The Estimation of Acoustic Parameters and Representations based on Room Impulse Responses

Yu, W.

DOI
10.4233/uuid:6923646e-09ea-4e06-9855-d3e487e5585f

Publication date
2024

Document Version
Final published version

Citation (APA)
Yu, W. (2024). *The Estimation of Acoustic Parameters and Representations based on Room Impulse Responses*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:6923646e-09ea-4e06-9855-d3e487e5585f

The Estimation of Acoustic Parameters and Representations based on Room Impulse Responses

The Estimation of Acoustic Parameters and Representations based on Room Impulse Responses

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Tuesday 21 May 2024 at 10:00 o'clock

by

Wangyang YU

Master of Science in Electrical Engineering,
Delft University of Technology, the Netherlands,
born in Dandong, China.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. W. B. Kleijn,	Delft University of Technology, the Netherlands, promotor Victoria University of Wellington, New Zealand
Prof. dr. ir. R. Heusdens,	Delft University of Technology, the Netherlands, promotor Netherlands Defence Academy, Netherlands
Independent members:	
Prof. dr. E. Eisemann,	Delft University of Technology, the Netherlands
Dr. ir. T. van Waterschoot,	Katholieke Universiteit Leuven, Belgium
Prof. dr. S. van de Par,	Carl von Ossietzky University, Germany
Dr. ir. R. C. Hendriks,	Delft University of Technology, the Netherlands
Prof. dr. ir. A. J. van der Veen,	Delft University of Technology, Netherlands, reserve member

Keywords: Room Impulse Responses, Room Acoustic Parameters, Higher Order Ambisonics, Deep Learning

ISBN 978-94-6384-586-1

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

"Research is to see what everybody else has seen and to think what nobody else has thought."

Albert Szent-Györgyi

Contents

Summary

We are surrounded by all kinds of sounds at all times. What we hear varies with the physical environment and our position. Room impulse responses (RIRs) characterize the effect of the environment on a sound produced by a source. A first goal of this dissertation is to analyze RIRs and investigate how to extract environmental information from RIRs. Immersive digital environments, such as virtual reality (VR) and augmented reality (AR), play an increasingly important role in society. Spatial audio, aiming to give listeners a 3D audio experience, is vital to immersive digital environments. Omnidirectional RIRs do not provide explicit spatial information for room acoustics applications. As a result, the description and reproduction of the sound field is of great importance for spatial audio. Specifically, we consider higher order ambisonics, which is the prevalent method to represent the sound field around a listener. The synthesis of ambisonics signals is a second goal of this dissertation.

Our first study applies deep learning based methods, including convolutional neural networks (CNNs) and multilayer perceptrons (MLPs), to estimate room acoustic parameters from omnidirectional RIRs. We estimate the room geometry and reflection coefficients, and we determine the link of the reflection coefficients to the corresponding walls. Different from the state-of-art methods, we only require a single omnidirectional RIR. For simulated environments, the proposed estimation method can achieve an RMSE accuracy of 0.04 m accuracy for each dimension in room geometry estimation and 0.09 accuracy in the reflection coefficients. For real-world environments, the room geometry estimation method achieves an RMSE accuracy of 0.07 m for each dimension.

In our second study, we divide the estimation of room acoustic parameters into a two-step process. The omnidirectional RIRs contain the room acoustic parameters implicitly. We assume the time of arrivals (TOAs) of specular reflections carry this information. We use transformer deep neural networks to estimate the TOAs of the direct path and specular reflections up to the second order. The image source method describes the behavior of the specular reflections. The TOAs of specular reflections (as determined by the image source method) generally do not correspond to the peaks of real-world measured RIRs. Hence we need to estimate the TOAs of specular reflections from omnidirectional RIRs. We then use analytical methods to estimate room acoustic parameters using these TOAs. Similarly to the first study, we only require a single omnidirectional RIR. The room acoustic parameter estimation is based on a symmetry analysis of RIRs. It is robust to erroneous pulses, non-specular reflections, and an unknown time offset. It can be applied to compute

the distance between two parallel walls in any room if there exists at least one connecting wall at 90 degrees. For real-world environments, the proposed method can achieve an accuracy of an average of 0.06 m, 0.07 m, and 0.08 m on each dimension of room geometry, source position, and receiver position, respectively, with a failure rate of 18.5%. Failures can be reduced by repeated measurements.

