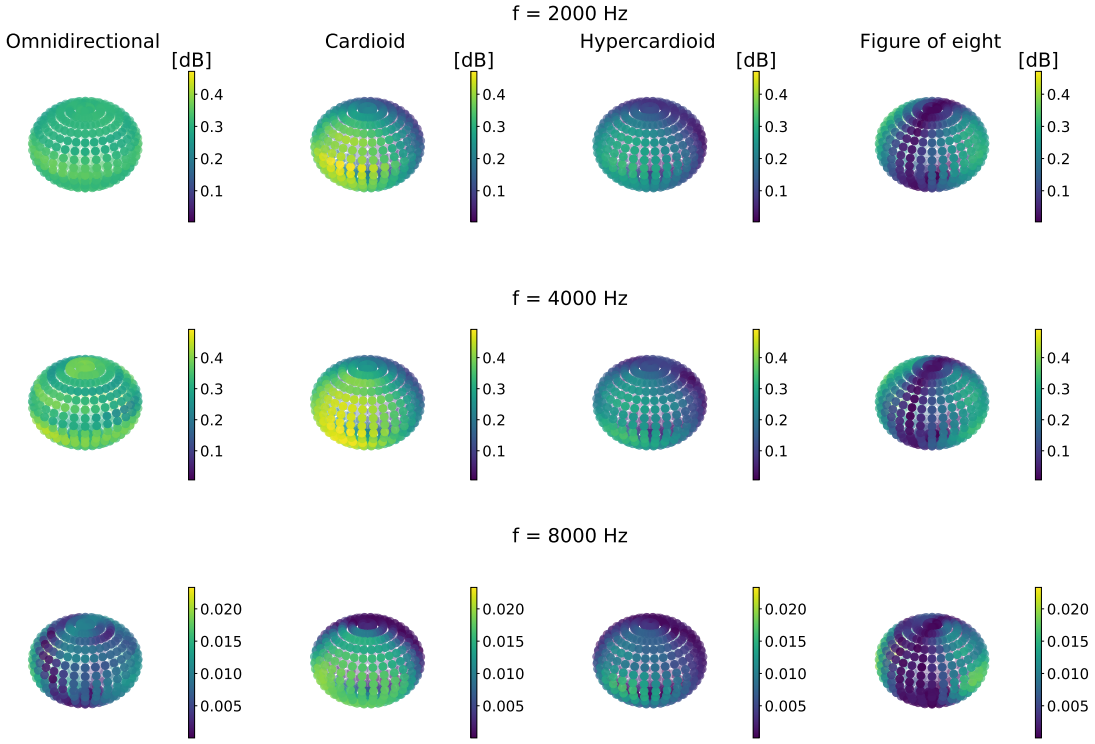

Figure 5.2: Spherical heatmap of AKG C414 from the DIRPAT dataset. The colorbars show normalized magnitudes of the filters.

5.3.1.2 DIRPAT and other datasets

The DIRPAT dataset consists of 2D and 3D directivity patterns of a variety of sources and receivers measured at the Institute of Electronic Music and Acoustics, University of Music and Performing Arts in Graz by Brandner et al. (2018). The dataset is freely available and stored in SOFA format for reproducible research. It consists of 3D measurements of four different types of receivers: the AKG C414 microphone with four directivity pattern

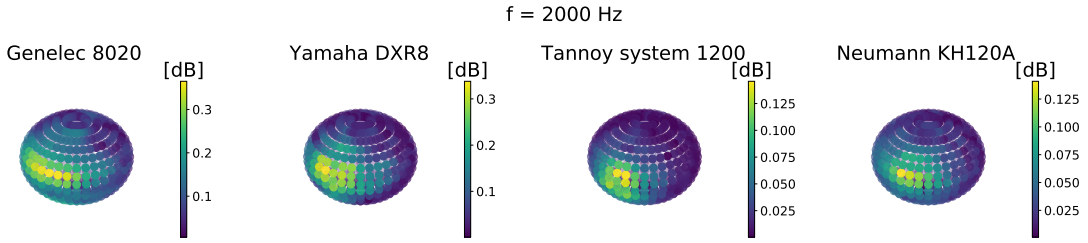

Figure 5.3: Spherical heatmap of 4 different loudspeaker manufacturers from the DIRPAT dataset. The colorbars show normalized magnitudes of the filters.

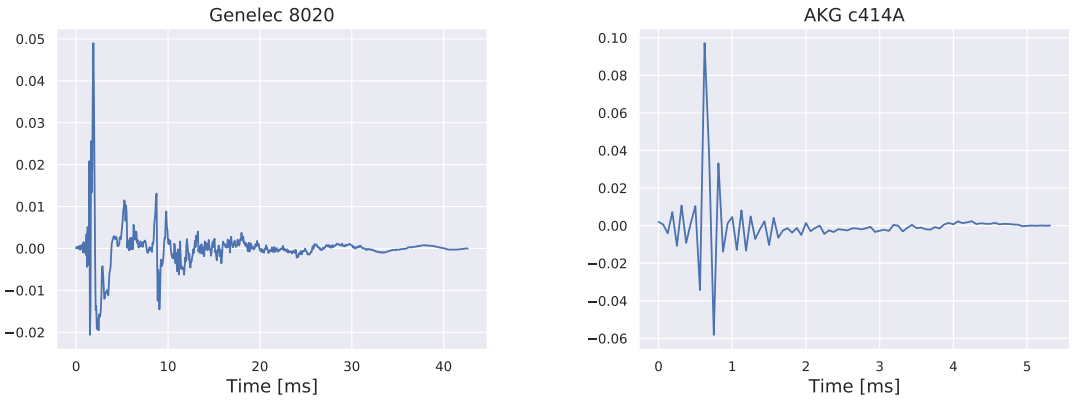

Figure 5.4: FIR Filters taken at a random point from a spherical function of a loudspeaker Genelec 8020 and microphone AKG C414A present in the DIRPAT dataset.

settings (omnidirectional, cardioid, hypercardioid, and figure-of-eight), the AKG C480 microphone, and two Soundfield microphones. Figure 5.2 shows the spherical heatmap of all the directivity patterns exhibited by the AKG C414 microphone at three different frequencies. The dataset also includes measured patterns of twelve sources, including generic loudspeakers, guitar amplifiers, and a HATS capable of approximating human speech directivity. Figure 5.3 shows the spherical heatmap of a few loudspeakers at $f = 2000$ Hz.

In the dataset, the directivities of sources and microphones are measured on two different grids. The grid for the sources consists of 540 measurement points while the grid for the microphones consists of 480 measurement points around the sphere. Grid points are further factored into 30 regularly-spaced azimuths and 16 or 18 regularly-spaced elevations. We denote these grids by $\mathcal{U}^{\text{meas}} = [(\theta_i^{\text{meas}}, \phi_i^{\text{meas}})]_{i=1}^{I^{\text{meas}}}$, where I^{meas} is the number of measurement points on the grid. The points on the grid are provided in the spherical coordinate system $[\theta^{\text{meas}}, \phi^{\text{meas}}, r']$ with $\theta^{\text{meas}} \in [0, 2\pi]$, $\phi^{\text{meas}} \in [0, \pi]$ and r' is the radius of the sphere. Each measurement point holds a Finite Impulse Response (FIR) filter sampled at 44.1 kHz for the incident sound path. In our implementation, the FIR filters are

downsampled to 16 kHz by default. Two such downsampled filters are presented in Figure 5.4. Tables 5.1 and 5.2 provide detailed specifications of the available sources and receivers in the dataset, with their grid coordinates and FIR length in samples. In addition to the DIRPAT dataset, we also utilized another directivity dataset that provides the directivity pattern of a 32-channel Eigenmike. This measurement was taken on a similar grid as the one specified for the DIRPAT receivers, and at the same location by Franz Zotter et al [1]. Therefore, the nomenclatures for geometric configuration and length of the FIR filter are the same as for the receivers in the DIRPAT dataset. We denote it as *"em32Directivity"* when used in our experiments.

Microphone	Number of azimuth	Number of elevation	FIR length
Oktava MK4012 (Soundfield)	30	16	160
Soundfield ST450			130
AKG c480/ck61 (cardioid)			256
AKG c414K (omnidirectional)			256
AKG c414N (wide-cardioid)			256
AKG c414S (hyper-cardioid)			256
AKG c414A (figure-of-8)			256

Table 5.1: Microphone directivity measurement specifications in DIRPAT.

5.3.2 DSHT and interpolation

The available grid points and corresponding filters in the measured receivers and sources of our selected directivity datasets are insufficient to represent the numerous incoming and exit angles of the image sources generated during simulations. Therefore, we perform an interpolation of the measurement grids $\mathcal{U}^{\text{meas}}$ onto more densely populated Fibonacci grids denoted as $\mathcal{U}^{\text{fibo}} = [(\theta_i^{\text{fibo}}, \phi_i^{\text{fibo}})]_{i=1}^{I^{\text{fibo}}}$, the two grids in 3D are shown in Figure 5.5. The Fibonacci grid consists of $I^{\text{fibo}} = 1,000$ measurement points, which provides a uniform distribution of points on the sphere (González, 2010).

The implementation of the spherical interpolation of the measurement grids is based on the DSHT, whose detailed description is given in Section 2.3.3. For the purpose of interpolation, we choose an order $L = 12$ for the forward and inverse DSHT. The values of spherical basis functions on the measured grid $\mathcal{U}^{\text{meas}}$ are calculated using Equation (2.10). We refer to this set of basis functions in matrix form as $\mathbf{Y}^{\text{meas}} \in \mathbb{C}^{I^{\text{meas}} \times (L+1)^2}$. Similarly, the basis functions on the Fibonacci grid $\mathcal{U}^{\text{fibo}}$ are denoted as $\mathbf{Y}^{\text{fibo}} \in \mathbb{C}^{I^{\text{fibo}} \times (L+1)^2}$. Measured directivity patterns in the directivity datasets are sampled spherical functions

[1] https://phaidra.kug.ac.at/o:69292.

Directional Source	Number of azimuth	Number of elevation	FIR length
Genelec 8020			
Lambda Labs CX-1A			
Tannoy System 1200			
Neumann KH120A			
Yamaha DXR8			
Bruel and Kjaer 4128C	36	15	2048
BM 1x12inch driver closed cabinet			
BM 1x12inch driver open cabinet			
BM open stacked with Crossover Network			
BM open stacked on closed fullrange			
Palmer 1x12 inch			
Vibrolux 2x10 inch			

Table 5.2: Source specification in DIRPAT. The first five are directive sources, the sixth is a HATS simulator, and the last six are the guitar amplifiers.

$\{g^{\text{meas}}(\theta_i^{\text{meas}}, \phi_i^{\text{meas}}, t)\}_{i\in\mathcal{U}^{\text{meas}}, t\in[1,\dots,T]}$. The FIR filters evaluated at the sampled spherical function for the measurement grids are transformed to the frequency domain by a DFT operation:

$$g^{\text{meas}}(\theta_i^{\text{meas}}, \phi_i^{\text{meas}}, t) \xrightarrow{\text{DFT}} G^{\text{meas}}(\theta_i^{\text{meas}}, \phi_i^{\text{meas}}, f)\,. \tag{5.5}$$

The FIR filters in the frequency domain are accumulated in a matrix denoted as $\mathbf{G}^{\text{meas}} = [G^{\text{meas}}(\theta_i^{\text{meas}}, \phi_i^{\text{meas}}, f)]_{i\in\mathcal{U}^{\text{meas}}, f\in[1,\dots,F']} \in \mathbb{C}^{I^{\text{meas}}\times F'}$, where $F' = N/2$ represents the number of positive frequency bins that is derived from the FIR filter length N. Zotter (2009) proposes a weighted least squares solution (see Section 2.3.4) for calculating a forward DSHT on the spherical grid with no exact inverse of the spherical basis , which is the case for our measurement grids $\mathcal{U}^{\text{meas}}$ and $\mathbf{Y}^{\text{meas}}$. Therefore, we write the forward DSHT as

$$\mathbf{Z}_L^{\text{meas}} = (\mathbf{Y}^{\text{meas}})^{+}\mathbf{Q}\mathbf{G}^{\text{meas}}\,, \tag{5.6}$$

where $\mathbf{Z}_L^{\text{meas}} \in \mathbb{C}^{(L+1)^2\times F'}$ are spherical harmonic coefficients for the projected spherical function $\mathbf{G}^{\text{meas}}$ on the spherical basis and $(\cdot)^+$ denotes the Moore-Penrose pseudo inverse. The quadrature weights are calculated based on Voronoi cells according to Zotter (2009) and is denoted by a diagonal matrix $\mathbf{Q} = \text{diag}\{\boldsymbol{q}\} \in \mathbb{R}^{I^{\text{meas}}\times I^{\text{meas}}}$. The weights are aligned in a diagonal matrix

$$\mathbf{Q} = \begin{bmatrix} q_0 & 0 & \cdots & 0 \\ 0 & q_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_{I^{\text{meas}}} \end{bmatrix}. \tag{5.7}$$

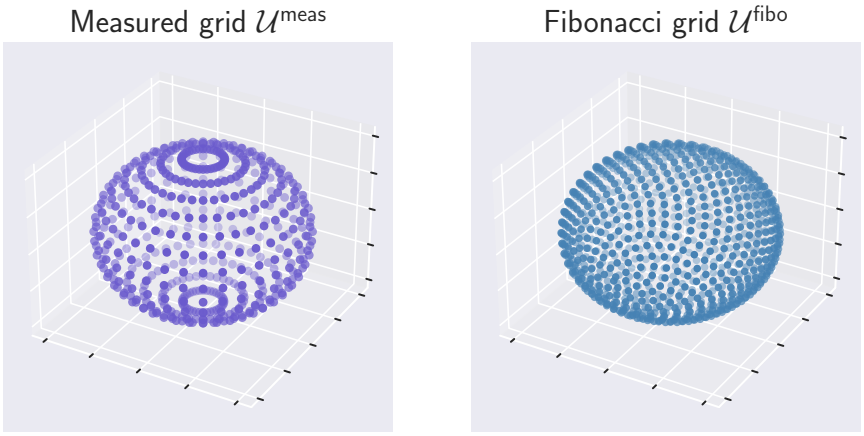

Figure 5.5: 3D scatter plot of the measured grid $\mathcal{U}^{\text{means}}$ (left) and the fibonacci grid $\mathcal{U}^{\text{fibo}}$ (right).

The value of the weights present as diagonal entries is associated with the area of each azimuth-elevation cell in the original grid. For the elevation angle, the geometrical formula is given by,

$$q_i = \frac{1}{2}\left[\cos\left(\frac{\text{left}(\phi_i^{\text{meas}}) + \phi_i^{\text{meas}}}{2}\right) - \cos\left(\frac{\phi_i^{\text{meas}} + \text{right}(\phi_i^{\text{meas}})}{2}\right)\right]_{i\in\mathcal{U}^{\text{meas}}}, \tag{5.8}$$

where $\text{left}(\phi_i^{\text{meas}})$ and $\text{right}(\phi_i^{\text{meas}})$ refer to the elevation points to the left and right of ϕ_i^{meas} on the grid. The spherical coefficients $\mathbf{Z}_L^{\text{meas}}$ of the measured grid are then used to calculate the interpolated transfer function on the Fibonacci grid

$$\mathbf{G}^{\text{fibo}} = \mathbf{Y}^{\text{fibo}}\mathbf{Z}_L^{\text{meas}}. \tag{5.9}$$

Finally, the IDFT inverts the frequency domain transfer functions to the time domain, resulting in an interpolated set of filters on a denser Fibonacci grid :

$$G^{\text{fibo}}(\theta_i^{\text{fibo}}, \phi_i^{\text{fibo}}, f) \xrightarrow{\text{IDFT}} g^{\text{fibo}}(\theta_i^{\text{fibo}}, \phi_i^{\text{fibo}}, t)_{i\in\mathcal{U}^{\text{fibo}}}. \tag{5.10}$$

Figure 5.6 shows the representation of spherical function interpolation in the form of a 2D spherical heatmap on the original grid $\mathcal{U}^{\text{meas}}$ and the corresponding heatmap on the denser Fibonacci grid $\mathcal{U}^{\text{fibo}}$.

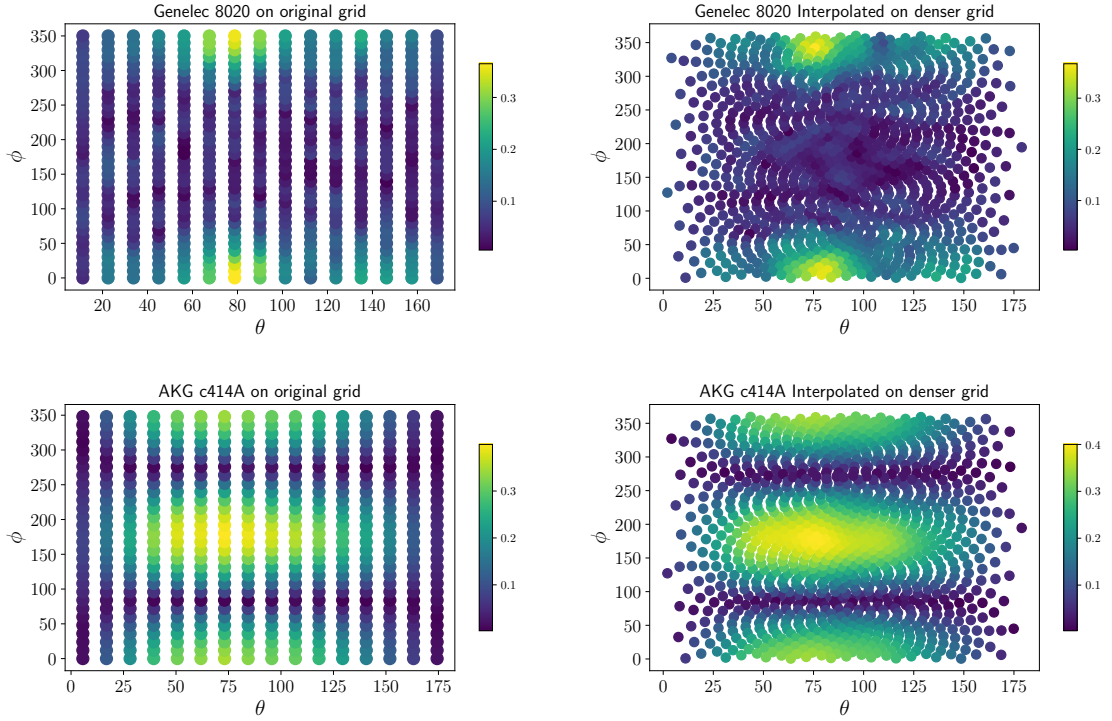

Figure 5.6: Input and output of DSHT interpolation of a spherical function presented as a 2D spherical heatmap at $f = 2000$ kHz. Left: shows the spherical function on the original grid. Right: Interpolated function on a Fibonacci grid. Top: Directivity pattern of the Genelec 8020 loudspeaker. Bottom: AKG C414A (Figure of eight) microphone.

5.3.3 Frequency domain RIR construction

Contrary to the original implementation of Pyroomacoustics, the following computations are made entirely in the DFT domain. Implementing frequency-domain RIR construction significantly changes the code and core functioning of Pyroomacoustics compared to the previous implementation of time domain RIR construction. We elucidate this new implementation in the form of pseudo-code for a single source-microphone pair. This contribution is built on top of the C++ engine, which provides us with image source positions $\mathbf{I}^{\text{pos}}$, wall attenuation in octave bands $\tilde{\mathbf{D}}$, the angle of arrival **AOA**, the angle of departure **AOD** and the delay $\mathbf{t}^{\text{delay}}$ for all image sources. Note that these variables are returned for the function calls of `get_pos_attn`, `get_angles` and `get_time_delay`, whose implementation is described in the previous pseudo-code in Implementation 1, thus the dimension of these matrices remains the same. In the psuedo-code in Implementation 2 below, the three functions `octave_band_fd_interp`, `fast_sinc_interp` and `get_response` provide individual parts for the generation of the RIR in the frequency domain.

`octave_band_fd_interp` takes one parameter, namely the wall attenuation in oc-

tave bands $\tilde{\mathbf{D}}$. It interpolates each row of the matrix to the full DFT frequency scale (F frequency bins) by

$$D_k(f) = \sum_{b=0}^{B-1} \tilde{D}_k(b)\zeta_b(f), \tag{5.11}$$

where $\zeta_b(f)$ denotes the half-cosine octave band filters used in the Pyroomacoustics toolbox. The interpolated filters are denoted by a matrix $\mathbf{D} \in \mathbb{C}^{K\times F}$. Equation (5.11) is illustrated by an example given in Figure 5.7, where the first figure shows the half cosine filters for multiple octave bands, the second figure illustrates the wall attenuation provided in octave bands for an image source $\tilde{D}_k(b)$, and the third figure shows the interpolated wall attenuation filter in the frequency domain.

`fast_sinc_interp` takes $\mathbf{t}^{\text{delay}}$ as a parameter. It decomposes the time delay for each image source k into an integer delay t_k^{int} and a fractional delay t_k^{frac}. t_k^{frac} is used to calculate the interpolated sinc function, which is windowed by the Hann function, and the results of the operation are returned as a windowed fractional delay filter denoted by $\mathbf{F} \in \mathbb{R}^{K\times F}$. To save computation time this function efficiently uses a pre-calculated sinc lookup table.

`get_response` requires either the angle of departure **AOD** or the angle of arrival **AOA**. With respect to the provided angles, the function does a *nearest neighbor search* on a K-d tree (Zhou et al., 2008) made out the angles $(\theta_i^{\text{fibo}}, \phi_i^{\text{fibo}})$ for $i \in \mathcal{U}^{fibo}$. It returns the filters $g^{\text{fibo}}(\theta_i^{\text{fibo}}, \phi_i^{\text{fibo}}, t)$ closest to the queried angles. The returned filter is denoted by matrices $\mathbf{G}^{\text{src}} \in \mathbb{C}^{K\times F}$ for the source and $\mathbf{G}^{\text{mic}} \in \mathbb{C}^{K\times F}$ for the microphone. Other variables such as $\mathbf{X}$, P' hold the computed values while the **rir** is an empty vector.

Implementation 2 Frequency domain RIR construction based on our extended ISM model

$\mathbf{I}^{\text{pos}}, \tilde{\mathbf{D}} \leftarrow$ `get_pos_attn()`
$\mathbf{AOD}, \mathbf{AOA} \leftarrow$ `get_angles`$(\mathbf{I}^{\text{pos}})$
$\mathbf{t}^{\text{delay}} \leftarrow$ `get_time_delay`$(\mathbf{I}^{\text{pos}})$
$\mathbf{G}^{\text{mic}} \leftarrow$ `get_response`$(\mathbf{AOD})$ *# retrieve frequency-dependent gain for microphone*
$\mathbf{G}^{\text{src}} \leftarrow$ `get_response`$(\mathbf{AOA})$
$\mathbf{D} \leftarrow$ `octave_band_fd_interp`$(\tilde{\mathbf{D}})$ *# interpolate attenuations to frequency scale*
$\mathbf{F} \leftarrow$ `fast_sinc_interp`$(\mathbf{t}^{\text{delay}})$ *# retrieve windowed fractional delay filter*
for $k \leftarrow 1$ to K **do** *# loop over all image sources*
  $\mathbf{X} \leftarrow \mathbf{D}[k,:] \odot \mathbf{F}[k,:] \odot \mathbf{G}^{\text{src}}[k,:] \odot \mathbf{G}^{\text{mic}}[k,:]$
  $P' \leftarrow (F + t_k^{\text{int}})$
  $\mathbf{rir}[t_k^{\text{int}} : P'] \leftarrow \text{IDFT}(\mathbf{X}) + \mathbf{rir}[t_k^{\text{int}} : P']$ *# add image source filter to the RIR*
end for

The functions `get_response`, `octave_band_fd_interp` and `fast_sinc_interp` do computations in a vectorized form for all the image sources calculated by the C++ engine. They output a matrix of size $K \times F$, that contains filters in the frequency domain for all

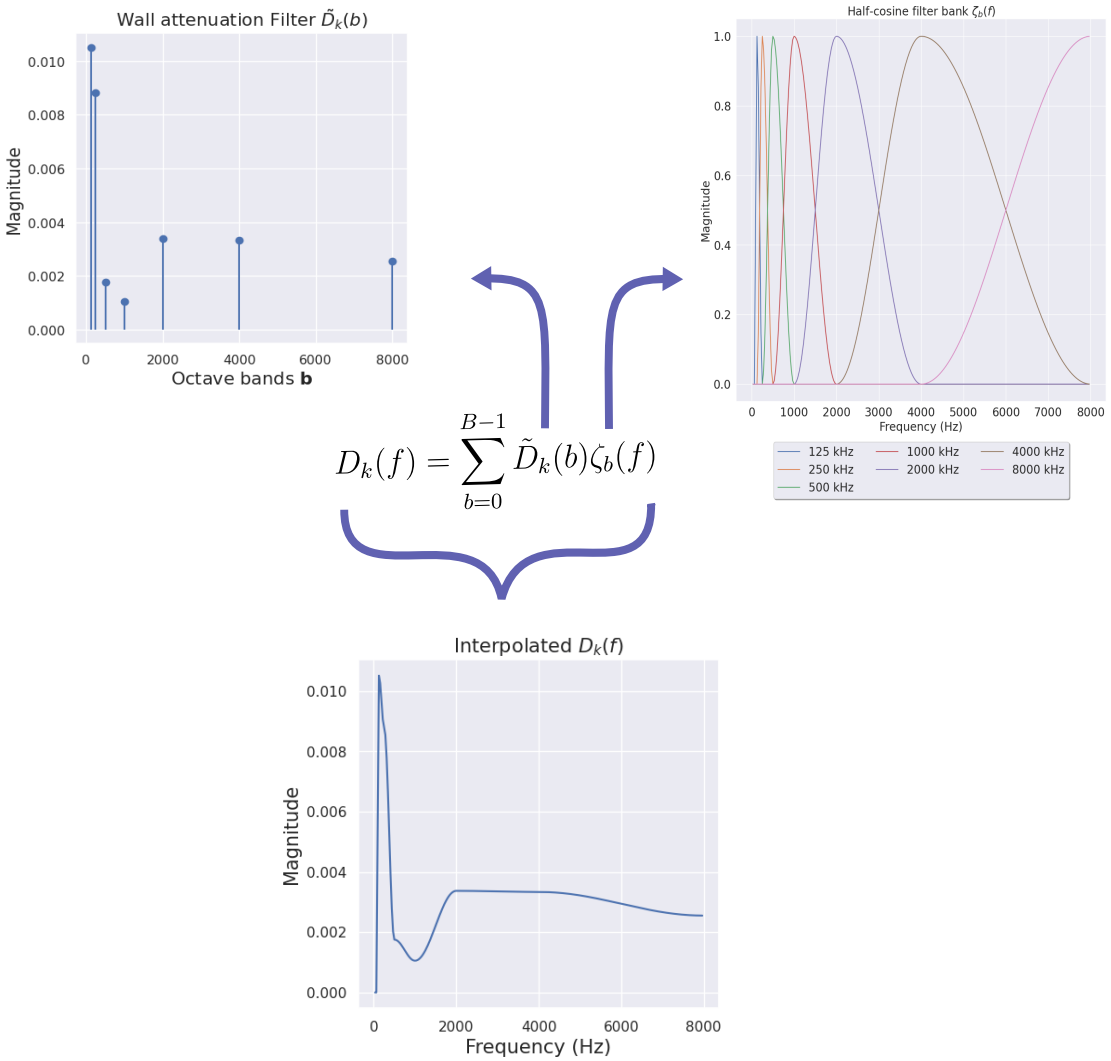

Figure 5.7: Example of interpolation of a wall attenuation filter from octave band $\tilde{D}_k(b)$ to frequency scale $D_k(f)$. Top: Half cosine filter banks and wall attenuations provided by pyroomacoustics. Bottom : Applying Equation (5.11) leads to the interpolated filter

image sources. The loop goes over all the image sources K and performs a linear convolution of all the filters in the frequency domain which results in an image source filter that is transformed into the time domain and placed at t_k^{int} in the vector **rir**. This approach also works in the following special cases:

1. For undefined source/microphone directivity, $\mathbf{G}^{\text{src/mic}} = 1$.
2. If one of the directivity patterns is chosen to be a frequency-independent pattern, $\mathbf{G}^{\text{src/mic}}$ becomes a scalar, whose value is obtained from Equation (5.2).

One important detail concerns the convolution of FIR filters in the frequency domain. A

convolution in the time domain is a multiplication in the frequency domain, and to be able to perform element-wise multiplication, the length of the filters should be the same and sufficiently large. In practice, the FIR filter length of the source and microphone directivities differ due to the differences in their measurement process. Therefore, we pad all time-domain filters with sufficient zeros, such that their lengths become equal and sufficient to obtain a linear convolution.

5.3.4 Added features

Apart from implementing the extended ISM in the frequency domain, we provide additional features to enhance the realism of the simulated RIRs. These additional features do not increase the computational overhead and have been implemented without modifying the aforementioned Algorithm 2.

Directivity Pattern Rotation

Directivity patterns are sensitive to the spatial orientation of the source or the receiver. It is therefore important to allow the user to specify the orientation of each source and receiver. When generating training data for virtually supervised learning, multiple random orientations can be simulated.

Each FIR filter on the Fibonacci grid is associated with a spherical angle $(\theta_i^{\text{fibo}}, \phi_i^{\text{fibo}})$. The spherical angles can be converted to Cartesian coordinates on the unit sphere, leading to the change in the function definition $\hat{g}^{\text{fibo}}(x_i, y_i, z_i, t)_{i \in \mathcal{U}^{\text{fibo}}}$. Let these new coordinates be denoted by the matrix $\mathbf{V} \in \mathbb{Z}^{3 \times I^{\text{fibo}}}$. Applying a suitable rotation matrix to $\mathbf{V}$ based on a requested orientation in azimuth $\theta^{\text{rotate}} \in [0, 2\pi]$ and in elevation $\phi^{\text{rotate}} \in [0, \pi]$ results in a transformed set of coordinates $\tilde{\mathbf{V}}$

$$\tilde{\mathbf{V}} = \begin{bmatrix} \cos(\phi^{\text{rotate}}) & 0 & \sin(\phi^{\text{rotate}}) \\ 0 & 1 & 0 \\ -\sin(\phi^{\text{rotate}}) & 0 & \cos(\phi^{\text{rotate}}) \end{bmatrix} \times \begin{bmatrix} \cos(\theta^{\text{rotate}}) & -\sin(\theta^{\text{rotate}}) & 0 \\ \sin(\theta^{\text{rotate}}) & \cos(\theta^{\text{rotate}}) & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \mathbf{V} \,. \tag{5.12}$$

This makes the FIR filters point to a new set of points in $\tilde{\mathbf{V}}$, resulting in a rotation of the directivity pattern without the need to recompute the spherical harmonic interpolation.

Minimum Phase Filters

As shown above, Equation (5.11) interpolates non-negative wall attenuations given in octave bands to the complex-valued discrete Fourier domain. The half-cosine interpolation scheme given in Equation (5.11) yields non-negative filters in the DFT domain, i.e., *zero-phase* filters. This implies that the wall attenuations are non-causal and cause an important artificial delay in the sound propagation (see Figure 5.8, left). Alternatively, Schimmel et al. (2009) in his room acoustic simulator proposes to use *minimum-phase* wall attenuation

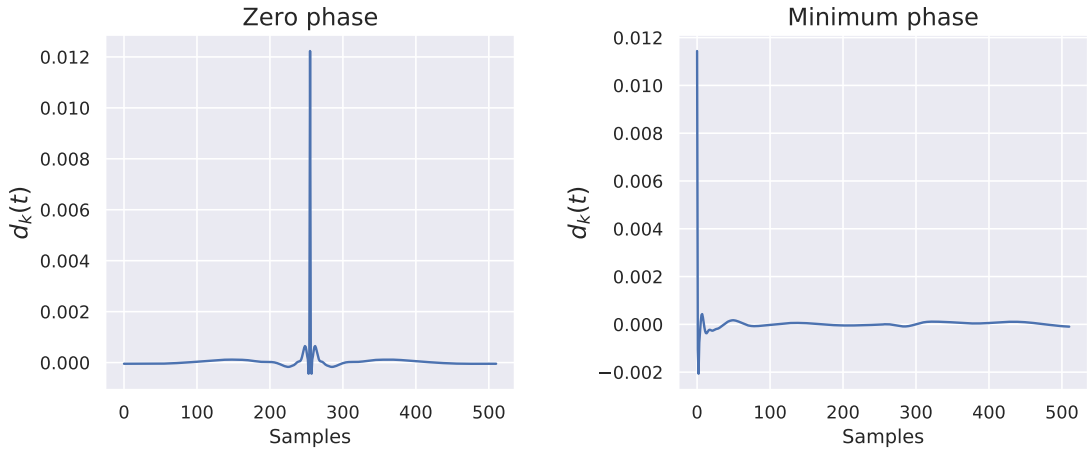

Figure 5.8: Example wall attenuation filter $d_k(t)$ with the "zero phase" (left) and "minimum phase" (right) settings.

filters. Minimum phase filters are, by definition, causal and stable with a causal and stable inverse, and they are the fastest decaying filters for a given set of Fourier magnitudes. For a given Fourier spectrum in the DFT domain $\mathbf{d} = [D(0), \ldots, D(F-1)] \in \mathbb{C}^F$, the corresponding minimum phase filter is obtained using the following formula :

$$\psi' = \mathrm{Im}[\mathrm{Hilbert}(-\log(|\mathbf{d}|))], \tag{5.13}$$

$$\tilde{\mathbf{d}} = |\mathbf{d}| \times e^{1j\cdot\psi'}, \tag{5.14}$$

where Hilbert denotes the Hilbert transform. The transformed minimum phase filter is shown in Figure 5.8 (right).

5.4 Qualitative analysis of obtained simulated RIRs

In Section 5.3 we presented the implementation of a frequency-domain RIR simulation method. To verify its functioning and its effectiveness in generating realistic RIRs, multiple qualitative tests have been performed, as described below.

5.4.1 Comparison between the original and the modified version of Pyroomacoustics

The original implementation of Pyroomacoustics has been used in a large number of works over the past few years. In this section, we take this implementation as a baseline and compare it to our proposed frequency-domain implementation under identical conditions.

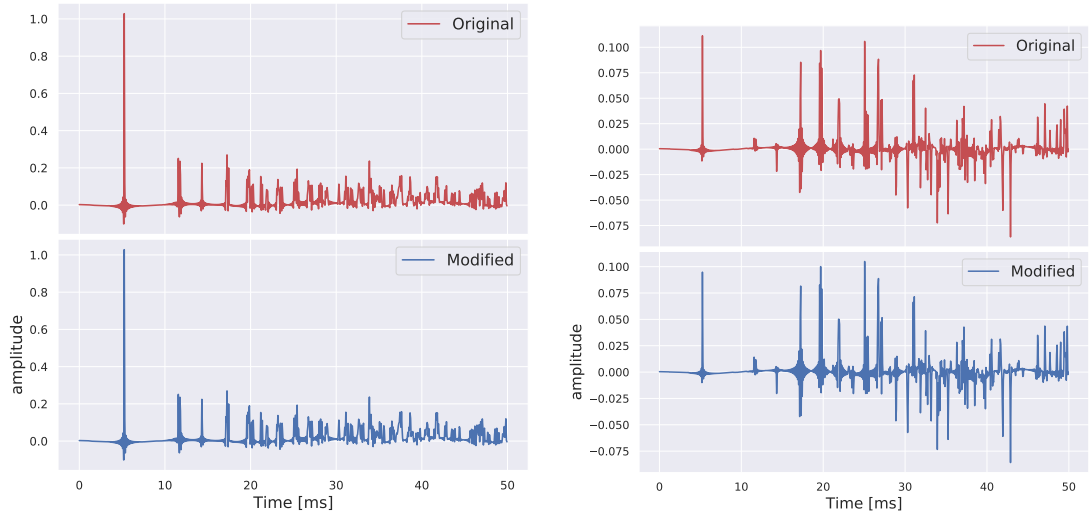

Figure 5.9: Qualitative comparison between the RIRs generated using the original and modified Pyroomacoustics. Left: Test condition 1, Right: Test condition 2.

This qualitative analysis helped us to debug and verify the functioning of our implementation of the advanced ISM in the simulator. These tests have been done under 2 different test conditions :

1. **Reflective omnidirectional.** In this condition, RIRs are simulated with omnidirectional sources and microphones in a shoebox room consisting of frequency-dependent walls. The aim is to verify the level of attenuation and the time of arrival of the image source filters between the two approaches. As it can be seen in Figure 5.9 (left) the image source filters exactly match each other, verifying that our frequency-domain implementation produces similar results compared to the time-domain one.
2. **Source and microphone directivity.** In these conditions, the purpose is to ensure the perfect working of the `get_response` function that performs DSHT interpolation and uses the nearest neighbor algorithm to return filters for queried angles of image source in Algorithm 2. The original version of Pyroomacoustics is able to simulate RIRs with frequency-independent source and microphone directivity patterns (see Section 5.1). We replicate the same with our modified version of Pyroomacoustics using the frequency-domain RIR simulation approach. One new SOFA file was created for each of the source and the receiver, using the same spherical grid as in the DIRPAT dataset with cardioid and figure-of-eight directivity pattern. The FIR filters at each point are one-sample filters whose values are given by the analytical directivity pattern in Equation (5.2). The two SOFA files implementing frequency-independent directivity patterns are employed to simulate RIR with our frequency-domain RIR construction method for a shoebox room with frequency-dependent walls. Another RIR is generated using the original version of Pyroomacoustics on similar acoustic

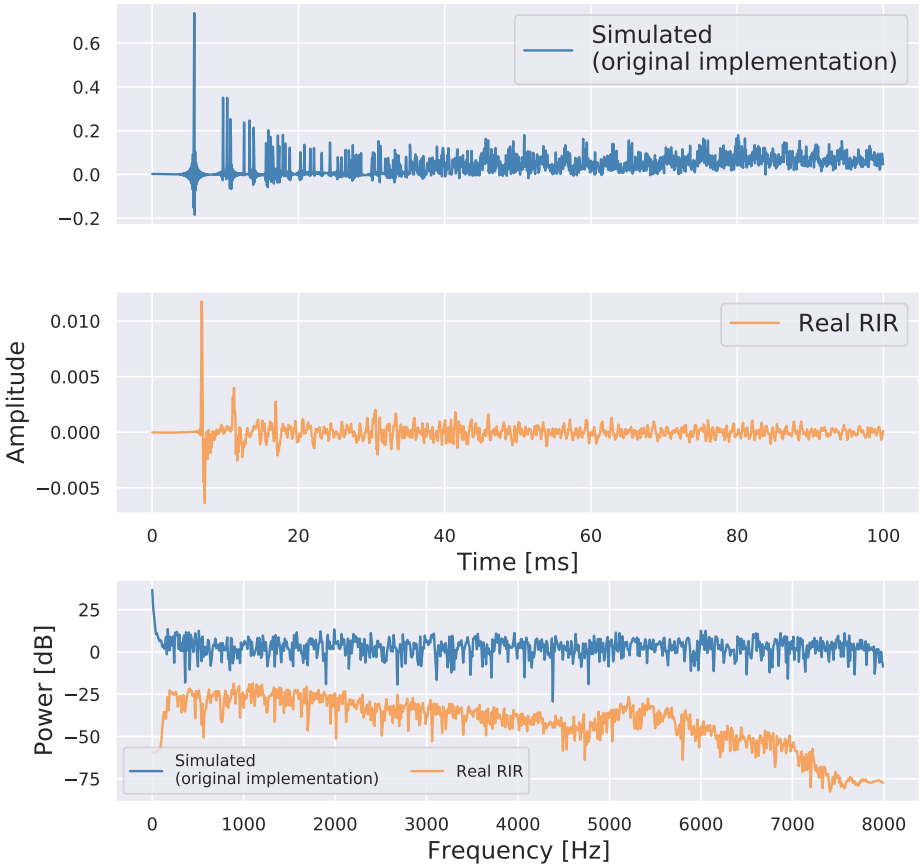

Figure 5.10: Qualitative comparison between a real RIR and a simulated RIR. The real RIR is taken from the "011111" room of the dEchorate dataset. Simultaneously, the simulation of the RIR is conducted within an acoustic environment that closely resembles the characteristics of the real RIR. This simulation employs time-domain processing and uses the original Pyroacoustics simulator. The first two rows are RIRs and the third row shows the frequency response of both the RIRs.

conditions and directivity patterns. The plot on Figure 5.9 (right) shows that RIRs simulated by both approaches have a high correlation with small coloration between the RIRs, despite the two RIR construction approaches being significantly different. This illustrates the proper functioning of the nearest neighbor, DSHT, and interpolation routines.

5.4.2 Similarity to measured RIRs

To validate our claim that incorporating real directivity patterns enhances the realism of RIRs, we conducted qualitative experiments. In our search for annotated RIR datasets

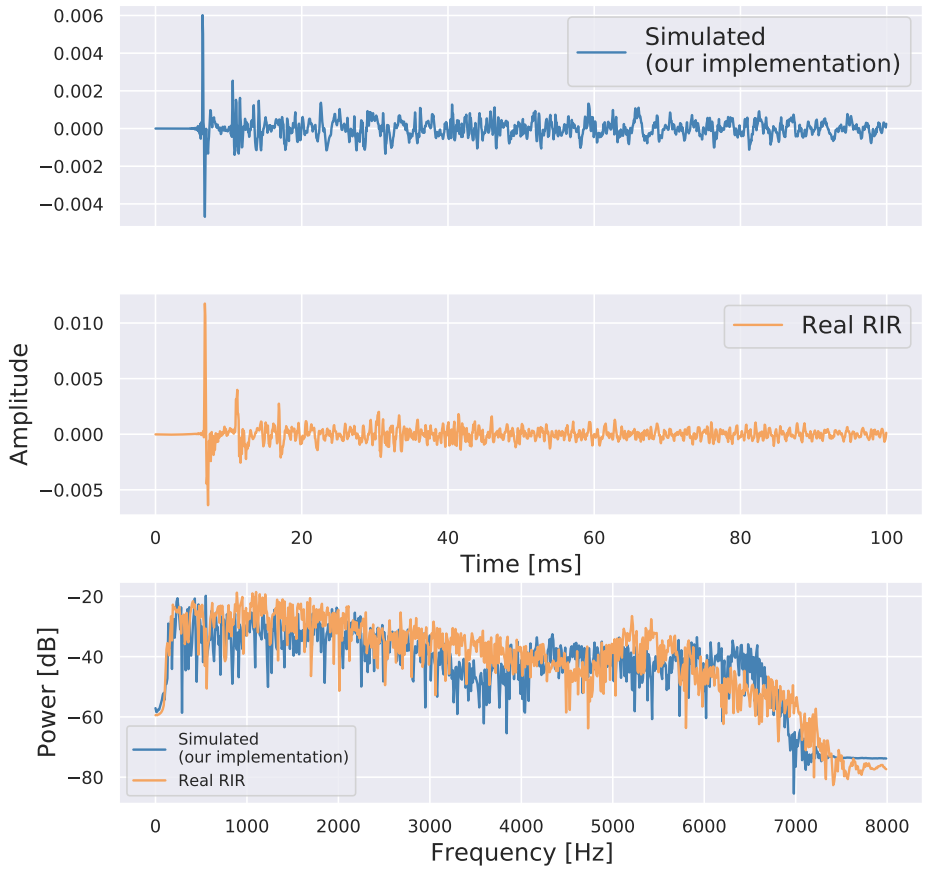

Figure 5.11: Qualitative comparison between a real RIR and a simulated RIR. The real RIR is taken from the "011111" room of the dEchorate dataset. Simultaneously, the simulation of the RIR is conducted within an acoustic environment that closely resembles the characteristics of the real RIR. This simulation employs frequency-domain processing and incorporates a measured source and microphone from our proposed modified Pyroomacoustics simulator. The first two rows are RIRs and the third row shows the frequency response of both the RIRs.

that could be mimicked by a room acoustic simulator, we discovered that the dEchorate dataset (Di Carlo et al., 2021) fits well our requirements. We replicated the acoustic scene of room "011111" from the dataset in our modified version of the simulator. The absorption profiles of the walls, room dimensions, and positions of microphone arrays and sources were meticulously recreated in the virtual acoustic environment. Although the directivity patterns of the microphones and sources used in recording the real RIRs were not provided in the dataset, the dataset did specify the references of the omnidirectional microphones and directive loudspeakers employed. Using this information, we selected an omnidirectional AKG c414 microphone and Genelec 8020 loudspeaker as substitutes in our virtual room

setup. These chosen directivity patterns closely resemble those employed in the measured RIR dataset.

With some post-processing, we aligned the RIRs on the time axis. Figure 5.11, shows the real RIR from the dataset and a RIR produced by our modified simulator. Comparing direct and first-order reflections, we see that both RIRs match well in the time domain. In the log-magnitude frequency domain, a high correlation can be observed throughout the frequency scale. This contrasts with Figure 5.10, where simulating the same acoustic scene with the original Pyroomacostic implementation shows poor resemblance to the real RIR in both time and frequency domains.

5.5 Further enhancements and improvements

Our implementation of the extended ISM in Pyroomacoustics improves the realism of simulated RIRs. Qualitative results presented in the previous section provide support to our claims. Still, there exist many fronts where the simulators can get better. Some directions include:

- combining ISM and stochastic ray tracing similarly to Schimmel et al. (2009), but with directive source and microphones;
- incorporating complex impedances of walls, resulting in directive wall responses;
- extension to non-shoebox rooms;
- inclusion of diffraction effects;
- support of a wider variety of measured source and microphone directivity patterns;

Computational time

While implementing the frequency-domain RIR construction (Algorithm 2), an important factor to consider was the computational time. Our goal was to achieve a computational time that is comparable to the time-domain RIR construction method. Due to the design of Algorithm 2, which involves iterating through all the image sources and multiple conversions between the time domain and the Fourier domain, the initial implementation in Python took around 20 seconds to simulate a RIR with an image source order of 20. While the original version of Pyroomacoustics generates RIR in the time domain and does not take into account the frequency-dependency of the source and microphone takes around 1 seconds.

To optimize computational time, we identified some bottlenecks in the Python code, such as high-dimensional array broadcasting, and ported them to Cython. Through multiple iterations of code refactoring, we also managed to vectorize certain parts of the algorithm, which further improved the performance. This significantly reduced the computational time and one RIR is now generated in 1.5 seconds. However, there is still room for improvement, as there are additional functions that could potentially be vectorized,

allowing for even faster execution times, especially with higher image source orders.

5.6 Summary

This chapter provides a detailed description of our contribution towards the Pyroomacoustics simulator. The chapter shows that employing an extended ISM with the inclusion of the measured source and receiver directivities improves the realism of the simulated RIRs without increasing the computational time. The key changes in this implementation were the DSHT interpolation and the frequency-domain RIR construction method. The results in Section 5.4 show that our implemented method makes one step in the direction of obtaining a digital twin of a real room. Observing positive results, this improved simulator will be used in the next chapter to train virtually-supervised DNN models on two different tasks and its effect on the generalization of the system will be assessed.

6 Impact of simulation realism on virtually supervised learning

This chapter details two contributions that focus on the effect of simulation realism on virtually supervised audio processing systems. Simulations of the training sets are made more realistic with the help of our advanced ISM simulator presented in Chapter 5. The impact of the realistic simulation is assessed on two different tasks, namely, room parameter estimation and sound source localization. Throughout this study, we enhance the simulation of the datasets by adding realism to the source, receiver, and walls of the virtual acoustic space. Realism in source and receivers is achieved by adding different types of frequency-dependent directivity patterns, while for the walls, floor, and ceiling, realistic distributions of absorption coefficients combined with minimum phase wall responses are considered. For the task of room parameter estimation, we use the same training procedures and DNN architecture as presented in Chapter 4. For sound source localization, we focus on direction of arrival (DOA) estimation using one microphone pair. A state-of-the-art, broadly applicable DNN architecture is used as the model. The main aim of this chapter is to study the generalization performance of models purely trained on various extensions of the ISM on real data. For this purpose, this study effectively tests the models on 4 different real test sets that include human speakers. Apart from the results obtained with a baseline, a naive and an advanced trained model, this chapter also presents an ablation study on both tasks to determine the impact of each layer of added realism to the training dataset. This chapter is divided in two parts. First, Section 6.1 presents the study on room parameter estimation, which is further divided into three subsections. Section 6.1.1 provides details on simulated data, Section 6.1.2 focuses on training and hyperparameters, and Section 6.1.3 presents the results and experiments on both simulated and real data. The second part in Section 6.2 concerns source localization, which starts with Section 6.2.1 explaining the principle of DOA estimation. Section 6.2.2 briefly describes the three real datasets, Section 6.2.3 provides details on scenario-based dataset simulation on three real datasets, Section 6.2.4 describes model selection and training parameters, and the experiment and results are presented in Section 6.2.5. At last Section 6.3 present a conclusion to both tasks and points to key takeaways.