Our third study focuses on listener perception. We investigate the necessary information to integrate a new sound source in an existing immersive environment to give listeners a plausible spatial audio experience. We consider localization and whether the new source is perceived as part of the existing environment. We assume an auditory-only environment and use ambisonics to reproduce the sound field. We focus on three attributes: the reflection order, the ambisonics order, and the reverberation time. We use listening tests to determine the quantitative relevance of these attributes. We conclude from the listening tests that at least third order ambisonics signals are required and that a finite order of reflections, for example ninth order RIRs, can perform as well as a full RIR. In addition, we find that knowledge of only the correct reverberation time is insufficient.

In our final study, we provide methods to estimate the ambisonics room impulse responses (ARRs) from omnidirectional RIRs using deep neural networks. The mapping from omnidirectional RIRs to ARRs is not always feasible due to the symmetry of RIRs. With a symmetry analysis of RIRs similar to that of the analytical method to estimate room acoustic parameters, we add a weak assumption to make this mapping possible. We assume the existence of at least two perpendicular walls in the environment. The ambisonics representation is then restricted to be one of a finite set, with known transformations between the set entries. We solve the estimation problem using CNNs and multi-task variational autoencoders (VAEs). Our method requires only a single room impulse response and obviates the need for specialized hardware for ambisonics measurement. The proposed method can achieve a signal to distortion ratio of 17.62 dB on estimated first order ARRs and 16.15 dB on estimated third order ARRs.

Samenvatting

We worden voortdurend omringd door allerlei geluiden. Wat we horen varieert met de fysieke omgeving en onze positie. Room impulse responses (RIRs) karakteriseren het effect van de omgeving op een geluid geproduceerd door een bron. Een eerste doel van dit proefschrift is het analyseren van RIR's en onderzoeken hoe je omgevingsinformatie uit RIR's kunt halen. Immersieve digitale omgevingen, zoals virtual reality (VR) en augmented reality (AR), spelen een steeds belangrijkere rol in de maatschappij. Ruimtelijke audio, met als doel luisteraars een 3D-audio-ervaring te geven, is van vitaal belang voor immersieve digitale omgevingen. Omnidirectionele RIR's bieden geen expliciete ruimtelijke informatie voor ruimteakoestiektoepassingen. Daarom is de beschrijving en weergave van het geluidsveld van groot belang voor ruimtelijke audio. Specifiek beschouwen we hogere orde ambisonics, de gangbare methode om het geluidsveld rond een luisteraar weer te geven. De synthese van ambisonicsignalen is een tweede doel van dit proefschrift.

Onze eerste studie past deep learning-gebaseerde methoden toe, waaronder convolutionele neurale netwerken (CNN's) en meerlagige perceptrons (MLP's), om de akoestische parameters van de ruimte te schatten op basis van omnidirectionele RIR's. We schatten de geometrie van de ruimte en de reflectiecoëfficiënten, en we bepalen het verband tussen de reflectiecoëfficiënten en de corresponderende muren. In tegenstelling tot de allernieuwste methoden hebben we slechts één omnidirectionele RIR nodig. Voor gesimuleerde omgevingen kan de voorgestelde schattingsmethode een RMSE-nauwkeurigheid bereiken van 0.04 m voor elke dimensie in de schatting van de ruimtegeometrie en 0.09 nauwkeurigheid in de reflectiecoëfficiënten. Voor echte omgevingen bereikt de methode voor het schatten van de ruimtegeometrie een RMSE-nauwkeurigheid van 0.07 m voor elke dimensie.

In ons tweede onderzoek verdelen we de schatting van de akoestische parameters van de ruimte in twee stappen. De omnidirectionele RIR's bevatten impliciet de akoestische parameters van de ruimte. We nemen aan dat de aankomsttijden (TOA's) van speculaire reflecties deze informatie bevatten. We gebruiken diepe neurale netwerken met transformatoren om de TOA's van het directe pad en speculaire reflecties tot de tweede orde te schatten. De beeldbronmethode beschrijft het gedrag van de speculaire reflecties. De TOA's van speculaire reflecties (zoals bepaald door de beeldbronmethode) komen over het algemeen niet overeen met de pieken van in de echte wereld gemeten RIR's. Daarom moeten we de TOA's van speculaire reflecties schatten op basis van omnidirectionele RIRs. Vervolgens gebruiken we analytische methoden om de akoestische parameters van de ruimte te schatten met behulp van deze TOA's. Net als in de eerste studie hebben we