6.1 Room parameter estimation

In this section of the chapter, we examine the generalization ability of a room parameter estimation model on real-world datasets. The model is trained using data generated from the extended ISM simulator presented in Chapter 5. Also, an ablation study is carried out on 7 different simulated training datasets, assessing the impact of experimenting with different source, receiver directivity profiles, and wall absorption coefficient distributions. The datasets are used within the room parameter estimation pipeline established in Chapter 4. Similar to the previous study, we estimate the room surface area S, volume V, and reverberation time $\mathrm{RT}_{60}(b)$ for all $b \in [125, 250, 500, 1000, 2000, 4000]$ Hz given a set of noisy speech signals. Contrary to the previous study, we exclude the estimation of mean absorption coefficients $\overline{\alpha}$ due to the lack of availability of real test sets with a proper annotation for this parameter.

6.1.1 Simulated data

The experiments on room parameter estimation shown in Chapter 4 use the Roomsim simulator (Schimmel et al., 2009) and simulate RIRs using the ISM combined with stochastic ray tracing. These experiments considered omnidirectional sources and receivers only. However, it has been noted in the study by Knüttel et al. (2013) that source and receiver directivities have a strong impact on a RIR. Therefore, in this chapter, we use our advanced Pyroomacoustics simulator described in Chapter 5. The datasets are simulated at a large scale using the frequency-dependent source and microphone directivity patterns from the DIRPAT directivity dataset (Brandner et al., 2018). The RIRs are also simulated using realistic wall responses.

6.1.1.1 Training sets used for the ablation study

Seven different datasets named $\{\mathrm{D1}, \ldots, \mathrm{D7}\}$ are simulated for this ablation study. Each dataset is structured to incrementally introduce varying levels of realism in the source, receivers, and walls. Table 6.1 summarizes the different levels of realism for each dataset using specific notations, which are explained below.

Source directivity. Three types of source directivity are considered in the training datasets, namely omnidirectional, frequency-independent, and measured frequency-dependent responses. Frequency-independent directivity patterns are implemented using the general analytical formula in Equation (5.2). For this experiment, a value $\psi \in [0.25, 0.5, 0.75]$ is randomly selected for each simulated source, respectively yielding hypercardioid, cardioid, and subcardioid patterns. In Table 6.1, such analytical patterns are denoted as $(\mathcal{A}_\psi)$ and are only used in dataset D4. Additionaly, omnidirectional sources correspond to the value $\psi = 1$ in Equation (5.2). This is denoted as $(\mathcal{O})$ and is used in multiple datasets from D1-D5. It should be noted that the directivity patterns formed by the analytical formula do not take into account the elevation dependency nor, more importantly, the

Datasets	Walls	Source	Microphones
D1	$\mathcal{N}$	$\mathcal{O}$	$\mathcal{O}$
D2	$\mathcal{RB}$	$\mathcal{O}$	$\mathcal{O}$
D3	$\mathcal{RB}$	$\mathcal{O}$	$\mathcal{M}$
D4	$\mathcal{RB}$	$\mathcal{A}_\psi$	$\mathcal{O}$
D5	$\mathcal{RB}$	$\mathcal{M}$	$\mathcal{O}$
D6	$\mathcal{N}$	$\mathcal{M}$	$\mathcal{M}$
D7	$\mathcal{RB}$	$\mathcal{M}$	$\mathcal{M}$

Table 6.1: Ablation study datasets and associated notations describing the level of realism of each dataset. In the table $\mathcal{N}$ and $\mathcal{RB}$ denote naive and reflectivity-biased wall sampling (Foy et al., 2021) . $\mathcal{O}$,$\mathcal{A}_\psi$ and $\mathcal{M}$ denotes omnidirectional, analytical and measured directivity patterns.

frequency dependency. Measured directivity patterns are taken from the DIRPAT dataset (Brandner et al., 2018), as described in details in Section 5.3.1. Measured source patterns were used in the datasets D5, D6, and D7, where a random directivity profile was selected for each simulated source from three loudspeakers: Genelec 8020, Neumann KH120A, and Yamaha DXR8. In Table 6.1, this is denoted by $(\mathcal{M})$. For each simulated source its directivity patterns is randomly rotated in the azimuth angle, parallel to the ground.

Receiver directivity. Omnidirectional $(\mathcal{O})$ and measured directivity patterns $(\mathcal{M})$ are considered for the receivers in the ablation study. Omnidirectional patterns are defined in the same way as for the source and are used in datasets D1, D2, D4, D5. Measured receivers are used in D3, D6, and D7, which utilized the frequency-dependent omnidirectional directivity pattern of the AKG C414 microphone from the DIRPAT dataset. In the ablation study, for receivers, we excluded any type of non-omnidirectional patterns, both frequency dependent and independent. This decision was made because the real test set only consists of omnidirectional receivers.

Absorption Profiles. Similarly, two different types of wall sampling strategies are considered: Naive sampling $(\mathcal{N})$ and *Reflectivity-Biased* sampling $(\mathcal{RB})$ (Foy et al., 2021). In the first strategy, each of the six surfaces in a shoebox room is associated with a single frequency independent coefficient $\alpha \in [0.02, 0.5]$, that is drawn uniformly at random. The second strategy yields realistic rooms as each wall of the room is associated with frequency-dependent absorption coefficients α given in six octave bands. These values are sampled uniformly in ranges defined according to a dataset of commonly encountered room surfaces (see Table 4.1). A similar sampling strategy on absorption coefficients is employed in our previous experiment on room parameter estimation, as detailed in Section 4.1.1.

6.1.1.2 RIR simulation and mixture generation

Each dataset consists of $30,000$ different rooms whose length, width and height are drawn at random from the box $[3, 10] \times [3, 10] \times [2, 4.5]$ in meters. In each room, 3 two-channel RIRs are simulated at a rate of 16 kHz. An array of two microphones placed 22.5 cm apart is used, similar to the one used in Chapter 4. The microphone array is placed at three different positions with a fixed source position. The directivity patterns for the microphones are randomly oriented on the sphere. The frequency-dependent and frequency-independent non-omnidirectional patterns of the sources were randomly oriented in only the azimuth direction. The RIRs are used to generate noisy mixtures. To do so, the two-channel RIRs are convolved with speech excerpts from the Librispeech corpus (Panayotov et al., 2015). Uncorrelated white Gaussian noise with SNR in $[60, 70]$ dB and diffuse speech-shaped noise with SNR in $[20, 50]$ dB are convolved with the late part of a random two-channel RIR in the room. Both noise components are added to the reverberated signal. To calibrate the noise levels, a reference speech signal is used, where the emitter is placed 1 M away and facing a receiver placed in the middle of a $5 \times 5 \times 3$ m room with an absorption coefficient of 0.2. This calibration process ensures that the signal-to-noise ratios across the datasets fall within the $[15, 75]$ dB range. 3 measurements per room for 30,000 different rooms resulted in 90,000 multichannel RIRs for each dataset. The 7 training datasets used in the ablation study sum up to a total of 630,000 speech signals.

6.1.2 Training and hyperparameters

To perform the task, we reuse our multi-channel multi-task state-of-the-art DNN model described in Section 4.2. The parameters to extract features from the speech signals and the pipelines to process the single/multi-channel features remain the same. The loss function and the Bayes formula used to fuse multiple observations from one room also remains the same, however instead of fusing multichannel estimates from five observation, we chose to fuse three observations to get a single estimate. The results presented in Chapter 4 points that the gain in performance with the fusion of 5 estimate compared to the fusion from 3 estimates is not huge, hence for this experiment we fused 3 estimates to get a single estimate.

The network is trained with the ADAM optimizer and a learning rate of 10^{-4} with a batch size of 16. To prevent overfitting, a dropout rate of 0.5 is applied between the convolutional blocks. Additionally, l_1 and l_2 regularization are applied on the network weights, with a weight of 10^{-5} and 10^{-3} respectively.

6.1.3 Experiments and results

Seven different models are trained on the seven training sets $\{\text{D1}, \ldots, \text{D7}\}$ summarized in Table 6.1. The models are tested on two different test sets and results are provided in terms of the mean absolute error on all the estimated quantities. The results for RT_{60}

are only provided in the octave band range of $[500\text{Hz} - 4\text{kHz}]$, due to the inaccuracy of geometric acoustic simulators below the Schroeder frequency, which is around 500 Hz for the real test room considered in this study.

6.1.3.1 Simulated test set

A simulated test set is created with 400 shoebox rooms consisting of 1,200 multi-channel recordings. The rooms are simulated with a similar level of realism to D7. The microphones are simulated using the measured omnidirectional directivity pattern of AKG C414. The source directivity in each test sample is randomly chosen from 3 out-of-training measured directivity patterns, namely, 2 loudspeakers Tannoy system 1200 and Lambda Labs CX-1A and 1 head and torso mouth simulator Bruel & Kjaer 4128C. We refer to this simulated test set as realistic. The results obtained with seven virtually supervised models are shown in the first row of Table 6.2.

6.1.3.2 Real test set

We also used a real test dataset to assess the impact of the different levels of realism in our training sets. The goal is to identify the specific level of realism that increases the generalization ability of the system. The dEchorate dataset (Di Carlo et al., 2021) is used as a real test set. The dataset is recorded in a modular rectangular room of $S = 125$ m^2 and $V = 82$ m^3, whose walls, floor and ceiling can be switched between the reflective and absorbing mode. Specific details on the source and microphone setup that performed the measurements in the dEchorate dataset are found in Section 3.5. Contrary to the experiments in Chapter 4, more rooms from the dataset were used for this test. Out of 11 rooms we selected 4 rooms with 2-5 reflective surfaces. The range of RT_{60} in these rooms is $[250, 810]$ ms. This range is commonly observed in rooms such as office rooms, classrooms, and meeting rooms. Six semi-anechoic rooms and one furnished room were excluded from this dataset. The reverberation times of the anechoic rooms do not represent commonly encountered conditions, while the furnished room seemed to feature inconsistent ground truth values[1]. The dataset is recorded with AKG CK32 omnidirectional microphones, while the speech is emitted by six directional sources and one omnidirectional source, namely, Avanton Mix Cubes and a Bruel & Kajer omnidirectional loudspeaker. Each source is emitting 4 different speech utterances from the Wall Street Journal (WSJ) dataset (Paul and Baker, 1992). The directivity patterns of these sources and microphones are not present in the DIRPAT dataset. Hence, none of our seven training datasets includes these directive patterns. We consider three two-microphone sub-arrays with an aperture of 22.5 cm from six microphone arrays consisting of 5 microphones each. Combining all sub-arrays, all sources and all utterances yields $7 \times 5 \times 4 = 140$ 2-channel real speech recordings that

[1]The measurements were taken in typical meeting room, however the RT_{60} ground truth measurements were inconsistent with the one taken in similar room dimension in the dataset.

are cut to three second length for each room. Combining the 4 rooms, there are hence 560 test cases in total.

6.1.3.3 Results

The mean absolute error on the two test sets for all the estimated quantities and the seven trained models is presented in Table 6.2. Upon observing the table, it is evident that the model trained on the most realistic training set, D7, yields the lowest errors as expected. This model consistently achieves the best or second-best performance for all estimated quantities on both the simulated and real test sets. Another noteworthy observation is the comparison between D5 and D7. While D5 shares similar wall and source characteristics as D7, it employs a simpler frequency-independent omnidirectional microphone model. Surprisingly, D5 outperforms D7 in terms of the estimation of S and V on the real test set. This discrepancy for D7 can be attributed to a mismatch in the microphone profile between the training and test sets, which seems to negatively impact the performance of the system. A possible solution could be to use a diverse set of microphone directivity patterns in the training set. Subsequently, we do three different comparisons between the trained models to highlight important findings. First, let's compare D2 and D3 on the one hand with D5 and D6 on the other hand. This highlights the improvement in RT_{60} estimation achieved by using measured microphone directivity. However, this improvement comes at the expense of degraded estimations for S and V, as mentioned before. Second, comparing D1 and D2 on the one hand with D6 and D7 on the other hand, we note that the realistic wall sampling strategy ($\mathcal{RB}$) employed in D2 and D7 consistently demonstrates better performance compared to the naive sampling strategy ($\mathcal{N}$) in D1 and D6 throughout all the quantities and in both test sets. Lastly, when comparing D2, D4, and D5, we observe a consistent improvement in estimation results by incorporating realism into the source directivity. This improvement is evident across all quantities and on both test sets. This claim is further supported by comparing D3 and D7, which demonstrates a significant improvement in the estimation of S and V when using a more realistic source directivity.

6.2 Sound source localization

We now turn our attention to the task of source localization to assess the performance improvements that can be achieved by training a localization system using simulated data from our improved version of the Pyroomacoustics simulator. The study by Vincent et al. (2017) on speech enhancement points out that a broader generalization of DNN models is achievable when the training set encompasses examples from a diverse range of acoustic environments, including those that are representative of the ones encountered in the test set. Simulations are the preferred approach to meet the demand for diverse training data, given the limited availability of real data for various types of arrays and acoustic conditions. For this purpose, room acoustic simulators have been widely used for the application

Training dataset	Realistic test set					
	RT_{60} (500 Hz)	RT_{60} (1 kHz)	RT_{60} (2 kHz)	RT_{60} (4 kHz)	S	V
D1	0.182	0.150	0.150	0.139	97.04	98.47
D2	0.186	0.189	0.228	0.226	69.28	75.56
D3	0.198	0.158	0.133	0.110	103.57	108.58
D4	0.170	0.138	0.151	0.155	77.61	84.89
D5	0.168	0.143	0.153	0.119	50.46	53.09
D6	0.152	0.177	0.177	**0.098**	37.88	41.08
D7	**0.134**	**0.105**	**0.116**	**0.092**	**25.22**	**28.63**

Training dataset	Real test set					
	RT_{60} (500 Hz)	RT_{60} (1 kHz)	RT_{60} (2 kHz)	RT_{60} (4 kHz)	S	V
D1	0.193	0.160	0.108	0.185	71.00	75.68
D2	0.182	0.140	0.128	0.198	45.11	55.16
D3	0.115	**0.098**	0.078	0.156	52.76	61.82
D4	0.167	0.134	0.121	0.197	37.91	48.95
D5	0.133	0.112	**0.066**	0.155	**21.46**	**18.57**
D6	0.151	0.133	0.084	0.159	35.88	31.11
D7	**0.080**	**0.103**	**0.064**	**0.140**	32.69	30.57

Table 6.2: Mean absolute errors in reverberation time (RT_{60}, in s), surface (S, in m^2) and volume (V, in m^3) estimation achieved over a realistic test set (top) and real test set (bottom) using the same model trained on 7 simulated training datasets. Bold numbers indicate the best statistically significant result per column, based on 98% confidence intervals.

of sound source localization (Wu et al., 2021; Subramanian et al., 2022), as they allow the modelling of different aspects of RIRs such as the amount of reverberation, interchannel level differences, and source-to-microphone distances which are crucial for localization purposes. Section 3.4 reviewed the vast variety of existing room acoustic simulators. Among them, ISM has seen widespread use in the literature because of its implementation simplicity and its ability to generate a variety of shoebox rooms with randomized dimensions, wall absorption coefficients, and source-receiver positions. The ISM has been shown to perform well on training sound source localization models that then give good performance on real test sets (Adavanne et al., 2018b; Chakrabarty and Habets, 2019; Perotin et al., 2019; Diaz-Guerra et al., 2020). The majority of these studies have used the simplest version of the ISM, simulating RIRs with omnidirectional receivers and sources in a virtual room with frequency-independent absorption coefficients for the walls. Only few studies on sound source localization using special multichannel receivers, such as spherical microphones (Koyama et al., 2022), Ambisonics (Adavanne et al., 2019; Perotin et al., 2019) and binaural receivers (Gaultier et al., 2017; Ding et al., 2020) incorporated the directivity of specific receivers in the training of these systems. Despite the widespread use of shoebox ISM-based simulators for training source localization systems, the effect of integrating more realistic simulation conditions during training and its performance on a real test set have hardly been studied.

More realistic conditions, such as directional sources, receivers, and frequency-dependent walls have already been successfully integrated as an extended ISM model in our improved version of the shoebox simulator. Section 6.1 shows that, when used for the purpose of training a virtually supervised model for acoustic parameter estimation, our advanced simulator shows improved generalization on real test scenarios. Similarly, for source localization, a recent work by Gelderblom et al. (2021) analyzes the effect of source directivity and diffuse late reverberation modeling in the RIR simulation. Source directivity was found to have a positive influence on the performance of the localization system while the inclusion of the latter showed no impact. Additionally, their results were not made on direct human speech in real acoustic conditions but rather on speech signals convolved with measured directive RIRs. Apart from these two levels of realism, source localization is influenced by other facets of the acoustic scene. One such important factor is receiver directivity. Simpler omnidirectional microphones could induce significant effect in localization performance, due to the observed variability in directivity pattern at different frequencies (see Figure 2.2). Moreover, frequency-dependent absorption coefficients affect the reverberation level reaching the microphones which can be observed in the spatial distribution and power spectrum. Similar to the previous section an ablation study is performed to quantify the effect of each added layer of realism at training time. We provide results on three distinct real test sets with different microphone arrays recording real human speakers in various acoustic conditions to consolidate our findings.

6.2.1 Angle of arrival estimation

Within the broad field of source localization, we choose the task of estimating the angle of arrival in $[0°, 180°]$ of the impinging source wave to the microphone array, also called the 1D direction of arrival (DOA). Using a two-microphone array, as in previous experiments, limits us to performing 1D DOA estimation, because the intrinsic symmetry of the array creates ambiguities around its rotation axis. Therefore, the array perceives sources lying on a so-called *cone of confusion* (Wallach, 1939).

The DOA estimation of a single far-field source by an array of two microphones separated by a distance of l_{21} cm is shown in Figure 6.1. In free field conditions (see Section 2.4.1), the system in the frequency domain is written as follows:

$$\mathbf{C}(f) = \mathbf{E}(f)S(f) + \mathbf{A}(f)\,, \tag{6.1}$$

where $\mathbf{C}(f)$ is the Fourier transform of a multichannel vector-valued signal $\mathbf{c}[n]$ containing the signal $c_m[n]$ of each microphone. The noise component $\mathbf{A}(f)$ is also a vector incorporating individual microphone noises $a_m(f)$. The single source signal is denoted as $S(f)$. $\mathbf{E}(f)$ is called the steering vector. Each element in the vector is the DFT of the free-field acoustic impulse response (see Equation (2.20)) from a source to a microphone m.

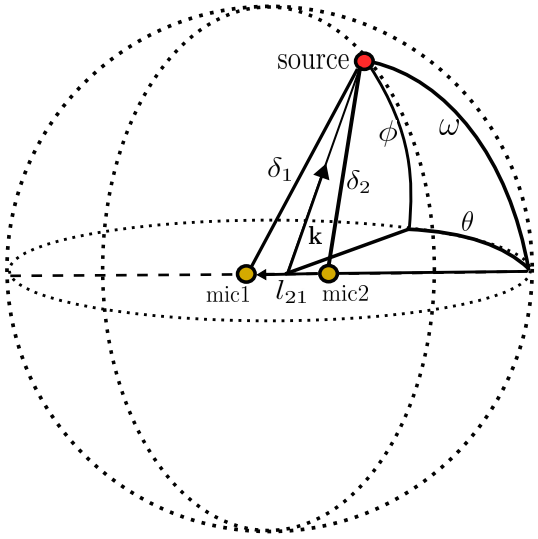

Figure 6.1: Illustration presenting DOA estimation of a far-field source from a pair of microphones on the horizontal plane. The distance between the microphones is l_{21}, i.e., the aperture. δ_1 and δ_2 are the source to microphone distances. The three angles are azimuth θ, elevation ϕ, and angle of arrival ω. Also, $\mathbf{k}$ defines a unit-norm vector pointing towards the source.

In the far-field setting,

$$\mathbf{E}(f) \approx \begin{bmatrix} e^{-2j\pi f\delta_1/cF} \\ e^{-2j\pi f\delta_2/cF} \end{bmatrix}, \tag{6.2}$$

where $\delta_m = \|\mathbf{r}^{\text{src}} - \mathbf{r}_m^{\text{mic}}\|_2$ is the Euclidean distance between the source and microphone m. The steering vector is directly related to the distance and DOA of the source relative to the microphone array. Each element $e_m(f)$ consists of a transfer function for a specific channel and has a specific amplitude and phase. If the phase and magnitude spectrum of the source is known, disambiguation between $e_m(f)$ and $S(f)$ can be exploited to do source localization. However, in practice, source signals are rarely known. In that case, it is sensible to calculate the ratio between the channels. By taking the first channel as a reference the relative steering vector is given as

$$\tilde{\mathbf{E}}(f) = \begin{bmatrix} 1 \\ e^{-2j\pi f\Delta_2/F} \end{bmatrix}, \tag{6.3}$$

where $\Delta_2 = \frac{\delta_2 - \delta_1}{c}$ is called the *time difference of arrival* (TDOA). In the far-field setting, the TDOA depends on the DOA of the source signal and can be expressed as

$$\Delta_2 \approx \frac{l_{21}\cos\omega}{c}, \tag{6.4}$$

where ω is the DOA of the source. GCC-PHAT is a commonly used algorithm for TDOA estimation from the signals of a microphone pair in reverberant conditions (see Equation (3.2)). With more than 2 microphones in a microphone array, a set of TDOAs is observed, which can be used by SRP-PHAT to localize a sound source (see Section 3.2.1).

6.2.2 DOA estimation on real test sets

To evaluate the influence of enhanced ISM realism during training, we assess the performance of virtually-supervised DOA estimation models on three real datasets with proper spatio-temporal annotations of the human speakers' activity and position relative to the microphone array. The aim was to select datasets that closely resemble real-world conditions. Therefore, we selected datasets that featured real human speakers rather than ones generated using anechoic speech signals convolved with measured directive RIRs. Guided by these requirements, we found three publically available datasets captured in a variety of rooms, each using a different microphone array. Namely, these are Distant-Speech Interaction for Robust Home Applications (DIRHA) (Cristoforetti et al., 2014), VoiceHome-2 (Bertin et al., 2019) and Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22) (Politis et al., 2022). The first two datasets are designed for the purpose of assessing smart home applications such as speech enhancement or speaker localization, while the latter is made for the DCASE sound event and localization challenge (Politis et al., 2020). A brief summary of these datasets can be found in Section 3.5. Additionally, from each dataset, we select a two-microphone sub-array for the task of DOA estimation, as detailed below :

1. The VoiceHome-2 corpus is recorded using a microphone array consisting of 8 MEMS that are placed near the corner of a cubic baffle. For this study, a two-channel sub-array with an aperture of 10.4 cm is selected, and 360 two-second speech recordings in quiet conditions are used.
2. The DIRHA corpus is captured using a network of omnidirectional wall-mounted microphone arrays placed on the walls and ceiling of many different rooms. For this study, a wall-mounted two-channel microphone array with an aperture of 30 cm placed in the living room is selected, and 410 two-second speech recordings from the living room are used.
3. The STARSS22 dataset is recorded using an Eigenmike spherical array[2] and is distributed in two formats: first-order Ambisonics and tetrahedral sub-array that selects the channels 6, 10, 26, and 22 of the Eigenmike. The arrangement of the selected individual microphones on the Eigenmike forms a tetrahedral shape. We carefully pre-processed the data to extract 2,100 two-second non-overlapping speech excerpts from microphones 6 and 10 out of the tetrahedral sub-array, with an aperture of 6.8 cm.

The combined duration of the three carefully curated test sets is 95 minutes, comprising real human speech recordings with two channels and DOA annotations. Details on the

[2]https://mhacoustics.com/products#eigenmike1

characteristics of each specific microphone array will be given in the next section.

6.2.3 Scenario-based data generation

6.2.3.1 RIR simulation

For each of the three test sets described in Section 6.2.2, one naive and one advanced simulated training set are built. The naive training sets consist of omnidirectional receivers and sources, where the apertures of the receivers are the same as the ones in their corresponding test sets. The wall absorption coefficients are assumed to be frequency-independent and equal for all six surfaces in the virtual room. This is similar to the simulation setting of D1, as described in Section 6.1.1.1. The advanced simulated training sets use our advanced version of the Pyroomacoustics simulator for RIR generation. These training sets incorporate more informed choices on the directivity and absorption components. For walls, the reflectivity-biased $(\mathcal{RB})$ absorption sampling strategy described in Section 6.1.1.1, is used. Regarding source directivity, the spatially interpolated measured directivities of a head-and-torso-with-mouth simulator (Bruel & Kjaer HATS 4128-C) and two directive loudspeakers (Genelec 8020 and YAMAHA DXR8) taken from the DIRPAT dataset (Brandner et al., 2018) are integrated into the simulation. Receiver directivities and aperture distance are associated with the particular scenarios found in the test sets. In the simulated training set designed for the Voicehome-2 test set, the receivers are set to be omnidirectional. Indeed, the directivity pattern of the micro electromechanical systems microphones (MEMS) used in VoiceHome-2 is not available, but is known to be close to omnidirectional. This choice is further supported by the results in our ablation study on room parameter estimation, where using mismatched microphone responses at training time was detrimental to the results. For DIRHA, the advanced simulation places the receivers on the room walls, which is equivalent to simulating microphones with a half-sphere directivity. Finally, in the simulated training set designed for STARS22, the advanced simulation utilizes the measured directivity pattern of the relevant sub-array of the Eigenmike (see Section 5.3.1).

The RIRs are simulated with an image source order of 20. A total of 40k shoebox rooms of sizes uniformly drawn at random in $[3, 10]\times[3, 10]\times[2, 4.5]$ meters are simulated, each containing a source and a two-microphone array placed uniformly at random with a minimum source-array and device-wall distance of 30 cm.

6.2.3.2 Mixture generation

The mixture generation process is similar to the one used in our study on room parameter estimation (see Section 4.1.2). The simulated RIRs are convolved with random speech signals from the Librispeech corpus. Diffuse speech-shaped noise and white Gaussian noise are added to the reverberated signals. The noise levels are tuned based on a reference signal that is simulated in an ideal acoustic environment, similar to the previous

experiment (see Section 4.1.2). However, according to the considered scenario the microphone array distance and directivity is adjusted, and a random source directivity pattern is chosen from the training set. This approach resulted in bell-shaped signal-to-noise ratio distributions in $[15, 75]$ dB with a peak at 40 dB for each of the six training sets considered in this study. This generation process yielded 40k two-second two-channel noisy reverberated speech samples, of which 38k are used for training and 2k for validation. We conducted experiments by supervising the model with datasets containing 10k to 60k samples. As expected, a steady performance improvement is observed from 10k to 60k, but we notice diminishing returns after reaching around 40k samples.

6.2.4 Model selection and hyperparameters

In search of a state-of-the-art learning-based DOA estimator, the long list of methods given by Grumiaux et al. (2022) is examined. The investigation is carried out with the objective to find an open-source method that could perform multisource DOA localization, irrespective of the microphone array configuration (i.e., the inter-microphone distance), and that has been tested on real test datasets.

Based on the mentioned criteria we opted for the model presented by He et al. (2019) and subsequently updated by He et al. (2021). We employed the DNN architecture from the latter study. This architecture involves a multi-task DNN that performs multi-source DOA estimation, speech detection and counting (not addressed in this study). The authors showed the ability to train an effectively deeper CNN thanks to residual blocks (Kaiming et al., 2016) that helped to achieve robustness against noisy signals and outperformed the MUSIC method. Moreover, their method uses the raw STFT as input features and the architecture is invariant towards any array configuration. Given the availability of the code on Github, this method was a clear selection.

The model is trained over different simulated training sets as described in the previous section. The training is performed with the ADAM optimizer and a learning rate of 10^{-4} over a batch of size 16 for a maximum of 110 epochs, with early stopping on validation sets. In this study, we employed the same input features as described by He et al. (2021), which involved concatenating STFT coefficients for each microphone. The STFT coefficients are calculated with 50% overlap and 42.7 ms time windows. For consistency, all the signals used in our study were down-sampled to 16 kHz, although He et al. (2021) used signals sampled at 48 kHz.

6.2.5 Experiments and results

6.2.5.1 Baseline system and evaluation metric

The trained virtually-supervised DOA estimation models are evaluated using two different metrics, namely, the mean angular error (MAE, in degrees) and Recall (in %). The recall is defined as the ratio of localized sources with an error below 10°. We tried the

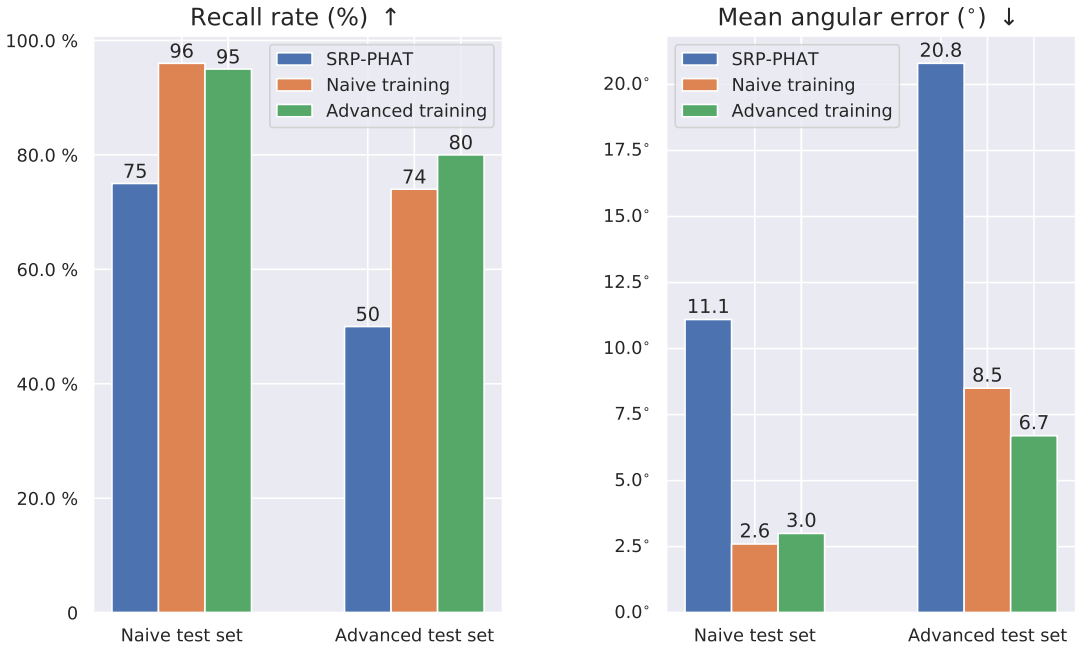

Figure 6.2: Localization results on naive and advanced simulated test sets following the VoiceHome-2 scenario.

recall metric with three error threshold: 5°, 10°, and 20°. The 10° threshold showed to be adequate to prune out the outliers. All three models trained on naive and advanced simulated data are compared to the classical learning-free SRP-PHAT localization method implemented by Scheibler et al. (2018).

6.2.5.2 Simulated test data

In this test, we compare three methods, namely, the SRP-PHAT and the model of He et al. (2021) trained using the naive and advanced training strategies described in the previous section, under the simulated VoiceHome-2 scenario. The results are shown in Figure 6.2. Both bar plots reveal that all the methods perform well on the naive test set, with both trained models achieving near-perfect recall rates. This implies that even under the same noise and reverberation-time distributions, the localization task is made much harder by the presence of realistic wall, source, and receiver responses. Although there is no prior research that directly supports this, we think that this could offer a useful guideline to enhance the evaluation of localization techniques on synthetic datasets. The bar plots on the advanced training set gave results as expected. The learning-based methods demonstrate superior performance compared to the learning-free approach, and the advanced model exhibits better generalization capabilities to advanced conditions compared to the model trained on the naive dataset. Moreover, we also observe that the advanced model performs

Real Test Sets →	VoiceHome-2		DIRHA		STARS22	
Methods	↑ Recall	↓ MAE (°)	↑ Recall	↓ MAE (°)	↑ Recall	↓ MAE (°)
SRP-PHAT	70%	9.9 ± 1.5	61%	15.0 ± 2.3	45%	14.9 ± 0.6
Naive training	78%	7.6 ± 1.2	77%	8.4 ± 1.4	57%	12.9 ± 0.6
Advanced training	**85%**	**5.8 ± 0.8**	**84%**	**6.3 ± 1.0**	61%	**11.4 ± 0.5**
Ablation study						
w/o wall realism	83%	6.2 ± 0.8	81%	7.5 ± 1.4	59%	12.1 ± 0.6
w/o source realism	82%	7.1 ± 1.1	80%	7.8 ± 1.2	**63%**	**11.4 ± 0.6**
w/o receiver realism	N/A	N/A	78%	8.3 ± 1.5	53%	13.4 ± 0.6

Table 6.3: Localization results on three real test sets achieved by the SRP-PHAT baseline and by the supervised model of He et al. (2021) trained using various simulation modes. All three real tests are recorded with different microphone arrays. Mean angular errors (MAE) are displayed with their 95% confidence interval. Bold numbers indicate the best system in each column and the systems statistically equivalent to it. Statistical significance was assessed using McNemar's test for the Recall metric and 95% confidence intervals over angular error differences for the MAE metric.

nearly as well as the naive model in the naive condition, despite the mismatch in the training conditions. This further supports the notion that speaker localization becomes inherently more challenging in more realistic conditions.

6.2.5.3 Real test data

The trained models and the baseline SRP-PHAT are compared on the three real datasets. The top part of Table 6.3 depicts the results. It is observed that the advanced training approach consistently outperforms the naive training approach, achieving 4 to 7 recall points higher and a 2° MAE margins across all three datasets, despite utilizing the exact same network architecture. The SRP-PHAT baseline is constantly outperformed by the advanced training approach by a margin of 15 to 23 recall points and 3° to 9° mean angular error. The results reveal that performing DOA estimation on the STARSS22 dataset is challenging, which aligns with the description of the dataset provided in Section 6.2.2. It is worth noting that even in the quiet and static conditions of VoiceHome-2 and DIRHA datasets, the results from the strong SRP-PHAT baseline are far from perfect. This clearly indicates that two-channel DOA estimation remains a challenging task in real-world settings.

Additionally, an ablation study on the proposed advanced simulation strategy is also presented in the lower half of Table 6.3. A general consensus drawn from these results is that the removal of any of the three layers of realism leads to a noticeable decline in performance. One notable exception occurs on the STARSS22 dataset, where the use of measured directive sources at training time does not seem to improve the performance. One possible explanation for this observation is that the human speakers in the STARS22 dataset exhibit significant head rotations, which are not accounted for in our framework. In contrast, the

results for this dataset are strongly improved by using the measured array directivity at training time, an observation which we have not encountered in previous literature for a two-microphone array of this nature. The positive impact of source directivity can be seen in the other two datasets which confirms the findings of Gelderblom et al. (2021). Using realistic wall absorptions appear to provide a similar enhancement in performance. To the best of our knowledge, this is a novel finding, which could be attributed to the presence of diverse real-world rooms in these datasets.

6.3 Summary

In this chapter, we used our advanced Pyroomacoustics simulator to simulate realistic RIRs for training sets and assess its impact on virtually supervised learning systems. We carefully crafted simulated training sets by incorporating various layers of realism in source, receiver, and wall responses to evaluate the generalization of the models for the tasks of room parameter estimation and speaker localization. The existing literature on learning-based systems for both tasks largely overlooks these aspects. However, our results on both tasks demonstrate that each added layer of realism significantly improves the estimation performance of the DNN models on real test sets. Notably, incorporating source directivity and reflectivity-biased wall sampling reduces errors on both tasks compared to the models trained on naive training datasets. This showcases that by employing an extension of ISM simulation, we can enhance the performance of virtually supervised systems without incurring additional computational costs.

7 Conclusion and perspectives

This thesis has presented contributions to DNN-based room parameter estimation and speaker localization, with an emphasis on virtually supervised learning, wherein the network is trained using simulated data. In this chapter, we wrap up the thesis by providing a summary of our contributions in Section 7.1 and outlining perspectives and potential future directions in Section 7.2.

7.1 Conclusion

In Chapter 4 we addressed the problem of room parameter estimation. Noting the dependence between the different acoustic parameters present in Sabine's law, we jointly estimate the room's surface area, volume, reverberation time, and mean absorption coefficients in six octave bands from noisy speech signals. We presented a novel multi-task DNN architecture inspired from Luo and Mesgarani (2019) that efficiently uses the data captured by a multi-microphone setup. Also, we developed a technique to fuse multiple observations (source-receiver pairs) in the room. The DNN is trained from the data simulated by a hybrid geometric acoustics simulator which comes under the strategy of virtually supervised learning. The results of this study highlight that the incorporation of interchannel cues significantly enhances the blind estimation of room volume and surface from noisy speech. However, for the estimation of reverberation and absorption parameters, a single channel suffices. Notably, the accuracy of absorption coefficient estimation is limited due to challenges in annotating real data and the constraints of the absorption models for walls used in the ISM. Additionally, the study underscores that combining multiple measurements leads to reduced estimation errors and variances across all parameters. The outcomes also show that a system trained on a meticulously simulated training set provides satisfactory generalization capabilities when applied to real-world data.

In Chapter 5 our focus centered on the enhancement of realism in the open-source Pyroomacoustics simulator (Scheibler et al., 2018). Despite the availability of numerous room acoustic simulators capable of generating various degrees of realism in room impulse responses (RIRs), most of them lack open-source accessibility and aren't programmed in Python (see Section 3.4). The Pyroomacoustics library satisfies both criteria. We expanded the simulator's capabilities by integrating an extended version of the ISM. Specifically, our attention was on incorporating measured source and receiver directivities and enhancing the realism of wall responses. This update fundamentally transformed the simulator's core

operation from time-domain to frequency-domain RIR computation. The incorporation of measured directivity was achieved through the utilization of the DSHT and interpolation. As demonstrated in Section 5.4, our implemented approach produces room impulse responses with a higher level of realism without significantly increasing computational costs. Also, this work is an evident step to improve the generalization of DNN-based method trained under the strategy of virtually supervised learning.

In Chapter 6 we took forward our advanced version of the Pyroomacoustics simulator and use it to simulate training data in order to train two DNN models. One of the tasks was room parameter estimation, as in Chapter 4, while the other task was speaker localization. The aim was to study the system's generalization ability to real-world datasets when trained with RIRs simulated using extended ISM. We created multiple training datasets with varying levels of realism in source, receiver, and wall responses to perform an ablation study for both tasks. Informed decisions were made for the inclusion of the measured source and receiver directivity with respect to each test dataset. Our results on both tasks indicate that each added layer of realism improves the performance of the DNN models on real test sets. Significantly, the inclusion of source directivity and the implementation of reflectivity-biased wall sampling (Foy et al., 2021) lead to lower errors on both tasks when compared to models trained on naive training datasets. This serves as a clear demonstration that through the utilization of an extended ISM simulation, we can augment the performance of virtually supervised systems without incurring any additional computational overhead.

7.2 Perspectives

The research conducted in this thesis paves the way for exploring various intriguing avenues for further investigation. Several of these potential directions are outlined below.

7.2.1 Advanced Pyroomacoustics simulator

In Chapter 5, we describe our implementation of extended ISM in the open-source Pyroomacoustics simulator (Scheibler et al., 2018). Moving further in the direction of improvements there is a possibility of data augmentation for measured source and microphone responses. Implementing it will lead to more diversity in directivity patterns which can directly effect the generalization of virtually supervised audio systems.

In our implementation of the extended ISM the late reverberation of the RIR is modeled using the ISM, however, more realism in the RIR can be achieved if it is modeled using stochastic ray tracing method (Gelderblom et al., 2021). Combining source and microphone directivity with the ray tracing methods is a challenging task, its implementation is shown in the simulator of Schimmel et al. (2009) using angle-time-frequency histograms.

Another avenue for enhancement involves introducing complex-valued wall impedance.

A study conducted by Meissner and Zielinski (2022) reveals that there are perceptual distinctions in reverberation between walls with real-valued impedance and those with complex-valued impedance, particularly in smaller rooms. Along the same lines, inclusion of diffraction effects in geometeric acoustics simulators will enhance the realism of the simulated RIR. In this thesis we worked with regular shoebox rooms because of the limitation of the simulators. Wider application opportunities can be observed if these techniques are extended to non-shoebox rooms. Keeping the computational time low while employing all these techniques would be a challenge, but achieving this would give this simulator an edge over wave-based simulation methods (Hamilton, 2021), which are capable of simulating RIRs with outmost realism but due to their high computational time can hardly be used to train virtually supervised learning methods.

7.2.2 Room parameter estimation

A DNN-based joint room parameter estimation model is presented in Chapter 4. The same model is employed in Chapter 6 to investigate the influence of enhancing realism in the training data through an ablation study. In terms of problem statement and the design of the pipeline, further exploration can be done on many fronts. This could include an extension of the current room parameter estimation model towards the joint estimation of local parameters such as position and properties of the source and the individual surfaces. This can be accompanied with experimentation on different DNN architecture designs and efficient input features used with specific microphone arrays such as circular arrays or an Eigenmike. A viable starting point is the recent study by Ick et al. (2023) that points towards the use of Gammatone spectograms combined with Gammatone phase spectograms for joint estimation of reverberation time and volume. Furthermore, data augmentation techniques on real data could be considered while training the system, which could be compared with the model trained purely with a virtually supervised approach. To realize this, efforts should be made in the measurement of a new dataset designed particularly for room parameter estimation involving different acoustic conditions with proper annotation of global and local acoustic parameters. The need for such datasets has been shown in various studies on room parameter estimation (Ick et al., 2023; Genovese et al., 2019; Xiong et al., 2018). Based on the results shown in Chapter 4 and 6, both contributions lack a detailed study of training the systems with data comprising different noise distributions and its impact on the real test sets. Also, the systems can be tested on a variety of real-world test sets to further bolster the claim of generalization. At last, conducting a thorough investigation employing a range of measured microphone responses is essential to substantiate comparable assertions to those made regarding source directivity, which has indeed demonstrated its utility in estimating room geometrical parameters.

A focus could also be given to the real-time estimation of the acoustic parameters in dynamic conditions, which could further be employed in different mobile devices or audio processing frameworks making it more accessible and usable in real-world scenarios. This

could integrate with various audio applications such as noise reduction, audio enhancement, virtual reality, and augmented reality, to enhance the user experience.

7.2.3 Sound source localization

Our contribution to speaker localization, as discussed in Chapter 6, aims to explore simulation techniques and their influence on the generalization of a system trained through virtually supervised methods. The outcomes of this study reveal certain findings that warrant further investigation through additional experiments. We lack a detailed study investigating the impact of various noise distributions within the training sets and their correlation with performance on the real test sets. During the training phase, we noticed that the model trained with advanced training sets attains convergence earlier compared to the one trained on naive datasets. However, a more thorough exploration is necessary to establish a conclusive claim in this direction. Additionaly, similar to other contributions, the impact of introducing more measured directivity patterns for microphones in the training of the system is an important research direction. Furthermore, this study's scope can be extended to more scenarios, such as localizing sound sources in dynamic acoustic environments with multi-source DOA estimation. Another promising avenue of experimentation involves the localization of sources in both 2D and 3D coordinates.

8 Résumé étendu

La Réalité Augmentée (AR) vise à intégrer du contenu virtuel généré par ordinateur dans l'environnement physique de manière à créer une fusion transparente, rendant le contenu virtuel semblable à une partie du monde réel (Azuma, 1997). L'AR a le potentiel d'améliorer la perception et l'interaction des personnes avec leur environnement, facilitant ainsi l'exécution de tâches du monde réel. Un avantage significatif de la technologie AR est sa capacité à augmenter les sens humains, permettant aux individus d'interagir avec des objets et des scènes virtuels aussi facilement qu'avec le monde physique (Azuma et al., 2001). Cependant, la grande majorité de la recherche en AR s'est principalement concentrée sur l'augmentation visuelle (Kim et al., 2018). La Réalité Augmentée Audio (AAR) a reçu moins d'attention par rapport à l'augmentation visuelle. En AAR, du contenu auditif virtuel est intégré de manière transparente dans l'environnement acoustique réel, améliorant ainsi l'expérience utilisateur. Afin de créer une perception réaliste de la direction, de la distance et de la réverbération pour les sons virtuels, les sources sonores virtuelles sont souvent spatialisées de manière binaurale. La technologie AAR est capable de créer des expériences audio immersives pour divers types de contenus sonores tels que la parole, la musique, les signaux et les alertes (Li et al., 2018). Par conséquent, elle peut transmettre différents types d'informations en fonction du contexte.

Des dispositifs informatiques puissants facilement disponibles tels que les téléphones mobiles ont la capacité de simuler des sources sonores virtuelles et de fournir une expérience audio immersive. L'utilisateur peut accéder à ces expériences audio immersives avec de nombreux écouteurs disponibles dans le commerce (Yang, 2021). Récemment, de nombreux casques et écouteurs AR-VR ont été introduits sur le marché, capables d'effectuer des calculs AAR en temps réel. Parmi eux figurent l'Apple Vision Pro, l'Apple AirPod Pro, le Samsung Galaxy Buds Pro et le Sony WH-1000XM5, ce qui a suscité un intérêt croissant dans ce domaine. Comparativement à l'augmentation audio dans les systèmes de réalité virtuelle (VR), les systèmes AAR sont plus difficiles à mettre en œuvre en raison de la technologie complexe nécessaire pour augmenter l'audio dans des conditions du monde réel. Plus spécifiquement, en VR audio, l'utilisation de scènes virtuelles préconçues peut simplifier le rendu de contenu audio, tandis qu'en AAR, la création de sons virtuels dans le monde physique et leur adaptation en temps réel à l'utilisateur est plus complexe. Un système AAR fonctionnel comprend cinq composants technologiques : le suivi de la position de l'objet de l'utilisateur, la modélisation acoustique de la salle, la synthèse sonore spatiale, l'interaction et la technologie d'affichage (Yang et al., 2022). La technologie d'interaction est nécessaire pour obtenir les entrées de l'utilisateur afin d'ajuster les paramètres de l'AAR, tandis que

la technologie d'affichage se réfère à la manière dont le son est restitué à l'utilisateur. Pour afficher le contenu visuel accompagné d'un son immersif, un écran est également nécessaire. Dans cette thèse, nous nous concentrons uniquement sur un aspect de l'AAR : la modélisation acoustique de la salle.

L'objectif de cette thèse est d'estimer les paramètres acoustiques globaux et locaux d'une pièce à partir de sources sonores inconnues. Les paramètres acoustiques globaux comprennent la surface de la salle S, le volume V, les coefficients d'absorption moyens $\overline{\alpha}$ et le temps de réverbération RT_{60}. En ce qui concerne les paramètres locaux, nous visons uniquement à estimer l'emplacement de la source sonore. Nous contribuons donc à deux tâches distinctes : l'estimation des paramètres de la salle et la localisation de la source sonore. Nous abordons ces deux tâches à l'aide d'une approche supervisée basée sur les réseaux de neurones profonds (DNN). Ces systèmes DNN sont entraînés exclusivement sur des données simulées, une approche appelée *apprentissage virtuellement supervisé*. La thèse se focalise notamment sur l'impact du réalisme du simulateur employé dans la généralisabilité des méthodes.

Ce résumé étendu est organisé comme suit. Dans la Section 8.1, nous développons notre première contribution, un cadre multi-tâche multicanal pour l'estimation des paramètres acoustiques de la salle. Dans la Section 8.2, nous décrivons notre contribution à l'amélioration du réalisme des RIR simulées par le simulateur acoustique de salle open source Pyroomacoustic en implémentant une version étendue de la méthode des sources images (ISM) avec des directivités de source et de récepteur dépendantes de la fréquence. Dans la Section 8.3, l'impact de la version modifiée du simulateur pyroomacoustics est analysé sur la tâche de localisation de source sonore et d'estimation des paramètres de la salle.

8.1 Estimation des paramètres de la pièce à canaux multiples en utilisant de multiples points de vue

Sélectionner les paramètres acoustiques pertinents de l'environnement utilisateur représente un défi. Le concept d'empreinte de réverbération proposé par Jot et Lee (2016) caractérise une salle en fonction de son volume et de son temps de réverbération par bande d'octave. Ces deux paramètres sont liés entre eux par la loi de Sabine dans des conditions idéales de champ sonore diffus, qui implique également la surface de la pièce et les coefficients d'absorption des parois. Cela nous a poussés à effectuer une estimation conjointe de ces quatre paramètres. Les systèmes proposés dans la littérature estiment généralement chacun de ces paramètres séparément et fonctionnent sur des caractéristiques à canal unique et estiment des valeurs à large bande pour les paramètres dépendants de la fréquence (Murgai et al., 2017; Kataria et al., 2017; Genovese et al., 2019).

Comme indiqué dans le chapitre 4, nous estimons simultanément les quatre paramètres

mentionnés dans six bandes d'octave. Pour ce faire, nous proposons une nouvelle architecture DNN multi-tâches capable de traiter des caractéristiques à canal unique et multi-canal. Ce réseau est entraîné sur des données simulées avec une fonction de perte basée sur la vraisemblance maximale qui produit des variances adaptatives pour chaque paramètre. Cela nous permet de fusionner de manière statistiquement motivée de multiples observations indépendantes d'une salle. Les résultats sont évalués sur des enregistrements de parole réverbérée simulée et réelle.