slechts één omnidirectionele RIR nodig. De schatting van de akoestische parameters voor de ruimte is gebaseerd op een symmetrieanalyse van de RIR's. De schatting is robuust voor foutieve RIR's. Het is robuust voor foutieve pulsen, niet-speculaire reflecties en een onbekende tijdafwijking. De methode kan worden toegepast om de afstand tussen twee parallelle muren in een ruimte te berekenen als er ten minste één verbindende muur op 90 graden bestaat. Voor echte omgevingen kan de voorgestelde methode een nauwkeurigheid bereiken van gemiddeld 0.06 m, 0.07 m en 0.08 m voor elke dimensie van respectievelijk ruimtegeometrie, bronpositie en ontvangerpositie, met een foutpercentage van 18.5%. Fouten kunnen worden verminderd door herhaalde metingen.

Onze derde studie richt zich op de perceptie van de luisteraar. We onderzoeken de informatie die nodig is om een nieuwe geluidsbron te integreren in een bestaande immersieve omgeving om luisteraars een plausibele ruimtelijke audio-ervaring te geven. We kijken naar lokalisatie en of de nieuwe bron wordt waargenomen als onderdeel van de bestaande omgeving. We gaan uit van een omgeving met alleen geluid en gebruiken ambisonics om het geluidsveld te reproduceren. We richten ons op drie attributen: de reflectievolgorde, de ambisonicsvolgorde en de nagalmtijd. We gebruiken luistertests om de kwantitatieve relevantie van deze attributen te bepalen. Uit de luistertests concluderen we dat ten minste derde-orde ambisonicsignalen nodig zijn en dat een eindige orde van reflecties, bijvoorbeeld negende-orde RIRs, net zo goed kan presteren als een volledige RIR. Daarnaast vinden we dat kennis van alleen de juiste nagalmtijd onvoldoende is.

In onze laatste studie bieden we methoden om de ambisonics kamerimpulsresponsies (ARR's) te schatten uit omnidirectionele RIR's met behulp van diepe neurale netwerken. De mapping van omnidirectionele RIR's naar ARR's is niet altijd haalbaar vanwege de symmetrie van RIR's. Met een symmetrieanalyse van RIR's die vergelijkbaar is met die van de analytische methode om akoestische parameters van ruimtes te schatten, voegen we een zwakke aanname toe om deze mapping mogelijk te maken. We gaan uit van het bestaan van ten minste twee loodrechte wanden in de omgeving. De ambisonische representatie is dan beperkt tot een eindige set, met bekende transformaties tussen de items in de set. We lossen het schattingsprobleem op met behulp van CNN's en multitask variationele autoencoders (VAE's). Onze methode vereist slechts een impulsrespons van één kamer en maakt gespecialiseerde hardware voor ambisonicsmetingen overbodig. De voorgestelde methode kan een signaal/vervormingsverhouding bereiken van 17.62 dB voor geschatte ARR's van de eerste orde en 16.15 dB voor geschatte ARR's van de derde orde.

1

Introduction

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

Sir William Lawrence Bragg

We hear all kinds of sounds from the world every day. Even if the source sound is identical, we can hear the difference when we are in different environments. In an outdoor environment, there is usually no reflection, and the sound we hear only depends on its direction and distance. In a church, we can clearly hear the reverberant sound from different directions due to the huge space and hard walls. In a small room, it is not easy to distinguish between the direct sound and the reverberation. This illustrates that the sound we hear carries environmental information. It is interesting and useful to investigate how environmental information is represented in the sound and how sound changes with the environment.

1.1 A brief overview of room acoustics

In this section, we give a brief overview of room acoustics. Specifically, we introduce room impulse responses and different sound field reproduction techniques. Room impulse responses characterize the sound propagation in a room. Room impulse responses can benefit many applications such as speech separation [1] and speech dereverberation [2]. Hence, we aim to study room impulse responses from different aspects in this dissertation. Sound field description and reproduction are essential for immersive audio-visual environments. The description and reproduction of sound field is also a topic in this dissertation. We discuss these two topics separately in this section.