8.1.1 Les données d'entraînement

Nous décrivons d'abord comment des RIR variées et réalistes ont été générées. Dans cette section, nous utilisons un simulateur acoustique de salle appelé Roomsim,développé par Schimmel et al. (2009). Roomsim est un simulateur d'acoustique géométrique hybride qui utilise des méthodes de traçage de rayons ISM et stochastiques pour simuler la réflexion spéculaire et diffuse. Le simulateur Roomsim utilise la technique de la *pluie diffuse* pour la réalisation du calcul de traçage stochastique des rayons Schröder (2011). La simulation est effectuée à une fréquence d'échantillonnage de 48 kHz, un ordre de source image de 10 et $2,000$ rayons pour les calculs stochastiques. Des simulations ont été effectuées pour $20,000$ différentes de salles "boîtes à chaussures" (shoebox) dont la longueur, la largeur et la hauteur sont uniformément échantillonnées à partir d'une gamme de $[3, 10]$ m, $[3, 10]$ m et $[2.5, 6]$ m. Il en est résulté $S \in [48, 440]$ m² et $V \in [18, 600]$ m³. Cinq positions aléatoires de sources et de microphones ont été simulées dans chaque pièce (c'est-à-dire, 5 points de vue). La source et le réseau de microphones ont été simulés comme étant omnidirectionnels et sont maintenus à au moins 30 cm l'un de l'autre et de chaque mur de la salle afin d'éviter les artefacts de champ proche. Le réseau de microphones considéré dans notre étude consiste en deux microphones situés à une distance de 22.5 cm. La distance ressemble beaucoup à la largeur d'une tête ou d'un casque. Pour obtenir du réalisme dans la scène acoustique de la salle échantillonnée, les parois des salles rectangulaires doivent avoir des coefficients d'absorption dépendant de la fréquence $\alpha_v(f)$ avec $v \in [1, \ldots, 6]$. Pour incorporer des coefficients dépendants de la fréquence, une représentation générale dans le domaine de Fourier de l'Équation (2.29) s'écrit comme suit :

$$H(\mathbf{r}_m^{\text{mic}}, f) = \sum_{k=0}^{K} \frac{D_{k,m}(f)}{4\pi \|\mathbf{r}_m^{\text{mic}} - \mathbf{r}_k^{\text{src}}\|_2} e^{-i2\pi f \frac{\|\mathbf{r}_m^{\text{mic}} - \mathbf{r}_k^{\text{src}}\|_2}{c}}, \tag{8.1}$$

où $D_{k,m}(f)$ est l'atténuation totale des parois de la $k-$ ieme source image définie dans le domaine fréquentiel. Les propriétés d'absorption du matériau utilisé dans les bâtiments sont généralement fournies par les fabricants sous forme de coefficients d'absorption donnés en valeurs $[0, 1]$ par bande d'octave. Par conséquent, il est nécessaire d'interpoler ces coefficients dans le domaine de Fourier et d'ajouter une réponse en phase aux parois.

Pour générer notre ensemble d'entraînement, les coefficients d'absorption pour six surfaces dans six bandes d'octaves sont échantillonnés à l'aide de la stratégie d'échantillonnage

	Murs	Planchers	Cloisons
125 Hz	[0.01 - 0.50]	[0.01 - 0.20]	[0.01 - 0.70]
250 Hz	[0.01 - 0.50]	[0.01 - 0.30]	[0.15 - 1.00]
500 Hz	[0.01 - 0.30]	[0.05 - 0.50]	[0.40 - 1.00]
1000 Hz	[0.01 - 0.12]	[0.15 - 0.60]	[0.40 - 1.00]
2000 Hz	[0.01 - 0.12]	[0.25 - 0.75]	[0.40 - 1.00]
4000 Hz	[0.01 - 0.12]	[0.30 - 0.80]	[0.30 - 1.00]

Table 8.1: Plages de $\alpha_v(b)$ utilisées pour la stratégie *reflectivity-biased* (Foy et al., 2021).

reflectivity-biased, telle que décrite dans Foy et al. (2021). Cette stratégie attribue à chaque paroi 50% chances d'être soit une paroi réfléchissante indépendante de la fréquence, soit une paroi absorbante dépendante de la fréquence, où les valeurs du profil de la paroi réfléchissante sont tirées uniformément au hasard dans une plage $\alpha_v(b) \in [0, 0.12]$, tandis que les valeurs du profil de la paroi absorbante dépendent du type de surface et sont tirées uniformément au hasard dans les fourchettes indiquées dans le tableau 8.1. La stratégie d'échantillonnage avancée conduit à une distribution réaliste de $\mathrm{RT}_{60}(b) \in [0.2, 3.2]$ s et $\overline{\alpha}(b) \in [0.02, 0.6]$ dans l'ensemble de données simulées.

Les signaux de réverbération à deux canaux résultants, d'une durée de 3 secondes chacun, subissent deux types de corruption par le bruit. Premièrement, il y a le bruit statique du microphone, qui comprend du bruit gaussien additif blanc indépendant sur chaque canal. Deuxièmement, il y a le bruit spatialement diffus, qui est constitué de *speech-shaped noise* convolu avec la partie tardive (> 50 ms) d'une Réponse Impulsionnelle de Salle (RIR) aléatoire dans la salle. Pour garantir des niveaux de bruit réalistes, les signaux provenant de sources situées plus loin du récepteur présentent des Rapports Signal-Bruit (SNR) plus faibles. Pour ce faire, pour chaque pièce, nous créons d'abord un signal de référence à l'aide d'une source de parole aléatoire située à 1 mètre devant un récepteur (ce signal n'est pas utilisé dans le jeu de données final). Les niveaux de bruit statique et diffus pour ce signal de référence sont configurés pour obtenir des SNR uniformément répartis dans la plage de [70, 90] dB et [30, 60] dB, respectivement. Ces niveaux de bruit restent constants pour les mélanges finaux, quelle que soit la distance entre la source de parole et le récepteur. Cette approche résulte en une plage SNR globale de [-10, 65] dB sur l'ensemble du jeu de données

8.1.2 Modèle de réseau de neurones

Nous proposons une nouvelle architecture de RN profond illustrée dans la Figure 8.1. Le composant de base de cette architecture utilise des blocs de convolution 1D, inspirés de l'architecture Conv-Tas Net utilisée pour la séparation de la parole dans Luo and Mesgarani (2019). Chaque bloc de convolution 1D est composé de différentes parties, à savoir une couche de convolution séparable suivie d'une couche ReLU et d'une normalisation de

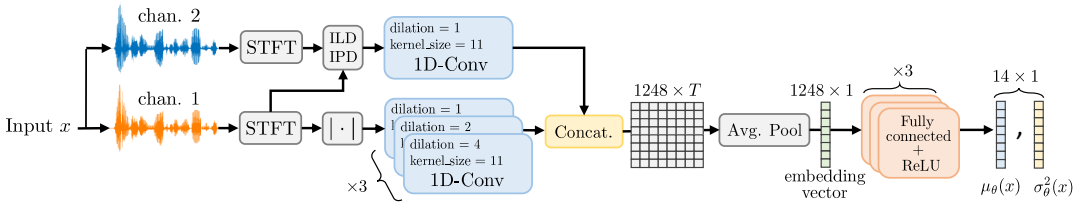

Figure 8.1: Architecture du réseau de neurones proposé

couche (Ba et al., 2016). Le réseau est divisé en 2 pipelines/parties qui traitent les spectres de magnitude, la partie inférieure comporte trois blocs de convolution 1D avec un taux de dilatation progressif de $[1, 2, 4]$ le long de l'axe de fréquence et une taille de noyau constante de 11. Les deux parties sont destinées à traiter différents ensembles de caractéristiques. La partie inférieure est destinée aux caractéristiques mono-canal (SC) et l'autre aux caractéristiques inter-canaux (IC). Les caractéristiques affinées obtenues à partir des deux pipelines sont concaténées le long de l'axe de fréquence et moyennement regroupées dans le temps pour former un vecteur d'*embedding* de taille (1248×1). Il est suivi de 3 couches entièrement connectées de dimensions respectives $96, 48, 28$, ce qui donne 28 sorties.

8.1.3 Caractéristiques d'entrée

Nous choisissons un signal d'entrée multicanal de 3 secondes $\mathbf{c}$ à partir d'une seule observation, même si notre architecture ne varie pas en fonction de la longueur du signal d'entrée. La STFT est réalisée avec une fenêtre de Hann de 96 ms et 50% de chevauchement, ce qui donne un spectrogramme $C_m(f,t)_{f,t=1}^{F,T}$ avec F =769 fréquences positives et T =63 périodes de temps sur chaque canal m. Ces données sont ensuite utilisées pour calculer les caractéristiques spécialisées monocanales et inter-canales. Quatre caractéristiques monocanal différentes ont été prises en compte : $|C_1(f,t)|^2$, $\sqrt{|C_1(f,t)|}$ et $\log|C_1(f,t)|$ et $|C_1(f,t)|$. La dernière s'est montrée plus efficace plus efficace en tant que caractéristique monocanale dans nos expériences.Les caractéristiques inter-canales sont obtenues en concaténant les différences de niveau interaurales (ILD) et le sinus des différences de phase interaurales (IPD) :

$$\mathrm{ILD}(f,t) = \log|C_1(f,t)| - \log|C_2(f,t)|\,, \tag{8.2}$$

$$\mathrm{IPD}(f,t) = \left[\mathrm{Re}, \mathrm{Im}\left(\frac{C_1(f,t)C_2^*(f,t)}{|C_1(f,t)C_2^*(f,t)|}\right)\right]. \tag{8.3}$$

8.1.4 Fonction de perte

Les Q paramètres des sorties sont modélisés comme une distribution gaussienne, le modèle estime donc 28 valeurs dont 14 sont la moyenne $\mu_\Theta(\mathbf{c})$, c'est-à-dire, les estima-

tions du régresseur. Le reste est $\sigma^2_\Theta(\mathbf{c})$, c'est-à-dire les incertitudes des paramètres estimés. Pour optimiser σ, une log-vraisemblance négative gaussienne est utilisée comme fonction de perte pour optimiser les paramètres du réseau Θ, c'est-à-dire,

$$\mathcal{L}_\Theta(\mathbf{c},\mathbf{y}) = -\log\mathcal{N}(\mathbf{y};\mu_\Theta(\mathbf{c}),\sigma^2_\Theta(\mathbf{c})) = \frac{1}{2}\sum_{q=1}^{Q}\log\sigma^2_{q,\Theta}(\mathbf{c}) + \frac{(y_q-\mu_{q,\Theta}(\mathbf{c}))^2}{\sigma^2_{q,\Theta}(\mathbf{c})}\,, \tag{8.4}$$

où la labélisation est représentée par $\mathbf{y}\in\mathbb{R}^Q$.

8.1.5 Fusion des estimations

Cette approche présente deux avantages. Premièrement, elle permet de pondérer de manière adaptative les erreurs sur les paramètres individuels. Deuxièmement, l'ensemble de données comprend plusieurs observations dans la même salle, et le réseau produit plusieurs estimations basées sur ces observations. Ces estimations peuvent être fusionnées en utilisant la variance que le réseau produit pour tous les paramètres, avec cette formule dérivée du théorème de Bayes,

$$p_\Theta(y_q|\mathbf{x}) = \mathcal{N}(y_q;\overline{\mu}_{q,\Theta}(\mathbf{x}),1/\overline{\gamma}^2_{q,\Theta}(\mathbf{x}))\,, \tag{8.5}$$

dans lequel $\overline{\mu}_{q,\Theta}(\mathbf{x})$ est l'estimation fusionnée dont la formule est donnée par :

$$\overline{\mu}_{q,\Theta}(\mathbf{x}) = \sum_{d=1}^{D}\frac{\gamma^2_{q,\Theta}(\mathbf{c}_d)}{\overline{\gamma}^2_{q,\Theta}(\mathbf{x})}\mu_{q,\Theta}(\mathbf{c}_d)\,, \tag{8.6}$$

où $\overline{\gamma}^2_{q,\Theta}(\mathbf{x}) = \sum_{d=1}^{D}\gamma^2_{q,\Theta}(\mathbf{c}_d)$, et $\gamma^2_{q,\Theta}(\mathbf{c}_d)$ est la précision estimée qui est l'inverse de la variance estimée $\gamma^2_{q,\Theta}(\mathbf{c}_d) = 1/\sigma^2_{q,\Theta}(\mathbf{c}_d)$.

8.1.6 Hyperparamètres et entraînement

Le réseau a été entraîné pendant 120 époques avec l'optimiseur ADAM (Kingma and Ba, 2014) en utilisant un taux d'apprentissage de 10^{-4}. Pour éviter un surajustement, des couches *dropout* sont utilisées entre chaque bloc convolutif avec une probabilité d'abandon de 0,2, et entre chaque couche entièrement connectée avec un taux d'abandon de 0,4. Le réseau, formé sur un ensemble de 80k exemples binauraux, converge généralement en 100-150 époques, avec une patience de 15 époques sur l'ensemble de validation comme condition préalable à l'arrêt précoce.

Les paramètres estimés par le réseau possèdent différentes échelles. Afin d'éviter les problèmes liés aux différences d'échelle, nous avons normalisé les valeurs de la vérité terrain y au moment d'apprentissage, avec les écarts types calculés sur l'ensemble de l'apprentissage. Au moment de l'inférence, ces valeurs sont réutilisées et multipliées par la sortie du réseau.

8.1.7 Résultats et expérimentations

8.1.7.1 Données simulées

À notre connaissance, aucune méthode existante n'est capable d'estimer la surface d'une salle, le temps de réverbération en fonction de la fréquence et le coefficient d'absorption moyen uniquement sur la base de signaux vocaux bruités. Cependant, il existe une méthode pour l'estimation aveugle du volume d'une salle à un canal et une seule position (Genovese et al., 2019). Nous utilisons ce travail comme référence pour comparer la tâche d'estimation du volume pour différents tests sur des données simulées et réelles. Tout au long des expériences, l'erreur absolue moyenne pour chaque paramètre est utilisée comme mesure.

Après l'entraînement, nous effectuons deux expériences différentes avec un ensemble de tests simulés composé de $2,000$ salle non vues. Le but de notre première expérience est de vérifier si la fusion de plusieurs observations prises dans la même salle en utilisant l'approche proposée dans la Section 8.1.4, améliore en effet les résultats pour les signaux d'entrée à une ou deux voies. C'est pourquoi, pour cette expérience, nous avons comparé deux approches :

1. L'architecture complète décrite dans la Figure 8.1, qui prend en compte l'entrée multicanal et sera dénommée (bicanal).
2. La même architecture est utilisée, mais sans la partie supérieure du réseau consacrée au traitement inter-canal, ce qui permet de traiter uniquement le canal unique, qui sera désigné par l'expression (un canal).

Les résultats présentés dans la Figure 8.2 pour $\overline{\alpha}(b)$ et $\mathrm{RT}_{60}(b)$ sont calculées en moyenne pour toutes les bandes d'octave afin de montrer les erreurs moyennes sur les bandes d'octave $\overline{\alpha}$ et RT_{60}. La figure montre clairement que la fusion de plusieurs observations par salle réduit l'erreur de manière progressive sur tous les paramètres. La fusion de 5 observations d'un signal d'entrée à deux canaux fournit la meilleure performance avec une erreur absolue moyenne de 0.052 pour $\overline{\alpha}$, 0.18 s pour RT_{60}, 42 m^2 sur S et 54 m^3 sur V, tandis que les plages de l'ensemble d'entraînement pour ces quantités sont les suivantes : $[0.02, 0.6]$ pour $\overline{\alpha}$, $[0.2, 3.6]$ s pour RT_{60}, $[48, 360]$ m^2 pour S et $[18, 400]$ m^3 pour V. Les résultats de l'estimation de V révèlent que la méthode proposée est plus performante que la méthode de référence pour les deux approches avec une seule observation. Un canal avec une observation réduit les erreurs de 13%, tandis que le système à deux canaux avec cinq observations réduit les erreurs de 31 % par rapport à la référence. L'approche à deux canaux permet d'obtenir une erreur absolue moyenne significativement plus faible pour l'estimation de S et V par rapport à l'approche à un seul canal. Ce résultat peut être associé aux caractéristiques inter-canales qui capturent efficacement les caractéristiques spatiales de la salle, ce qui facilite l'estimation de S et V qui sont des quantités spatiales dépendant des premières réflexions. Toutefois, en raison d'intervalles de confiance se chevauchant, l'utilisation de deux canaux au lieu d'un seul ne semble pas faciliter l'estimation des $\overline{\alpha}$ et RT_{60}.

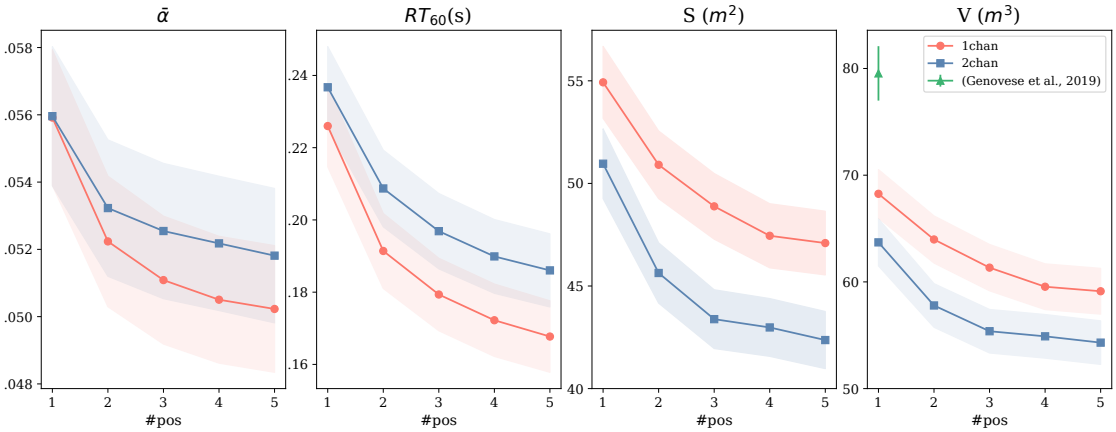

Figure 8.2: Erreur absolue moyenne obtenue sur des données simulées par Genovese et al. (2019) contre le modèle proposé avec des entrées à un ou deux canaux, en fonction du nombre de positions source-récepteur fusionnées dans chaque salle. Les intervalles de confiance à 95%.

Méthode	Caractéristiques	# pos	$\bar{\alpha}$	RT_{60}	S	V
Genovese et al. (2019)	SC	1	-	-	-	137.8
Notre	SC	1	**0.061**	0.134	129.6	154.5
Notre	SC	5	**0.060**	0.097	125.8	149.1
Notre	SC+IC	1	0.084	0.101	89.4	107.6
Notre	SC+IC	5	0.094	**0.062**	**50.2**	**68.8**

Table 8.2: Erreur absolue moyenne obtenue sur 3 salles à partir de l'ensemble de données réelles dEchorate. Les chiffres en gras indiquent le meilleur résultat statistiquement significatif par colonne, sur la base d'intervalles de confiance à 95%..

8.1.7.2 Données réelles

Pour tester la capacité de généralisation du modèle entraîné dans des conditions acoustiques réelles inédites, des enregistrements de parole réverbérée de l'ensemble de données dEchorate (Di Carlo et al., 2021) ont été utilisés comme un test réel. Sur les 11 configurations de salles dans l'ensemble de données dEchorate, nous choisissons 3 configurations de salles qui ont 3 surfaces réfléchissantes ou plus avec $\overline{\alpha}(b) \in [0.16 - 0.35]$ et $\mathrm{RT}_{60}(b) \in [0.25{-}0.66]$ sec. Ces salles ressemblent à des salles acoustiques réelles et relèvent du scénario choisi pour l'ensemble d'entraînement. Nous avons choisi une combinaison de 30 paires de microphones pour chacune des trois salles, ce qui donne $3 \times 30 = 90$ signaux de parole de 3 secondes chacun.

Les résultats sur l'ensemble de tests réels sont présentés dans le Tableau 8.2. Les erreurs absolues moyennes sont reportées dans le tableau en utilisant la méthode proposée pour les approches à un canal et à deux canaux. Elle fournit également une comparaison avec

la méthode de base sur l'estimation de V. Ils ont été calculés sur 1 ou 5 observations avec des antennes de microphones fixes et des positions de source mobiles qui ont été choisies de manière aléatoire parmi les 6 sources présentes dans la pièce. Un point important que l'on peut observer dans le tableau est que les erreurs générées par notre approche se situent dans la même fourchette que celles obtenues sur les données simulées. De plus, les erreurs pour l'estimation de V en utilisant l'approche à un canal avec une observation sont comparables à la méthode de référence. Les résultats obtenus en utilisant l'approche à deux canaux avec des caractéristiques inter-canaux renforcent également l'hypothèse selon laquelle les caractéristiques inter-canales sont importantes pour l'estimation de S et V car leur inclusion diminue significativement les erreurs. De plus, les chiffres montrent à nouveau que l'augmentation du nombre d'observations réduit les erreurs sur RT_{60}, S et V. Cependant, les erreurs obtenues sur $\overline{\alpha}$ ne présentent pas de cohérence et sont plus variées que celles obtenues sur l'ensemble de données simulées. Cela pourrait s'expliquer par deux raisons : Tout d'abord, la difficulté inhérente à l'annotation des coefficients d'absorption dans une pièce réelle. Deuxièmement, la modélisation des coefficients d'absorption dans la méthode ISM utilisée pour former le modèle est moins valide aux basses fréquences.

8.1.8 Résumé

Les résultats de cette étude démontrent que l'incorporation de signaux intercanaux peut grandement améliorer l'estimation aveugle du volume et de la surface d'une pièce à partir de parole bruitée. Cependant, lors de l'estimation des paramètres de réverbération et d'absorption, un seul canal s'avère suffisant. Il convient de noter que l'estimation des coefficients d'absorption n'est pas très fiable, en raison des défis liés à l'annotation de données réelles et des limitations des modèles d'absorption pour les parois utilisés dans la méthode ISM. De plus, l'étude souligne que la combinaison de plusieurs mesures réduit les erreurs et les variances d'estimation pour tous les paramètres. Les résultats montrent également qu'un système formé sur un ensemble d'entraînement méticuleusement simulé présente des capacités de généralisation satisfaisantes lorsqu'il est appliqué à des données du monde réel.

8.2 ISM étendue et implémentation sous Pyroomacoustics

La contribution décrite dans le Chapitre 5 se concentre sur l'amélioration du réalisme du simulateur acoustique de salle open-source Pyroomacoustics en augmentant le simulateur avec la mise en œuvre d'une version plus réaliste de la méthode des sources images (ISM) sans compromettre son temps de calcul. Bien qu'il existe une grande variété de simulateurs acoustiques de salle capables de simuler des RIR à différents degrés de réalisme (voir Section 3.4), la plupart d'entre eux ne sont soit pas open source, soit ne sont pas écrits en Python. La bibliothèque Pyroomacoustics répond à ces deux exigences. Cette contribu-

tion est une étape nécessaire pour améliorer la généralisation des méthodes d'apprentissage virtuellement supervisées basées sur des réseaux de neurones profonds.

8.2.1 ISM étendu

Pour rendre le simulateur et les RIRs plus réalistes dans le but de l'apprentissage virtuellement supervisé, nous proposons d'ajouter la capacité de charger les diagrammes de directivité mesurés des sources et des microphones réels dans le simulateur. Afin de tenir compte de l'ensemble de l'échelle des fréquences, notre contribution au simulateur modifie le modèle de construction de base des RIRs dans Pyroomacoustics, ce qui conduit à la construction des RIRs dans le domaine fréquentiel. La plupart des simulateurs d'acoustique des salles n'implémentent pas l'ISM dans sa forme étendue, bien que cela puisse être fait avec peu de surcharge computationnelle. Pour tenir compte des modèles de directivité des sources et des récepteurs dans la simulation des RIR, nous étendons l'Equation 4.1 comme suit :

$$H(\mathbf{r}_m^{\text{mic}}, f) = \sum_{k=0}^{K} \frac{D_{k,m}(f)}{4\pi \|\mathbf{r}_m^{\text{mic}} - \mathbf{r}_k^{\text{src}}\|_2} e^{-i2\pi f t_{k,m}^{\text{delay}}} D^{\text{air}}(f) G^{\text{src}}(-\theta_{k,m}, -\phi_{k,m}, f) \\ G^{\text{mic}}(\theta_{k,m}, \phi_{k,m}, f), \quad (8.7)$$

où l'atténuation atmosphérique dans le domaine fréquentiel est notée comme $D^{\text{air}}(f)$. Le retard temporel d'arrivée à un microphone m pour une source image k est notée par $t_{k,m}^{\text{delay}}$ Les modèles de directivité des microphones et des sources sont notés par G^{mic} et G^{src}, les angles d'arrivée en azimut et en élévation d'une source d'image k vers un microphone m sont notés par $(\theta_{k.m}, \phi_{k,m})$. L'angle de départ d'une source donnée est l'opposé de l'angle d'arrivée au microphone. Cette équation sera désignée sous le nom de modèle ISM étendu. Une formulation similaire a été donnée par Schröder (2011).

8.2.2 Jeux de données de directivité

Deux jeux de données ont été utilisés dans cette étude. Tout d'abord, il y a le jeu de données DIRPAT qui se compose de modèles de directivité 2D/3D de diverses sources et microphone mesurés à l'Institut de musique électronique et d'acoustique de l'Université de musique et des arts du spectacle de Graz par Brandner et al. (2018). Il comprend des mesures en 3D de quatre types différents de réceivers : le microphone AKG C414 avec quatre configurations de directivité (omnidirectionnelle, cardioïde, supercardioïde et en huit), le microphone AKG C480 et deux microphones Soundfield. La Figure 8.3 montre la carte de chaleur sphérique de l'ensemble des motifs de directivité présentés par le microphone AKG C414 à trois fréquences différentes. Le jeu de données comprend également des motifs mesurés pour douze sources, notamment des haut-parleurs génériques, des amplificateurs de guitare et un HATS capable de simuler la parole humaine. La Figure 8.4 montre la carte

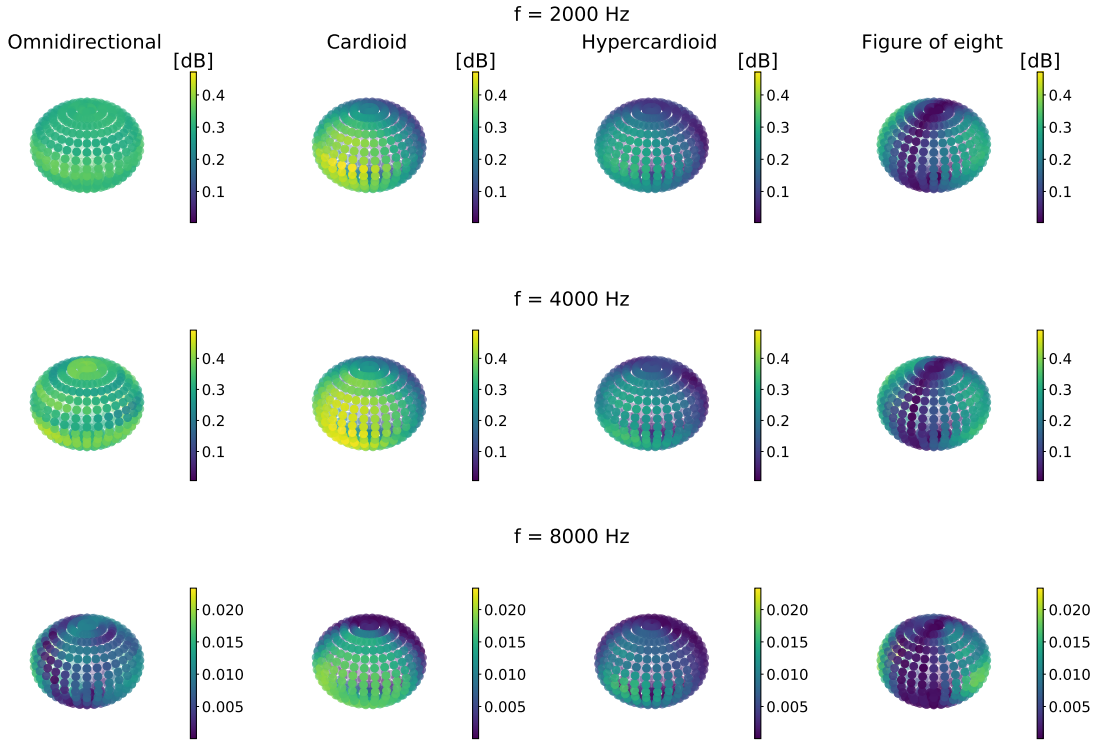

Figure 8.3: Carte sphérique de directivité du microphone AKG C414 issue de l'ensemble de données DIRPAT. La barre de couleur montre la magnitude normalisée des filtres.

de chaleur sphérique de quelques haut-parleurs à $f = 2000$ Hz. Le deuxième jeu de données directivité fournit le motif de directivité du microphone à 32 canaux Eigenmike. La mesure de ce motif de directivité est effectuée sur la même grille sphérique que DIRPAT et au même endroit par Franz Zotter[1].

8.2.3 Construction de RIR dans le domaine des fréquences

Contrairement à l'implémentation originale de Pyroomacoustics, les calculs dans Pyroomacoustics modifié sont effectués entièrement dans le domaine DFT. L'implémentation de la construction de la RIR en domaine fréquentiel modifie considérablement le code et le fonctionnement central de Pyroomacoustics par rapport à l'implémentation précédente de la construction de la RIR en domaine temporel. Les détails techniques sur cette implémentation sous forme de pseudo-code sont présentés dans la Section 5.3.3.

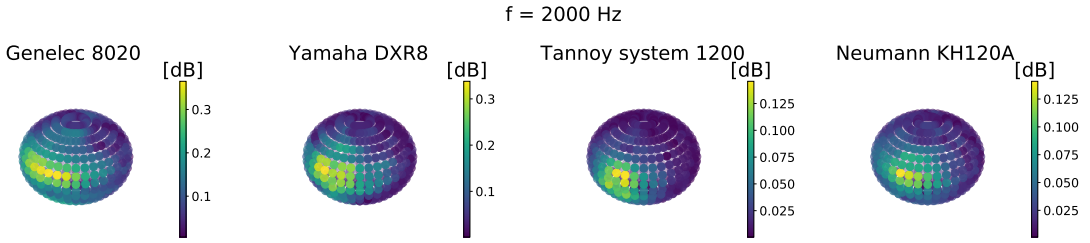

Figure 8.4: Cartes sphériques de directivité de 4 fabricants différents de haut-parleurs issues de l'ensemble de données DIRPAT. La barre de couleur montre la magnitude normalisée des filtres.

8.2.4 Similarité avec les RIR mesurées

Pour valider notre affirmation selon laquelle l'incorporation de modèles de directivité réels améliore le réalisme des RIR, nous avons réalisé des expériences qualitatives. Les RIRs du jeu de données dEchorate sont utilisées pour cette expérience. Ces RIRs sont annotées et peuvent être ainsi simulées dans une scène acoustique virtuelle à l'aide d'un simulateur acoustique de salle. Nous avons reproduit la scène acoustique de la salle "011111" du jeu de données dans notre version modifiée du simulateur. Bien que les diagrammes de directivité des microphones et des sources utilisés pour enregistrer les RIR réelles n'aient pas été fournis dans le jeu de données, le jeu de données a spécifié les références des microphones omnidirectionnels et des haut-parleurs directifs utilisés. En utilisant ces informations, nous avons sélectionné un microphone AKG c414 omnidirectionnel et un haut-parleur Genelec 8020 en tant que substituts dans notre configuration de salle virtuelle. Ces diagrammes de directivité choisis ressemblent étroitement à ceux utilisés dans le jeu de données des RIR mesurées. Avec un peu de post-traitement, nous avons aligné les RIR sur l'axe du temps. La Figure 8.5 montre la RIR réelle du jeu de données et une RIR produite par notre simulateur modifié. En comparant les réflexions directes et de premier ordre, nous constatons que les deux RIR correspondent bien dans le domaine temporel. Dans le domaine de la fréquence en magnitude logarithmique, une forte corrélation peut être observée sur l'ensemble de l'échelle de fréquence.

8.2.5 Résumé

Notre contribution présentée dans le chapitre 5 montre que l'utilisation d'un ISM étendu avec l'inclusion des directivités des sources et des récepteurs mesurées améliore le réalisme des RIR simulées sans augmenter le temps de calcul. Les principaux changements dans cette mise en œuvre étaient l'utilisation d'harmoniques sphériques et la méthode de construction des RIR en domaine fréquentiel. Les résultats montrent que notre méthode

[1]https://phaidra.kug.ac.at/o:69292.

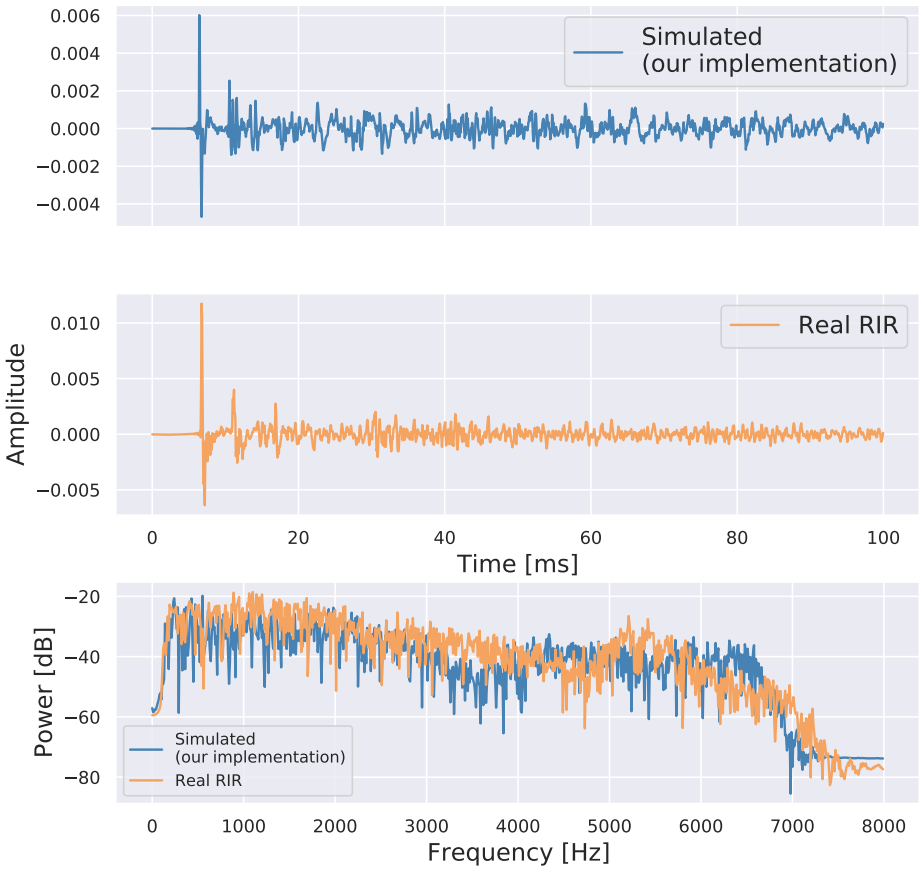

Figure 8.5: Comparaison qualitative entre une RIR réelle et une RIR simulée. La RIR réelle est extrait de la salle "011111" de l'ensemble de données dEchorate. La simulation de la RIR est effectuée dans un environnement acoustique qui ressemble étroitement aux caractéristiques de la RIR réelle. Cette simulation utilise un traitement en domaine fréquentiel et intègre une source et un microphone mesurés grâce au simulateur Pyroomacoustics modifié. Les deux premières rangées sont des RIR et la troisième rangée montre la réponse en fréquence des deux RIR.

mise en œuvre fait un pas dans la direction de l'obtention d'une réplique numérique d'une pièce réelle. En observant des résultats positifs, ce simulateur amélioré est utilisé dans le Chapitre 6 pour entraîner des modèles DNN supervisés de manière virtuelle sur deux tâches différentes, et son effet sur la généralisation du système est évalué.

8.3 Impact du réalisme de la simulation sur l'apprentissage virtuellement supervisé

Dans le chapitre 6, nous détaillons deux contributions qui se concentrent sur l'effet du réalisme de la simulation sur les systèmes de traitement audio supervisés virtuellement. Les simulations des ensembles d'entraînement sont rendues plus réalistes grâce à notre simulateur ISM avancé présenté dans le Chapitre 5. L'impact de la simulation réaliste est évalué sur deux tâches différentes, à savoir l'estimation des paramètres de la salle et la localisation de la source sonore. Tout au long de cette étude, nous améliorons la simulation des ensembles de données en ajoutant du réalisme à la source, au récepteur et aux parois de l'espace acoustique virtuel. Le réalisme des sources et des récepteurs est obtenu en ajoutant différents types de diagrammes de directivité dépendants de la fréquence, tandis que pour les parois, le sol et le plafond, nous considérons des distributions réalistes de coefficients d'absorption combinées à des réponses minimales de phase des parois. Le principal objectif de Chapitre 6 est d'étudier les performances de généralisation des modèles purement entraînés sur diverses extensions de l'ISM sur des données réelles. À cette fin, cette étude teste efficacement les modèles sur 4 ensembles de tests réels différents qui comprennent des locuteurs humains. Outre les résultats obtenus avec un modèle de référence, un modèle naïf et un modèle avancé, le Chapitre 6 présente également une étude d'ablation sur les deux tâches pour déterminer l'impact de chaque couche de réalisme ajoutée à l'ensemble de données d'entraînement.

8.3.1 Estimation des paramètres de la salle

Nous examinons la capacité de généralisation d'un modèle d'estimation des paramètres de la salle sur des ensembles de données du monde réel. Le modèle est formé à l'aide de données générées à partir du simulateur ISM étendu présenté au Chapitre 5. De plus, une étude d'ablation est menée sur 7 ensembles de données d'entraînement simulés différents, évaluant l'impact de l'expérimentation avec différentes configurations de directivité de source, de récepteur et de distributions de coefficients d'absorption des murs. Les ensembles de données sont utilisés dans le cadre du modèle d'estimation des paramètres de salle établi au Chapitre 4. Comme dans l'étude précédente, nous estimons la surface de la salle S, volume V, et temps de réverbération RT_{60} pour tous les $b \in [125, 250, 500, 1000, 2000, 4000]$ Hz étant donné un ensemble de signaux vocaux bruités.

8.3.1.1 Ensembles d'entraînement utilisés pour l'étude d'ablation

Sept ensembles de données différents nommés {D1,. . . , D7} sont simulés pour cette étude d'ablation. Chaque ensemble de données est structuré de manière à introduire progressivement différents niveaux de réalisme dans la source, les récepteurs et les murs. Tableau 8.3 résume les différents niveaux de réalisme pour chaque ensemble de données en utilisant

des notations spécifiques, qui sont expliquées ci-dessous.

Directivité de la source Trois types de directivité de la source sont pris en compte dans les ensembles de données d'entraînement, à savoir les réponses omnidirectionnelles, indépendantes de la fréquence et dépendantes de la fréquence mesurée. Les modèles de directivité indépendants de la fréquence sont mis en œuvre à l'aide d'une formule analytique. Pour cette expérience, une valeur $\psi \in [0.25, 0.5, 0.75]$ est choisi au hasard pour chaque source simulée, ce qui donne respectivement des motifs hypercardioïdes, cardioïdes et sous-cardioïdes. Dans le Tableau 8.3, de tels directivités analytiques sont désign par $(\mathcal{A}_\psi)$ et ne sont utilisées que dans l'ensemble de données D4. En outre, les sources omnidirectionnelles correspondent à la valeur $\psi = 1$ dans Equation 5.2. Cela est désigné par l'abréviation $(\mathcal{O})$ et est utilisé dans les ensembles de données D1-D5. Il convient de noter que les motifs de directivité formés par la formule analytique ne tiennent pas compte de la dépendance à l'élévation ni, plus important encore, de la dépendance en fréquence. Les modèles de directivité mesurés sont extraits de l'ensemble de données DIRPAT (Brandner et al., 2018), comme décrit en détail dans la Section 8.2.2. Les modèles de source mesurés ont été utilisés dans les ensembles de données D5, D6, D7 et est désigné par $\mathcal{M}$ dans le Table 8.3. Le motif de directivité de la source est choisi de manière aléatoire parmi trois haut-parleurs : Genelec 8020, Neumann KH120A, and Yamaha DXR8.

Directivité du récepteur Des directivités omnidirectionnelles $\mathcal{O}$ et mesurées $\mathcal{M}$ sont prises en compte pour les récepteurs dans l'étude d'ablation. Les récepteurs mesurés utilisent le motif de directivité omnidirectionnelle dépendant de la fréquence du microphone AKG C414 du jeu de données DIRPAT. Dans l'étude d'ablation, pour les récepteurs, nous avons exclu tout type de motifs non-omnidirectionnels, à la fois dépendants de la fréquence et indépendants. Cette décision a été prise car l'ensemble de test réel ne comprend que des récepteurs omnidirectionnels.

Profil d'absorption De même, deux différentes stratégies d'échantillonnage des parois sont considérées. Premièrement, l'échantillonnage naïf $(\mathcal{N})$ où chacune des six surfaces dans une pièce rectangulaire est associée à un unique coefficient indépendant de la fréquence $\alpha \in [0.02, 0.5]$, tirée uniformément au hasard. Deuxièmement, l'échantillonnage par réflectivité biaisé $(\mathcal{RB})$ (Foy et al., 2021), cette stratégie donnant des salle réalistes car chaque paroi de la salle est associée à un coefficient d'absorption dépendant de la fréquence par bandes de six octaves.

8.3.1.2 Simulation des RIR et génération de mélanges

Chaque ensemble de données est composé de $30,000$ différentes salles dont la longueur, la largeur et la hauteur sont tirées au hasard dans la boîte $[3, 10]\times[3, 10]\times[2, 4.5]$ en mètres. Dans chaque salle, 3 RIR à deux canaux sont simulés à une fréquence de 16 kHz. Un réseau de deux microphones de $22,5$ cm, similaire à celui utilisé au Chapitre 4, est placé à trois positions différentes avec une position de source fixe. Les diagrammes de directivité des microphones sont orientés de manière aléatoire sur la sphère. Les modèles de directivité non

Jeu de données	Murs	Source	Microphones
D1	$\mathcal{N}$	$\mathcal{O}$	$\mathcal{O}$
D2	$\mathcal{RB}$	$\mathcal{O}$	$\mathcal{O}$
D3	$\mathcal{RB}$	$\mathcal{O}$	$\mathcal{M}$
D4	$\mathcal{RB}$	$\mathcal{A}_\psi$	$\mathcal{O}$
D5	$\mathcal{RB}$	$\mathcal{M}$	$\mathcal{O}$
D6	$\mathcal{N}$	$\mathcal{M}$	$\mathcal{M}$
D7	$\mathcal{RB}$	$\mathcal{M}$	$\mathcal{M}$

Table 8.3: Ensembles de données d'études d'ablation et notations associées décrivant le niveau de réalisme de chaque ensemble de données.

Jeu de données Entraînement	Jeu d'essai réel					
	RT_{60} (500 Hz)	RT_{60} (1 kHz)	RT_{60} (2 kHz)	RT_{60} (4 kHz)	S	V
D1	0.193	0.160	0.108	0.185	71.00	75.68
D2	0.182	0.140	0.128	0.198	45.11	55.16
D3	0.115	**0.098**	0.078	0.156	52.76	61.82
D4	0.167	0.134	0.121	0.197	37.91	48.95
D5	0.133	0.112	**0.066**	0.155	**21.46**	**18.57**
D6	0.151	0.133	0.084	0.159	35.88	31.11
D7	**0.080**	**0.103**	**0.064**	**0.140**	32.69	30.57

Table 8.4: Erreurs absolues moyennes dans le temps de réverbération (RT_{60}, en s), surface (S, en m^2) et volume (V, en m^3) obtenue sur l'ensemble de test réel en utilisant le même modèle entraîné sur 7 ensembles de données d'entraînement simulés. Les chiffres en gras indiquent le meilleur résultat statistiquement significatif par colonne, sur la base d'intervalles de confiance à 98%.

omnidirectionnels, à la fois dépendants de la fréquence et indépendants de la fréquence, sont orientés de manière aléatoire uniquement dans la direction azimutale.

8.3.1.3 Résultats

En observant le tableau 8.4 qui décrit les résultats sur données réelles, il est évident que le modèle entraîné sur l'ensemble d'entraînement le plus réaliste, D7, produit les erreurs les plus faibles comme prévu. Ce modèle atteint de manière constante les meilleures performances, voire les deuxièmes meilleures performances, pour toutes les quantités estimées sur les ensembles de test réels. Une autre observation importante concerne la comparaison entre D5 et D7. Bien que D5 partage des caractéristiques similaires en ce qui concerne les murs et les sources avec D7, il utilise étonnamment un modèle de microphone omnidirectionnel plus simple et indépendant de la fréquence. De manière surprenante, D5 surpasse D7 en ce qui concerne l'estimation de S et de V sur l'ensemble de test réel. Cette disparité pour D7 peut être attribuée à une inadéquation du profil du microphone entre les ensem-

bles d'entraînement et de test, ce qui semble avoir un impact négatif sur les performances du système. Une solution possible pourrait être d'utiliser un ensemble diversifié de modèles de directivité de microphone dans l'ensemble d'entraînement.

8.3.2 Localisation des sources sonores

Nous nous tournons maintenant vers la tâche de localisation de source pour évaluer les améliorations de performance qui peuvent être obtenues en formant un système de localisation en utilisant des données simulées provenant de notre version améliorée du simulateur Pyroomacoustics.

Des conditions plus réalistes, telles que des sources directionnelles, des récepteurs et des parois dépendantes de la fréquence, ont déjà été intégrées avec succès en tant que modèle ISM étendu dans notre version améliorée du simulateur de salle. La Section 8.3.1 montre que, lorsqu'il est utilisé dans le but de former un modèle supervisé virtuellement pour l'estimation des paramètres acoustiques, notre simulateur avancé montre une meilleure généralisation sur des scénarios de test réels. De même, pour la localisation de la source, un travail récent de Gelderblom et al. (2021) analyse l'effet de la directivité de la source et de la modélisation de la réverbération tardive diffuse dans la simulation de RIR. Il a été constaté que la directivité de la source avait une influence positive sur les performances du système de localisation, tandis que l'inclusion de la réverbération tardive n'a montré aucun impact. De plus, leurs résultats n'ont pas été obtenus avec de la parole humaine directe dans des conditions acoustiques réelles, mais plutôt avec des signaux vocaux convolutés avec des RIRs directrices mesurées. Outre ces deux niveaux de réalisme, la localisation de la source est influencée par d'autres aspects de la scène acoustique. Un facteur important est la directivité du récepteur. Des microphones omnidirectionnels plus simples pourraient avoir un effet significatif sur les performances de localisation, en raison de la variabilité observée dans le motif de directivité à différentes fréquences. De plus, les coefficients d'absorption dépendant de la fréquence influencent le niveau de réverbération atteignant les microphones, ce qui peut être observé dans la distribution spatiale et le spectre de puissance. Tout comme dans la section précédente, une étude d'ablation est réalisée pour quantifier l'effet de chaque couche de réalisme ajoutée lors de l'entraînement. Nous présentons des résultats sur trois ensembles de tests réels distincts avec différentes configurations de microphones enregistrant de vrais locuteurs humains dans diverses conditions acoustiques afin de consolider nos conclusions.