1.1.1 Room impulse responses

Room acoustics describes sound propagation behavior in an enclosed space. When a sound hits a boundary, reflection and absorption happen simultaneously. Upon reflection, both the amplitude and the phase of the sound wave will change. As shown in Figure 1.1, the reflection can be further divided into specular reflection and diffuse reflection. Specular reflection follows the law of reflection where the incidence sound, the reflection sound, and the normal of the surface lie in the same plane, and the incidence angle equals the reflection angle [3]. The reflection coefficient is defined as the portion of the reflected energy to the incident energy. Reflection coefficients are angle and frequency dependent. Diffuse reflection can propagate in any direction from the surface [4], and results from the roughness of the surface or reflecting on the edge of the surface.

The study of room acoustics is of great importance in a variety of fields. Room acoustics design helps people build rooms for various purposes. For example, lecture rooms require clear speech and concert halls may benefit from a long reverberation. Room acoustics is also important in virtual reality and augmented reality, aiming to give users an immersive experience. Room acoustic parameters, such as reverberation time, clarity, and late lateral sound level, can be used to evaluate the nature of the sound field.

Room impulse responses (RIRs) characterize sound propagation from the source to the receiver in an enclosed space. The pulses that are apparent in a RIR signal correspond

1

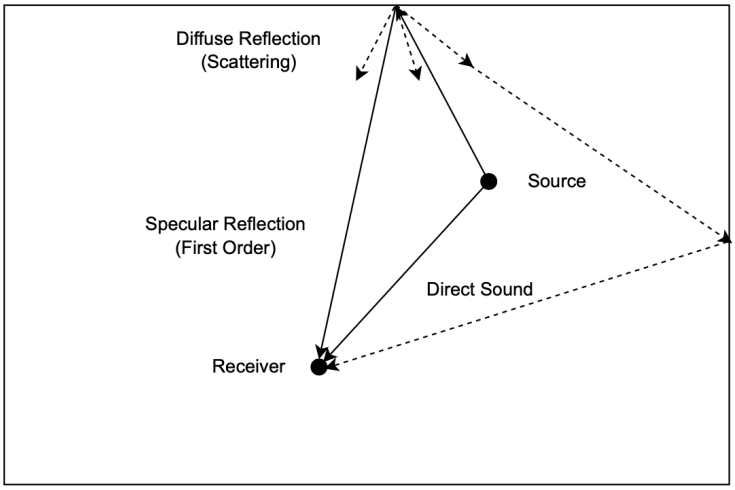


Figure 1.1: A simplified 2D example of sound propagation and reflection in a 2D rectangular room.

to the direct sound and reflections. The room acoustic attributes lie in the RIRs. Room impulse response measurement is usually time-consuming and requires specific hardware and procedures. The basic principle behind a single measurement is that a source emits an excitation signal and a receiver records the received signal simultaneously. The room impulse response can then be estimated by various algorithms, such as deconvolution, cross correlation, or a maximum a posteriori formulation [5]. The commonly used signals include the maximum length sequence, the inverse repeated sequence, the time stretched pulses, the sine sweep, and the choice of the emitted signal depends on the properties of the environment [6]. Real-measured room impulse responses are of limited size and only cover a limited combination of room acoustic parameters. Hence, simulated room impulse responses are also of great importance in room acoustics applications. A range of simulation methods exist [4], including the Finite-Element Method (FEM) [7–9], and the image source method [10–13] .

1.1.2 Sound field description and reproduction

Immersive digital environments, such as virtual reality and augmented reality, are rapidly becoming commonplace. The term *immersive digital environment* refers to an artificial interactive scenario, allowing users to immerse themselves [14]. This new technology has many applications, and relevant technologies are under development. Augmented reality (AR) is a specific immersive audio-visual environment that provides users with

an interactive and enhanced experience in the real world with added artificial objects [15]. It can be used in various applications, such as education and entertainment. One primary attribute of a believable AR system is spatial audio, which aims to create a 3D audio experience. As a consequence, the description and reproduction of the acoustical environment are of great importance. Sound field reproduction aims to minimize the distance between an original sound field and an artificial sound field synthesized with loudspeakers. A good sound field reproduction system should be able to give listeners the same listening experience as the original recording environment. The original sound field is usually represented by the sound pressure in the listening area over time or frequency. An array of loudspeakers is referred to as a set of secondary sources, which are responsible for reproducing the sound field. For the sound field reproduction problem, the target sound field is given over a predefined region, and the signals of the secondary sources need to be determined. In this subsection, we introduce the representation of sound field and sound field rendering methods using loudspeakers.

Higher order ambisonics (HOA) is a commonly used representation of the sound field. HOA is an extension of the original first-order ambisonics system developed by Gerzon [16]. It describes the sound field around the listener by means of a small set of temporal signals. In other words, ambisonics [16–18] takes the listener central view instead of depending on the description of specific sound sources. Ambisonics is based on the spherical harmonic decomposition of the sound field. For a spherical volume, the spherical harmonic expansion of the sound field and the Kirchhoff–Helmholtz integral are equivalent [18]. The order of ambisonics refers to the truncation order of the spherical harmonics expansion. Zero-th order ambisonics representations contain the pressure information, first-order ambisonics contain the acoustic velocity information, and the higher order ambisonics representations include higher order derivatives of the sound field. Ambisonics representations can be recorded by a sound field microphone, for example, a B-format microphone [19]. Recording of higher order ambisonics representations can also use multipole microphones or circular microphone arrays [20]. Ambisonics represents the sound field for the so-called interior case, where all sources lie outside the region of interest. Thus, ambisonics is a particular representation of the interior-case solution to the acoustic wave equation or, equivalently, the Helmholtz equation.

Rendering the ambisonics representations often uses the mode matching method, which aims to match the modes, i.e., spherical wavefunctions [21]. In other words, the driving signals are solved to make sure the expansion coefficients of the spherical wavefunctions of the reproduced sound field are equal to that of the ground truth sound field [18, 22]. It uses a flexible number of secondary sources at a distance. Generally, the secondary sources are assumed to emit plane waves that synthesize the sound field. Ambisonics can only accurately reproduce a limited area. The size of the sweet zone is related to the number of secondary sources, frequency, and ambisonics order. Ambisonics rendering can suffer from truncation and aliasing errors [18]. Ambisonics can be applied to reproduce the sound

field within a reverberant environment by considering image sources of loudspeakers as secondary sources. Still, the implementation requires further research [18]. Near field higher order ambisonics [23, 24] was developed from higher order ambisonics where the secondary sources are assumed to be monopoles at a finite distance. Distance-coding filters are designed in [23] to compensate for the near field effect. An efficient and parametric algorithm to realize a recursive filter is proposed in [24] but only applies to a 2.5-dimensional sound field, where 3D secondary sources are used to reproduce a 2D sound field. Higher order loudspeakers, i.e., compact loudspeaker arrays with different radiation patterns, are employed for the mode-matching method [25]. Higher order loudspeakers can reduce the number of required loudspeakers. A weighted mode matching method is proposed in [21] for higher accuracy, where the sound field reproduction is formulated as an optimization problem, and the driving signal is computed with the spherical wavefunction expansion. The boundary surface control technique can also be used for immersive sound field reproduction [26]. It also requires a large number of loudspeakers and can accurately reproduce the sound field. Experiments show the algorithm performs well on horizontal localization but needs improvement on vertical localization and distance recognition [26].

Wave field synthesis (WFS) is one commonly used sound field method. WFS was first proposed by Berkhout [27, 28] where the sound field generated by a virtual source within a region is reproduced by a large number of loudspeakers placed on the boundary of the listening area. The original sound field can be recorded by source-oriented directive microphones placed on the boundary of the original area [28]. Wave field synthesis is based on the Huygens–Fresnel principle, the Kirchhoff–Helmholtz integral, and the Rayleigh I Integral. The Huygens–Fresnel principle states that any wavefront can be represented as a superposition of elementary spherical waves [29]. The Kirchhoff–Helmholtz integral states that the sound field of a source-free region is known if the sound pressure and velocity on and point of its surface are known [30]. The Rayleigh I Integral states that continuous planar monopole secondary sources can synthesize a source-free sound field driven by any distribution of original sources [31]. Monopole or dipole secondary sources are densely and equally placed on the boundary of the region. The secondary source driving signal depends on the virtual source and the geometry of the sound field reproduction system and can be calculated by the normal derivative of the sound pressure [30]. The virtual source is usually placed behind the loudspeaker arrays. In addition to