8.3.2.1 Localisation sur des ensembles de tests réels

Pour évaluer l'influence du réalisme amélioré de l'ISM pendant l'entraînement, nous évaluons les performances des modèles de localisation supervisés virtuellement sur trois ensembles de données réels avec des annotations spatio-temporelles appropriées de l'activité des locuteurs humains et de leur position par rapport au réseau de microphones. Il s'agit de

DIRHA (Distant-Speech Interaction for Robust Home Applications) (Cristoforetti et al., 2014), VoiceHome-2 (Bertin et al., 2019) et Sony-Tau Realistic Spatial Soundscapes 2022 (STARSS22) (Politis et al., 2020). Les deux premiers ensembles de données sont conçus dans le but d'évaluer des applications pour les *smarthomes* telles que l'amélioration de la parole ou la localisation des locuteurs, tandis que le dernier est destiné au défi DCASE sur les événements sonores et la localisation (Politis et al., 2020). De plus, à partir de chaque ensemble de données, nous sélectionnons un sous-ensemble de deux microphones pour la tâche d'estimation de la direction d'arrivée (DOA), comme expliqué ci-dessous :

1. Le corpus VoiceHome-2 est enregistré à l'aide d'un réseau de microphones composé de 8 MEMS placés aux coins d'une antenne cubique. Pour cette étude, un sous-réseau à deux canaux avec une ouverture de 10.4 cm est sélectionné, et 360 des enregistrements vocaux de deux secondes dans des conditions silencieuses sont utilisés
2. Le corpus DIRHA est capturé à l'aide d'un réseau de microphones omnidirectionnels montés sur les murs et le plafond de différentes pièces. Pour cette étude, un réseau de microphones à deux canaux montés sur le mur avec une ouverture de 30 cm placé dans la salle de séjour est sélectionné, et 410 des enregistrements audio de deux secondes provenant du salon sont utilisés.
3. Le jeu de données STARSS22 est enregistré à l'aide d'un réseau sphérique Eigenmike[2] et est distribué sous deux formats : Ambisonics du premier ordre et sous-ensemble tétraédrique qui sélectionne les canaux 6, 10, 26 et 22 de l'Eigenmike. La disposition des éléments sélectionnés, les microphones individuels de l'Eigenmike forment une forme tétraédrique. Nous avons soigneusement prétraité les données pour extraire $2,100$ extraits de discours de deux secondes sans chevauchement des microphones 6 et 10 du sous-ensemble tétraédrique, avec une ouverture de 6.8 cm.

La durée totale des trois ensembles de tests soigneusement sélectionnés est de 95 minutes, comprenant des enregistrements de discours humain réel en deux canaux avec des annotations de DOA.

8.3.2.2 Génération de données sur la base de scénarios

Pour chacun des trois ensembles de test, une ensemble d'entraînement simulé naïf et un ensemble d'entraînement simulé avancé sont créés. Les ensembles d'entraînement naïfs sont composés de récepteurs et de sources omnidirectionnels, où les ouvertures des récepteurs sont les mêmes que celles de leurs ensembles de test correspondants. Les coefficients d'absorption des parois sont supposés être indépendants de la fréquence et égaux pour les six surfaces de la salle virtuelle. Les ensembles d'entraînement simulés avancés utilisent notre version avancée du simulateur Pyroomacoustics pour générer les réponses impulsionnelles des pièces. Ces ensembles d'entraînement intègrent des choix plus informés concernant les composantes de directivité et d'absorption. Pour les murs, la stratégie d'échantillonnage d'absorption biaisée vers la réflexion ($\mathcal{RB}$), telle que décrite dans Section 8.3 est utilisé.

[2] https://mhacoustics.com/productseigenmike1

En ce qui concerne la directivité de la source, les directivités mesurées spatialement d'un simulateur de tête et de torse avec bouche (Bruel & Kjaer HATS 4128-C) et de deux haut-parleurs directionnels (Genelec 8020 et YAMAHA DXR8) provenant de l'ensemble de données DIRPAT (Brandner et al., 2018) sont intégrées dans la simulation. Les directivités des récepteurs et la distance d'ouverture sont associées aux scénarios particuliers que l'on trouve dans les ensembles de tests. Dans l'ensemble d'entraînement simulé conçu pour l'ensemble de tests Voicehome-2, les récepteurs sont réglés en mode omnidirectionnel. La directivité du microphone MEMS utilisé dans le Voicehome-2 n'est pas disponible, mais est connue pour être omnidirectionnelle. Pour DIRHA, la simulation avancée place les récepteurs sur les murs de la pièce, ce qui équivaut à simuler des microphones avec une directivité en demi-sphère. Enfin, dans l'ensemble d'entraînement simulé conçu pour STARS22, la simulation avancée utilise le modèle de directivité mesurée de la sous-antenne pertinente de l'Eigenmike.

Les RIR sont simulées avec un ordre d'image source de 20. Un total de 40 000 salles rectangulaires de tailles uniformément tirées au hasard dans $[3, 10] \times [3, 10] \times [2, 4, 5]$ (en mètres) sont simulées, chacune contenant une source et un réseau de deux microphones placés uniformément au hasard avec une distance minimale entre la source et le réseau et une distance minimale entre les appareils et les murs de 30 cm. Le processus de génération des mélanges est similaire à celui utilisé dans notre étude sur l'estimation des paramètres de la pièce (voir la Section 8.1).

8.3.2.3 Sélection de modèles et hyperparamètres

À la recherche d'un estimateur DOA basé sur l'apprentissage de pointe, la longue liste de méthodes présentée par Grumiaux et al. (2022) est examinée. L'enquête est menée dans le but de trouver une méthode open source capable d'effectuer une localisation DOA multi-source, indépendamment de la configuration du réseau de microphones (c'est-à-dire de la distance entre les microphones), et qui a été testée sur des ensembles de données réels. Nous avons opté pour le modèle présenté par He et al. (2019), mis à jour par la suite par He et al. (2021). Nous avons utilisé l'architecture DNN de cette dernière étude. Cette architecture implique un DNN multi-tâches qui effectue une estimation DOA multi-source, une détection de la parole et un comptage (qui ne sont pas abordés dans cette étude).

Le modèle est entraîné sur différents ensembles d'entraînement simulés comme décrit dans la section précédente. L'entraînement est effectué avec l'optimiseur ADAM et un taux d'apprentissage de 10^{-4} sur des batchs de taille 16 pour un maximum de 110 époques, avec un arrêt anticipé sur les ensembles de validation. Dans cette étude, nous avons utilisé les mêmes caractéristiques d'entrée que celles décrites par He et al. (2021), ce qui implique la concaténation des coefficients STFT pour chaque microphone. Les coefficients STFT sont calculés avec un chevauchement de 50% et des fenêtres temporelles de 42.7 ms. Pour des raisons de cohérence, tous les signaux utilisés dans notre étude ont été sous-échantillonnés à 16 kHz.

Jeu de tests réels →	VoiceHome-2		DIRHA		STARS22	
Méthodes	↑ Recall	↓ MAE (°)	↑ Recall	↓ MAE (°)	↑ Recall	↓ MAE (°)
SRP-PHAT	70%	9.9 ± 1.5	61%	15.0 ± 2.3	45%	14.9 ± 0.6
Entraînement naïf	78%	7.6 ± 1.2	77%	8.4 ± 1.4	57%	12.9 ± 0.6
Entraînement avancée	**85%**	**5.8 ± 0.8**	**84%**	**6.3 ± 1.0**	61%	**11.4 ± 0.5**
Étude d'ablation						
sans réalisme mur	83%	6.2 ± 0.8	81%	7.5 ± 1.4	59%	12.1 ± 0.6
sans réalisme source	82%	7.1 ± 1.1	80%	7.8 ± 1.2	**63%**	**11.4 ± 0.6**
sans réalisme récepteur	N/A	N/A	78%	8.3 ± 1.5	53%	13.4 ± 0.6

Table 8.5: Résultats de localisation sur trois ensembles de tests réels obtenus par la méthode SRP-PHAT de référence et par le modèle supervisé de He et al. (2021) formé en utilisant différents modes de simulation. Les erreurs angulaires moyennes (MAE) sont affichées avec leur intervalle de confiance à 95%. Les chiffres en gras indiquent le meilleur système dans chaque colonne et les systèmes statistiquement équivalents. La signification statistique a été évaluée à l'aide du test de McNemar pour la métrique de Recall et des intervalles de confiance à 95% sur les différences d'erreur angulaire pour la métrique de MAE.

8.3.2.4 Expériences et résultats

Système de référence et mesures d'évaluation

Les modèles d'estimation DOA supervisés de manière virtuelle sont évalués à l'aide de deux métriques différentes, à savoir l'erreur angulaire moyenne (MAE, en degrés) et le rappel (en %). Le rappel est défini comme le rapport des sources localisées avec une erreur inférieure à 10°. Les trois modèles entraînés sur des données simulées, à la fois naïves et avancées, sont comparés à la méthode de localisation SRP-PHAT classique sans apprentissage mise en œuvre par Scheibler et al. (2018).

Jeu de tests réels

Les modèles entraînés et la méthode SRP-PHAT de base sont comparés sur les trois ensembles de données réels. La partie supérieure du Tableau 8.5 présente les résultats. On observe que l'approche d'entraînement avancée surpasse de manière cohérente l'approche d'entraînement naïve, atteignant de 4 à 7 points de rappel supplémentaires et une marge de 2° MAE sur l'ensemble des trois ensembles de données, malgré l'utilisation de la même architecture de réseau. La référence SRP-PHAT est constamment surpassée par l'approche d'entraînement avancée avec une marge de 15 à 23 points de rappel et de 3° à 9° erreurs angulaires moyennes. Les résultats révèlent également que l'estimation du DOA sur l'ensemble de données STARSS22 est difficile, ce qui correspond à la description de l'ensemble de données. De plus, une étude d'ablation sur la stratégie de simulation avancée proposée est également présentée dans la moitié inférieure du Tableau 8.5. Le consensus général tiré de ces résultats est que la suppression de l'une des trois couches de réalisme entraîne une baisse perceptible des performances. Une exception notable se produit sur l'ensemble de données

STARSS22, où l'utilisation de sources directionnelles mesurées lors de l'apprentissage ne semble pas améliorer les performances. Une explication possible de cette observation est que les locuteurs humains de l'ensemble de données STARS22 effectuent des rotations significatives de la tête, ce qui n'est pas pris en compte dans notre approche.

8.3.3 Résumé

Dans ce chapitre, nous avons utilisé notre simulateur avancé Pyroomacoustics pour simuler des RIR réalistes pour les ensembles d'entraînement, et évalué son impact sur des systèmes d'apprentissage virtuellement supervisés. Nous avons soigneusement élaboré des ensembles d'entraînement simulés en incorporant diverses couches de réalisme dans les réponses des sources, des récepteurs et des parois pour évaluer la généralisation des modèles pour les tâches d'estimation des paramètres d'une pièce et de localisation de locuteurs. Nos résultats pour les deux tâches démontrent que chaque couche de réalisme ajoutée améliore considérablement les performances d'estimation des modèles DNN sur les ensembles de tests réels. En particulier, l'incorporation de la directivité de la source et de l'échantillonnage des parois biaisé vers la réflectivité réduit les erreurs pour les deux tâches par rapport aux modèles formés sur des ensembles d'entraînement naïfs. Cela démontre que, en utilisant une extension de la simulation ISM, nous pouvons améliorer les performances des systèmes virtuellement supervisés sans entraîner de coûts informatiques supplémentaires.

8.4 Les perspectives

Les recherches menées dans cette thèse ouvrent la voie à l'exploration de plusieurs avenues intrigantes pour de futures investigations. Plusieurs de ces orientations potentielles sont décrites ci-dessous.

8.4.1 Simulateur avancé Pyroomacoustics

Dans le chapitre 5, nous décrivons notre mise en œuvre de l'ISM étendu dans le simulateur open-source Pyroomacoustics (Scheibler et al., 2018). En continuant dans la voie des améliorations, il existe la possibilité d'augmenter les données pour les réponses mesurées des sources et des microphones. La mise en œuvre de cette méthode conduirait à une plus grande diversité dans les motifs de directivité, ce qui peut directement affecter la généralisation des systèmes audio supervisés de manière virtuelle. Dans notre mise en œuvre de l'ISM étendu, la réverbération tardive du RIR est modélisée à l'aide de l'ISM, cependant, une plus grande réalisme dans le RIR peut être atteint en le modélisant à l'aide de la méthode de traçage stochastique des rayons (Gelderblom et al., 2021). La combinaison de la directivité des sources et des microphones avec les méthodes de traçage des rayons est une tâche complexe, dont la mise en œuvre est présentée dans le simulateur de Schimmel et al. (2009) en

utilisant des histogrammes angle-temps-fréquence.

Une autre avenue d'amélioration consiste à introduire une impédance murale à valeurs complexes. Une étude menée par Meissner and Zielinski (2022) révèle qu'il existe des distinctions perceptuelles dans la réverbération entre les murs à impédance réelle et ceux à impédance à valeurs complexes, notamment dans les petites pièces. Dans la même optique, l'inclusion des effets de diffraction dans les simulateurs d'acoustique géométrique améliorera le réalisme du RIR simulé. Dans cette thèse, nous avons travaillé avec des pièces de forme rectangulaire standard en raison des limitations des simulateurs. Des opportunités d'application plus vastes peuvent être observées si ces techniques sont étendues aux pièces de forme non rectangulaire. Maintenir un temps de calcul réduit tout en utilisant toutes ces techniques serait un défi, mais cela donnerait à ce simulateur un avantage par rapport aux méthodes de simulation basées sur les ondes (Hamilton, 2021), qui sont capables de simuler des RIR avec le plus grand réalisme, mais en raison de leur temps de calcul élevé, peuvent difficilement être utilisées pour former des méthodes d'apprentissage supervisées de manière virtuelle.

8.4.2 Estimation des paramètres de la pièce

Un modèle d'estimation conjointe des paramètres de la pièce basé sur un DNN est présenté au Chapitre 4. Le même modèle est utilisé au Chapitre 6 pour étudier l'influence de l'amélioration du réalisme des données d'entraînement grâce à une étude d'ablation. En ce qui concerne l'énoncé du problème et la conception de la chaîne de traitement, des explorations plus poussées peuvent être menées sur de nombreux fronts. Cela pourrait inclure une extension du modèle actuel d'estimation des paramètres de la pièce vers l'estimation conjointe de paramètres locaux tels que la position et les propriétés de la source et des surfaces individuelles. Cela peut s'accompagner d'expérimentations sur différentes conceptions d'architecture de DNN et des caractéristiques d'entrée efficaces utilisées avec des configurations de microphones spécifiques, telles que les réseaux circulaires ou un Eigenmike.

Un point de départ viable est l'étude récente menée par Ick et al. (2023) qui suggère l'utilisation de spectrogrammes Gammatone combinés avec des spectrogrammes de phase Gammatone pour l'estimation conjointe du temps de réverbération et du volume. De plus, des techniques d'augmentation des données sur des données réelles pourraient être envisagées lors de la formation du système, ce qui pourrait être comparé au modèle formé uniquement avec une approche supervisée virtuellement. Pour concrétiser cela, des efforts devraient être déployés dans la mesure d'un nouvel ensemble de données conçu spécifiquement pour l'estimation des paramètres de la pièce, impliquant différentes conditions acoustiques avec une annotation appropriée des paramètres acoustiques globaux et locaux. Le besoin de telles bases de données a été démontré dans diverses études sur l'estimation des paramètres de la pièce (Ick et al., 2023; Xiong et al., 2018; Genovese et al., 2019). Sur la base des résultats présentés au Chapitre 4 et 6, les deux contributions manquent d'une étude détaillée sur la formation des systèmes avec des données comprenant différentes dis-

tributions de bruit et leur impact sur les ensembles de tests réels. De plus, les systèmes peuvent être testés sur divers ensembles de tests du monde réel pour renforcer davantage l'affirmation de la généralisation. Enfin, mener une enquête approfondie en utilisant une gamme de réponses de microphones mesurées est essentiel pour étayer des assertions comparables à celles faites concernant la directivité des sources, qui a effectivement démontré son utilité dans l'estimation des paramètres géométriques de la pièce.

Un accent pourrait également être mis sur l'estimation en temps réel des paramètres acoustiques dans des conditions dynamiques, ce qui pourrait être utilisé dans différents dispositifs mobiles ou cadres de traitement audio, le rendant ainsi plus accessible et utilisable dans des scénarios du monde réel. Cela pourrait s'intégrer à diverses applications audio telles que la réduction du bruit, l'amélioration audio, la réalité virtuelle et la réalité augmentée, afin d'améliorer l'expérience utilisateur.

8.4.3 Localisation de la source sonore

Notre contribution à la localisation des orateurs, telle que discutée au Chapitre 6, vise à explorer les techniques de simulation et leur influence sur la généralisation d'un système formé grâce à des méthodes supervisées virtuellement. Les résultats de cette étude révèlent certaines conclusions qui nécessitent une investigation plus approfondie à travers des expériences supplémentaires. Nous manquons d'une étude détaillée examinant l'impact de différentes distributions de bruit au sein des ensembles d'entraînement et leur corrélation avec les performances sur les ensembles de tests réels. Au cours de la phase d'entraînement, nous avons observé que le modèle formé avec des ensembles d'entraînement avancés atteint la convergence plus rapidement que celui formé sur des ensembles de données plus simples. Cependant, une exploration plus approfondie est nécessaire pour établir une conclusion définitive dans cette direction. De plus, à l'instar d'autres contributions, l'impact de l'introduction de plus de motifs de directivité mesurés pour les microphones dans la formation du système est une direction de recherche importante. De plus, la portée de cette étude peut être étendue à davantage de scénarios, tels que la localisation des sources sonores dans des environnements acoustiques dynamiques avec une estimation de l'angle d'arrivée de multiples sources. Une autre avenue prometteuse d'expérimentation concerne la localisation des sources en coordonnées 2D et 3D.

Bibliography

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2018a). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48.

Adavanne, S., Politis, A., and Virtanen, T. (2018b). Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466.

Adavanne, S., Politis, A., and Virtanen, T. (2019). Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network. *arXiv preprint arXiv:1904.12769*.

Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

Antonacci, F., Filos, J., Thomas, M. R., Habets, E. A., Sarti, A., Naylor, P. A., and Tubaro, S. (2012). Inference of room geometry from acoustic impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2683–2695.

Antonello, N., van Waterschoot, T., Moonen, M., and Naylor, P. A. (2015). Evaluation of a numerical method for identifying surface acoustic impedances in a reverberant room. In *10th European Congress and Exposition on Noise Control Engineering*, pages 1–6.

Arberet, S., Ozerov, A., Duong, N. Q., Vincent, E., Gribonval, R., Bimbot, F., and Vandergheynst, P. (2010). Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, pages 1–4.

Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., and MacIntyre, B. (2001). Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, 21(6):34–47.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Barumerli, R., Bianchi, D., Geronazzo, M., and Avanzini, F. (2021). SofaMyRoom: a fast and multiplatform" shoebox" room simulator for binaural room impulse response dataset generation. *arXiv preprint arXiv:2106.12992*.

Bechler, D. and Kroschel, K. (2003). Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array. In *International Workshop on Acoustic Echo and Noise Control*, pages 315–318.

Behboodi, B. and Rivaz, H. (2019). Ultrasound segmentation using U-Net: learning from simulated data and testing on real data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6628–6631.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K. (2010). Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39.

Benedetto, J. J. (2021). *Frame decompositions, sampling, and uncertainty principle inequalities*. CRC Press.

Bertin, N., Camberlein, E., Lebarbenchon, R., Vincent, E., Sivasankaran, S., Illina, I., and Bimbot, F. (2019). VoiceHome-2, an extended corpus for multichannel speech processing in real homes. *Speech Communication*, 106:68–78.

Bertin, N., Camberlein, E., Vincent, E., Lebarbenchon, R., Peillon, S., Lamandé, É., Sivasankaran, S., Bimbot, F., Illina, I., Tom, A., et al. (2016a). A french corpus for distant-microphone speech processing in real homes. In *Interspeech 2016*.

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

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C.-A. (2019). Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5):3590–3628.

Bird, J. J., Faria, D. R., Ekárt, A., and Ayrosa, P. P. (2020). From simulation to reality: CNN transfer learning for scene classification. In *2020 IEEE 10th International Conference on Intelligent Systems (IS)*, pages 619–625.

Bohlender, A., Spriet, A., Tirry, W., and Madhu, N. (2021). Exploiting temporal context in CNN-based multisource DOA estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1594–1608.

Bologni, G., Heusdens, R., and Martinez, J. (2021). Acoustic reflectors localization from stereo recordings using neural networks. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Borish, J. (1984). Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75(6):1827–1836.

Bork, I. (2000). A comparison of room simulation software-the 2nd round robin on room acoustical computer simulation. *Acta Acustica united with Acustica*, 86(6):943–956.

Brandão, E., Lenzi, A., and Paul, S. (2015). A review of the in situ impedance and sound absorption measurement techniques. *Acta Acustica united with Acustica*, 101(3):443–463.

Brandner, M., Frank, M., and Rudrich, D. (2018). DirPat—Database and viewer of 2D/3D directivity patterns of sound sources and receivers. In *Audio Engineering Society Convention 144*.

Brinkmann, F., Erbes, V., and Weinzierl, S. (2019). *Extending the closed form image source model for source directivity*. Technische Universität Berlin.

Bryan, N. J. (2020). Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Busso, C., Hernanz, S., Chu, C.-W., Kwon, S.-i., Lee, S., Georgiou, P. G., Cohen, I., and Narayanan, S. (2005). Smart Room: Participant and speaker localization and identification. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Campbell, D., Palomaki, K., and Brown, G. (2005). A MATLAB simulation of "shoebox" room acoustics for use in research and teaching. *Computing and Information Systems*, 9(3):48.

Candes, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223.

Chakrabarty, S. and Habets, E. A. (2019). Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):8–21.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017). Matterport3d: Learning from RGB-D data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Chazan, S. E., Hammer, H., Hazan, G., Goldberger, J., and Gannot, S. (2019). Multi-microphone speaker separation based on deep DOA estimation. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Chen, C., Jain, U., Schissler, C., Gari, S. V. A., Al-Halah, Z., Ithapu, V. K., Robinson, P., and Grauman, K. (2020). Soundspaces: Audio-visual navigation in 3D environments. In *2020 European Conference on Computer Vision (EECV)*, pages 17–36.

Christensen, J. H., Hornauer, S., and Stella, X. Y. (2020). Batvision: Learning to see 3D spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1581–1587.

Chu, S., Narayanan, S., and Kuo, C.-C. J. (2009). Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158.

Couvreur, L. and Couvreur, C. (2000). On the use of artificial reverberation for ASR in highly reverberant environments. In *2nd IEEE Benelux Signal Processing Symposium (SPS)*, pages S001–S004.

Cox, T. J., Li, F., and Darlington, P. (2001). Extracting room reverberation time from speech using artificial neural networks. *Journal of the Audio Engineering Society*, 49(4):219–230.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M., and Maragos, P. (2014). The DIRHA simulated corpus. In *2014 The International Conference on Language Resources and Evaluation*, pages 2629–2634.

Crocco, M., Cristani, M., Trucco, A., and Murino, V. (2016). Audio surveillance: A systematic review. *ACM Computing Surveys*, 48(4):1–46.

Crocco, M. and Del Bue, A. (2016). Estimation of TDOA for room reflections by iterative weighted l1 constraint. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3201–3205.

Del Vallado, J., de Lima, A. A., Prego, T. d. M., and Netto, S. L. (2013). Feature analysis for the reverberation perception in speech signals. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8169–8173.

Deleforge, A. and Horaud, R. (2012). 2D sound-source localization on the binaural manifold. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6.

Deleforge, A., Horaud, R., Schechner, Y. Y., and Girin, L. (2015). Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):718–731.

Deng, S., Mack, W., and Habets, E. A. (2020). Online Blind Reverberation Time Estimation Using CRNNs. In *Interspeech*, pages 5061–5065.

Di Carlo, D. (2020). *Echo-Aware Signal Processing for Audio Scene Analysis*. PhD thesis, Université de Rennes 1; INRIA-IRISA-PANAMA.

Di Carlo, D., Deleforge, A., and Bertin, N. (2019). Mirage: 2D source localization using microphone pair augmentation with echoes. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 775–779.

Di Carlo, D., Tandeitnik, P., Foy, C., Bertin, N., Deleforge, A., and Gannot, S. (2021). dEchorate: a calibrated room impulse response dataset for echo-aware signal processing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021:1–15.

Diaz-Guerra, D., Miguel, A., and Beltran, J. R. (2020). Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:300–311.

Diaz-Guerra, D., Miguel, A., and Beltran, J. R. (2021). gpuRIR: A python library for room impulse response simulation with GPU acceleration. *Multimedia Tools and Applications*, 80:5653–5671.

DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (2001). Robust localization in reverberant rooms. pages 157–180.

Dilungana, S., Deleforge, A., Foy, C., and Faisan, S. (2021). Learning-based estimation of individual absorption profiles from a single room impulse response with known positions of source, sensor and surfaces. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, number 1, pages 5623–5630.

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

Ding, J., Ke, Y., Cheng, L., Zheng, C., and Li, X. (2020). Joint estimation of binaural distance and azimuth by exploiting deep neural networks. *The Journal of the Acoustical Society of America*, 147(4):2625–2635.

Dokmanić, I., Lu, Y. M., and Vetterli, M. (2011). Can one hear the shape of a room: The 2-D polygonal case. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 321–324.

Driscoll, J. R. and Healy, D. M. (1994). Computing Fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics*, 15(2):202–250.

Eaton, J., Gaubitch, N. D., Moore, A. H., and Naylor, P. A. (2016). Estimation of room acoustic parameters: The ACE challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1681–1693.

El Baba, Y., Walther, A., and Habets, E. A. (2016). Reflector localization based on multiple reflection points. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1458–1462.

El Baba, Y., Walther, A., and Habets, E. A. (2017). 3D room geometry inference based on room impulse response stacks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5):857–872.

Emmanuel, P., Parrish, N., and Horton, M. (2021). Multi-scale network for sound event localization and detection. Technical report.

Evers, C., Löllmann, H. W., Mellmann, H., Schmidt, A., Barfuss, H., Naylor, P. A., and Kellermann, W. (2020). The LOCATA challenge: Acoustic source localization and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1620–1643.

Evers, C. and Naylor, P. A. (2018). Acoustic Slam. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1484–1498.

Fernandez-Grande, E., Bianco, M. J., Gannot, S., and Gerstoft, P. (2021). DTU three-channel room impulse response dataset for direction of arrival estimation 2020. *IEEE Dataport*.

Foy, C., Deleforge, A., and Di Carlo, D. (2021). Mean absorption estimation from room impulse responses using virtually supervised learning. *The Journal of the Acoustical Society of America*, 150(2):1286–1299.

Fu, Z.-h. and Li, J.-w. (2016). GPU-based image method for room impulse response calculation. *Multimedia Tools and Applications*, 75:5205–5221.

Gamper, H. and Tashev, I. J. (2018). Blind reverberation time estimation using a convolutional neural network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 136–140.

Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730.

Gao, S., Wu, X., and Qu, T. (2022). Room geometry blind inference based on the localization of real sound source and first order reflections. *arXiv preprint arXiv:2207.10478*.

Gaubitch, N. D., Loellmann, H. W., Jeub, M., Falk, T. H., Naylor, P. A., Vary, P., and Brookes, M. (2012). Performance comparison of algorithms for blind reverberation time estimation from speech. In *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pages 1–4.

Gaultier, C., Kataria, S., and Deleforge, A. (2017). VAST: The virtual acoustic space traveler dataset. In *13th International Conference on Latent Variable Analysis and Signal Separation*, pages 68–79.

Gelderblom, F. B., Liu, Y., Kvam, J., and Myrvoll, T. A. (2021). Synthetic data for DNN-based DoA estimation of indoor speech. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4390–4394.

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

Geronazzo, M., Fantin, J., Sorato, G., Guido, B., Avanzini, F., et al. (2016). The selfear project: a mobile application for low-cost pinna-related transfer function acquisition. In *Proceedings of the International Conference on Sound and Music Computing*, pages 164–171.

Geronazzo, M., Granza, F., Spagnol, S., and Avanzini, F. (2013). A standardized repository of head-related and headphone impulse response data. In *Audio Engineering Society Convention 134*.

Gerzon, M. A. (1975). The design of precisely coincident microphone arrays for stereo and surround sound. In *Audio Engineering Society Convention 50*.

González, Á. (2010). Measurement of areas on a sphere using fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, 42:49–64.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

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

Grondin, F., Lauzon, J.-S., Michaud, S., Ravanelli, M., and Michaud, F. (2020). BIRD: Big impulse response dataset. *arXiv preprint arXiv:2010.09930*.

Grumiaux, P.-A., Kitić, S., Girin, L., and Guérin, A. (2021a). Improved feature extraction for CRNN-based multiple sound source localization. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 231–235.

Grumiaux, P.-A., Kitić, S., Girin, L., and Guérin, A. (2022). A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, 152(1):107–151.

Grumiaux, P.-A., Kitić, S., Srivastava, P., Girin, L., and Guérin, A. (2021b). SALADnet: Self-attentive multisource localization in the Ambisonics domain. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 336–340.

Habets, E. A. (2006). Room impulse response generator. *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4):1.

Habets, E. A. and Gannot, S. (2007). Generating sensor signals in isotropic noise fields. *The Journal of the Acoustical Society of America*, 122(6):3464–3470.

Habets, E. A. P. (2007). Single- and multi-microphone speech dereverberation using spectral enhancement.

Hadad, E., Heese, F., Vary, P., and Gannot, S. (2014). Multichannel audio database in various acoustic environments. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317.

Hald, J., Song, W., Haddad, K., Jeong, C.-H., and Richard, A. (2019). In-situ impedance and absorption coefficient measurements using a double-layer microphone array. *Applied Acoustics*, 143:74–83.

Hamilton, B. (2021). Pffdtd software. https://github.com/bsxfun/pffdtd.

Hao, Y., Küçük, A., Ganguly, A., and Panahi, I. M. (2020). Spectral flux-based convolutional neural network architecture for speech source localization and its real-time implementation. *IEEE Access*, 8:197047–197058.

He, W., Motlicek, P., and Odobez, J.-M. (2019). Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 770–774.

He, W., Motlicek, P., and Odobez, J.-M. (2021). Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1303–1317.

Heller, F., Krämer, A., and Borchers, J. (2014). Simplifying orientation measurement for mobile audio augmented reality applications. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 615–624.

Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., and Nagamatsu, K. (2020). End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. *arXiv preprint arXiv:2005.09921*.

Hu, Y., Chen, C., Zou, H., Zhong, X., and Chng, E. S. (2023). Unifying speech enhancement and separation with gradient modulation for end-to-end noise-robust speech separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Huang, J. and Bocklet, T. (2019). Intel Far-Field Speaker Recognition System for VOiCES Challenge 2019. In *Interspeech*, pages 2473–2477.

Huang, J., Ohnishi, N., and Sugie, N. (1997). Sound localization in reverberant environment based on the model of the precedence effect. *IEEE Transactions on Instrumentation and Measurement*, 46(4):842–846.

Ick, C., Mehrabi, A., and Jin, W. (2023). Blind Acoustic Room Parameter Estimation Using Phase Features. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

ISO.354:2003 (2003). *Acoustics - measurement of sound absorption in a reverberation room*. International Organization for Standardization, Vernier, Geneva, Switzerland, ISO 354:2003 edition.

ISO.ASTM:E1050-9 (2006). *Standard Test Method for Impedance and Absorption of Acoustical Materials Using a Tube, Two Microphones and a Digital Frequency Analysis System*. American Society for Testing and Materials.

Jager, I., Heusdens, R., and Gaubitch, N. D. (2016). Room geometry estimation from acoustic echoes using graph-based echo labeling. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jarrett, D. P., Habets, E. A., Thomas, M. R., and Naylor, P. A. (2012). Rigid sphere room impulse response simulation: Algorithm and applications. *The Journal of the Acoustical Society of America*, 132(3):1462–1472.

Jenrungrot, T., Jayaram, V., Seitz, S., and Kemelmacher-Shlizerman, I. (2020). The cone of silence: Speech separation by localization. *Advances in Neural Information Processing Systems*, 33:20925–20938.

Jot, J.-M. and Lee, K. S. (2016). Augmented reality headphone environment rendering. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*.

Jot, J.-M., Lee, K. S., and Stein, E. (2018). Augmented reality headphone environment rendering. US Patent 10,038,967.

Kagami, S., Sasaki, Y., Thompson, S., Fujihara, T., Enomoto, T., and Mizoguchi, H. (2008). Loudness measurement of human utterance to a robot in noisy environment. In *the 3rd ACM/IEEE international conference on Human robot interaction*, pages 217–224.

Kaiming, H., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645.

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

Kearney, G., Gorzel, M., Rice, H., and Boland, F. (2012). Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields. *Acta Acustica united with Acustica*, 98(1):61–71.

Kim, K., Billinghurst, M., Bruder, G., Duh, H. B.-L., and Welch, G. F. (2018). Revisiting trends in augmented reality research: A review of the 2nd decade of ismar (2008–2017). *IEEE transactions on visualization and computer graphics*, 24(11):2947–2962.

Kim, Y. and Ling, H. (2011). Direction of arrival estimation of humans with a small sensor array using an artificial neural network. *Progress In Electromagnetics Research B*, 27:127–149.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitić, S., Bertin, N., and Gribonval, R. (2014). Hearing behind walls: localizing sources in the room next door with cosparsity. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3087–3091.

Knüttel, T., Witew, I. B., and Vorländer, M. (2013). Influence of "omnidirectional" loudspeaker directivity on measured room impulse responses. *The Journal of the Acoustical Society of America*, 134(5):3654–3662.

Kowalczyk, K., Habets, E. A., Kellermann, W., and Naylor, P. A. (2013). Blind system identification using sparse learning for TDOA estimation of room reflections. *IEEE Signal Processing Letters*, 20(7):653–656.

Koyama, Y., Shigemi, K., Takahashi, M., Shimada, K., Takahashi, N., Tsunoo, E., Takahashi, S., and Mitsufuji, Y. (2022). Spatial data augmentation with simulated room impulse responses for sound event localization and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8872–8876.

Kreković, M., Dokmanić, I., and Vetterli, M. (2016). EchoSLAM: Simultaneous localization and mapping with acoustic echoes. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11–15.

Kuttruff, H. (2016). *Room acoustics*. Crc Press.

Lathoud, G., Odobez, J.-M., and Gatica-Perez, D. (2005). AV16. 3: An audio-visual corpus for speaker localization and tracking. In *Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers 1*, pages 182–195.

Lebart, K., Boucher, J.-M., and Denbigh, P. N. (2001). A new method based on spectral subtraction for speech dereverberation. *Acta Acustica united with Acustica*, 87(3):359–366.

Lee, M. and Chang, J.-H. (2016). Blind estimation of reverberation time using deep neural network. In *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 308–311.

Lee, M. and Chang, J.-H. (2018). Deep neural network based blind estimation of reverberation time based on multi-channel microphones. *Acta Acustica united with Acustica*, 104(3):486–495.

Lee, W. S. and Lee, S. S. (2008). Piezoelectric microphone built on circular diaphragm. *Sensors and Actuators A: Physical*, 144(2):367–373.

Li, X., Yi, W., Chi, H.-L., Wang, X., and Chan, A. P. (2018). A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Automation in Construction*, 86:150–162.

Lovedee-Turner, M. and Murphy, D. (2019). Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array. *The Journal of the Acoustical Society of America*, 146(5):3339–3352.

Luo, Y. and Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266.

Luo, Y. and Yu, J. (2022). FRA-RIR: Fast Random Approximation of the Image-source Method. *arXiv preprint arXiv:2208.04101*.

Lyons, R. (2004). *Understanding Digital Signal Processing*. Prentice Hall professional technical reference. Prentice Hall/PTR.

Ma, N., Brown, G., and May, T. (2015). Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. In *Interspeech*, volume 2015, pages 160–164.

Ma, W. and Liu, X. (2019). Phased microphone array for sound source localization with deep learning. *Aerospace Systems*, 2(2):71–81.

Mabande, E., Kowalczyk, K., Sun, H., and Kellermann, W. (2013). Room geometry inference based on spherical microphone array eigenbeam processing. *The Journal of the Acoustical Society of America*, 134(4):2773–2789.

Madhu, N. and Martin, R. (2008). A scalable framework for multiple speaker localization and tracking. *Proceedings of the International Workshop on Acoustic Echo and Noise Control*, pages 1–4.

Maeda, I., DeGraw, D., Kitano, M., Matsushima, H., Sakaji, H., Izumi, K., and Kato, A. (2020). Deep reinforcement learning in agent based financial market simulation. *Journal of Risk and Financial Management*, 13(4):71.

Majdak, P., Iwaya, Y., Carpentier, T., Nicol, R., Parmentier, M., Roginska, A., Suzuki, Y., Watanabe, K., Wierstorf, H., Ziegelwanger, H., et al. (2013). Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*.

Malek, A., Chazan, S. E., Malka, I., Tourbabin, V., Goldberger, J., Tzirkel-Hancock, E., and Gannot, S. (2017). Speaker extraction using LCMV beamformer with DNN-based SPP and RTF identification scheme. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2274–2278.

Malmgren-Hansen, D., Kusk, A., Dall, J., Nielsen, A. A., Engholm, R., and Skriver, H. (2017). Improving SAR automatic target recognition models with transfer learning from simulated data. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1484–1488.

Mascia, M., Canclini, A., Antonacci, F., Tagliasacchi, M., Sarti, A., and Tubaro, S. (2015). Forensic and anti-forensic analysis of indoor/outdoor classifiers based on acoustic clues. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 2072–2076.

Mattheiss, E., Regal, G., Vogelauer, C., and Furtado, H. (2020). 3D Audio Navigation-Feasibility and Requirements for Older Adults. In *2020 17th International Conference on Computers Helping People with Special Needs (ICCHP)*, pages 323–331.

Meissner, M. and Zielinski, T. (2022). Impact of Wall Impedance Phase Angle on Indoor Sound Field and Reverberation Parameters Derived from Room Impulse Response. *Archives of Acoustics*, 47(3):343–353.

Moore, A. H., Brookes, M., and Naylor, P. A. (2014). Room identification using roomprints. In *Audio Engineering Society Conference: 54th International Conference: Audio Forensics*.

Moore, A. H., Evers, C., and Naylor, P. A. (2016). Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):178–192.

Murgai, P., Rau, M., and Jot, J.-M. (2017). Blind estimation of the reverberation fingerprint of unknown acoustic environments. In *Audio Engineering Society Convention 143*.

Murphy, D. T. and Shelley, S. (2010). Openair: An interactive auralization web resource and database. In *Audio Engineering Society Convention 129*.

Nakadai, K., Kumon, M., Okuno, H. G., Hoshiba, K., Wakabayashi, M., Washizaki, K., Ishiki, T., Gabriel, D., Bando, Y., Morito, T., et al. (2017). Development of microphone-array-embedded uav for search and rescue task. In *2017 International Conference on Intelligent Robots and Systems (IROS)*, pages 5985–5990.

Nakamura, S., Hiyane, K., Asano, F., and Endo, T. (1999). Sound scene data collection in real acoustical environments. *Journal of the Acoustical Society of Japan (E)*, 20(3):225–231.

Nam, S., Davies, M. E., Elad, M., and Gribonval, R. (2013). The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56.

Nava, G. P., Yasuda, Y., Sato, Y., and Sakamoto, S. (2009). On the in situ estimation of surface acoustic impedance in interiors of arbitrary shape by acoustical inverse methods. *Acoustical science and technology*, 30(2):100–109.

Nguyen, T. N. T., Gan, W.-S., Ranjan, R., and Jones, D. L. (2020). Robust source counting and doa estimation using spatial pseudo-spectrum and convolutional neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2626–2637.

Okawa, Y., Watanabe, Y., Ikeda, Y., and Oikawa, Y. (2021). Estimation of acoustic impedances in a room using multiple sound intensities and fdtd method. In *Advances in Acoustics, Noise and Vibration-Proceedings of the 27th International Congress on Sound and Vibration, ICSV*.

Pan, X., You, Y., Wang, Z., and Lu, C. (2017). Virtual to real reinforcement learning for autonomous driving. *arXiv preprint arXiv:1704.03952*.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210.

Pariente, M. (2021). *Implicit and explicit phase modeling in deep learning-based source separation*. PhD thesis, Université de Lorraine.

Park, S. and Choi, J.-W. (2021). Iterative echo labeling algorithm with convex hull expansion for room geometry estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1463–1478.

Paul, D. B. and Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. In *workshop Speech and Natural Language: Held at Harriman, New York, February 23-26, 1992*.

Peng, F., Wang, T., and Chen, B. (2015). Room shape reconstruction with a single mobile acoustic sensor. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1116–1120.

Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2018). CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 241–245.

Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2019). CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):22–33.

Pertilä, P. and Cakir, E. (2017). Robust direction estimation with convolutional neural networks based steered response power. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6125–6129.

Politis, A., Mesaros, A., Adavanne, S., Heittola, T., and Virtanen, T. (2020). Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:684–698.

Politis, A., Shimada, K., Sudarsanam, P., Adavanne, S., Krause, D., Koyama, Y., Takahashi, N., Takahashi, S., Mitsufuji, Y., and Virtanen, T. (2022). STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv preprint arXiv:2206.01948*.

Poschadel, N., Hupke, R., Preihs, S., and Peissig, J. (2021). Direction of arrival estimation of noisy speech using convolutional recurrent neural networks with higher-order ambisonics signals. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 211–215.

Prawda, K., Schlecht, S., Välimäki, V., et al. (2021). Room acoustic parameters measurements in variable acoustic laboratory arni. *Proceedings of Akustiikkapäivät*, pages 24–25.

Pujol, H., Bavu, E., and Garcia, A. (2019). Source localization in reverberant rooms using Deep Learning and microphone arrays. In *23rd International Congress on Acoustics (ICA 2019 Aachen)*.

Purushwalkam, S., Gari, S. V. A., Ithapu, V. K., Schissler, C., Robinson, P., Gupta, A., and Grauman, K. (2021). Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1183–1192.

Raghuvanshi, N., Lloyd, B., Govindaraju, N., and Lin, M. C. (2009). Efficient numerical acoustic simulation on graphics processors using adaptive rectangular decomposition. In *Proceedings of the EAA Symposium on Auralization*.

Ratnam, R., Jones, D. L., Wheeler, B. C., O'Brien Jr, W. D., Lansing, C. R., and Feng, A. S. (2003). Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892.

Ratnarajah, A., Zhang, S.-X., Yu, M., Tang, Z., Manocha, D., and Yu, D. (2021). FAST-RIR: Fast neural diffuse room impulse response generator. *arXiv preprint ARXIV.2110.04057*.

Remaggi, L., Jackson, P. J., Coleman, P., and Wang, W. (2016). Acoustic reflector localization: Novel image source reversion and direct localization methods. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):296–309.

Richter, J., Welker, S., Lemercier, J.-M., Lay, B., and Gerkmann, T. (2023). Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Richter, S. R., AlHaija, H. A., and Koltun, V. (2022). Enhancing photorealism enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1700–1715.

Röber, N., Spindler, M., and Masuch, M. (2006). Waveguide-based Room Acoustics through Graphics Hardware. In *ICMC*.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Savioja, L. and Svensson, U. P. (2015). Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730.

Scheibler, R., Bezzam, E., and Dokmanić, I. (2018). Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 351–355.

Schimmel, S. M., Muller, M. F., and Dillier, N. (2009). A fast and accurate "shoebox" room acoustics simulator. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 241–244.

Schissler, C., Mehra, R., and Manocha, D. (2014). High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *ACM Transactions on Graphics (TOG)*, 33(4):1–12.

Schröder, D. (2011). *Physically based real-time auralization of interactive virtual environments*, volume 11. Logos Verlag Berlin GmbH.

Schroeder, M. R. (1979). Integrated-impulse method measuring sound decay without using impulses. *The Journal of the Acoustical Society of America*, 66(2):497–500.

Schroeder, M. R. (1987). Statistical parameters of the frequency response curves of large rooms. *Journal of the Audio Engineering Society*, 35(5):299–306.

Schwartz, O. and Gannot, S. (2013). Speaker tracking using recursive em algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):392–402.

Shabtai, N. R., Zigela, Y., and Rafaely, B. (2009). Estimating the room volume from room impulse response via hypothesis verification approach. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 717–720.

Shih, O. and Rowe, A. (2019). Can a phone hear the shape of a room? In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, pages 277–288.

Shimada, K., Koyama, Y., Takahashi, N., Takahashi, S., and Mitsufuji, Y. (2021). ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 915–919.

Shlomo, T. and Rafaely, B. (2021). Blind amplitude estimation of early room reflections using alternating least squares. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 476–480.

Siltanen, S., Lokki, T., and Savioja, L. (2010). Rays or waves? understanding the strengths and weaknesses of computational room acoustics modeling techniques. In *Proceedings of the International Symposium on Room Acoustics, ISRA*, pages 29–31.

Singh, N., Mentch, J., Ng, J., Beveridge, M., and Drori, I. (2021). Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 286–295.

Sneeuw, N. (1994). Global spherical harmonic analysis by least-squares and numerical quadrature methods in historical perspective. *Geophysical Journal International*, 118(3):707–716.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333.

Sprunck, T., Deleforge, A., Privat, Y., and Foy, C. (2022). Gridless 3D Recovery of Image Sources from Room Impulse Responses. *IEEE Signal Processing Letters*, 29:2427–2431.

Stiefelhagen, R., Bernardin, K., Bowers, R., Rose, R. T., Michel, M., and Garofolo, J. (2008). The CLEAR 2007 evaluation. In *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, pages 3–34.

Su, J., Jin, Z., and Finkelstein, A. (2020). Acoustic matching by embedding impulse responses. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 426–430.

Subramanian, A. S., Weng, C., Watanabe, S., Yu, M., and Yu, D. (2022). Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition. *Computer Speech & Language*, 75:101360.

Suvorov, D., Dong, G., and Zhukov, R. (2018). Deep residual network for sound source localization in the time domain. *arXiv preprint arXiv:1808.06429*.

Svensson, P. and Kristiansen, U. R. (2002). Computational modelling and simulation of acoutic spaces. In *Audio engineering society conference: 22nd international conference: Virtual, synthetic, and entertainment audio*. Audio Engineering Society.

Szöke, I., Skácel, M., Mošner, L., Paliesek, J., and Černockỳ, J. (2019). Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876.

Tang, Z., Aralikatti, R., Ratnarajah, A. J., and Manocha, D. (2022). GWA: A large high-quality acoustic dataset for audio processing. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9.

Tervo, S. and Korhonen, T. (2010). Estimation of reflective surfaces from continuous signals. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 153–156.

Tho, N. T. N., Zhao, S., and Jones, D. L. (2014). Robust DOA estimation of multiple speech sources. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2287–2291.

Tjandra, A., Sakti, S., and Nakamura, S. (2017). Listening while speaking: Speech chain by deep learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 301–308.

Traer, J. and McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865.

Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. (2014). The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506.

Valimaki, V., Franck, A., Ramo, J., Gamper, H., and Savioja, L. (2015). Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments. *IEEE Signal Processing Magazine*, 32(2):92–99.

Varanasi, V., Gupta, H., and Hegde, R. M. (2020). A deep learning framework for robust DOA estimation using spherical harmonic decomposition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1248–1259.

Venkatesan, R. and Ganesh, A. B. (2017). Full sound source localization of binaural signals. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 213–217.

Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., and Piazza, F. (2016). A neural network based algorithm for speaker localization in a multi-room environment. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Vincent, E., Virtanen, T., and Gannot, S. (2018). *Audio source separation and speech enhancement*. John Wiley & Sons.

Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J., and Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557.

Wabnitz, A., Epain, N., Jin, C., and Van Schaik, A. (2010). Room acoustics simulation for multichannel microphone arrays. In *Proceedings of the International Symposium on Room Acoustics*, pages 1–6.

Wallach, H. (1939). On sound localization. *The Journal of the Acoustical Society of America*, 10(4):270–274.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1973). The precedence effect in sound localization (tutorial reprint). *Journal of the audio engineering society*, 21(10):817–826.

Wang, Z.-Q., Zhang, X., and Wang, D. (2018). Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):178–188.

Wen, J. Y., Habets, E. A., and Naylor, P. A. (2008). Blind estimation of reverberation time based on the distribution of signal decay rates. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 329–332.

Woodruff, J. and Wang, D. (2012). Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1503–1512.

Wu, Y., Ayyalasomayajula, R., Bianco, M. J., Bharadia, D., and Gerstoft, P. (2021). Sslide: Sound source localization for indoors based on deep learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4680–4684.

Xiong, F., Goetze, S., Kollmeier, B., and Meyer, B. T. (2018). Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):255–267.

Yalta, N., Nakadai, K., and Ogata, T. (2017). Sound source localization using deep learning models. *Journal of Robotics and Mechatronics*, 29(1):37–48.

Yang, J. (2021). *Audio-Facilitated Human Interaction with the Environment: Advancements in Audio Augmented Reality and Auditory Notification Delivery*. PhD thesis, ETH Zurich.

Yang, J., Barde, A., and Billinghurst, M. (2022). Audio Augmented Reality: A Systematic Review of Technologies, Applications, and Future Research Directions. *Journal of the Audio Engineering Society*, 70(10):788–809.

Yiwere, M. and Rhee, E. J. (2017). Distance estimation and localization of sound sources in reverberant conditions using deep neural networks. *Int. J. Appl. Eng. Res*, 12(22):12384–12389.

Yu, W. and Kleijn, W. B. (2020). Room acoustical parameter estimation from room impulse responses using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:436–447.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhou, B., Elbadry, M., Gao, R., and Ye, F. (2017). BatMapper: Acoustic sensing based indoor floor plan construction using smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 42–55.

Zhou, K., Hou, Q., Wang, R., and Guo, B. (2008). Real-time kd-tree construction on graphics hardware. *ACM Transactions on Graphics (TOG)*, 27(5):1–11.

Zotter, F. (2009). *Analysis and Synthesis of Sound-Radiation with Spherical Arrays*. na.

Zotter, F. and Frank, M. (2019). *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature.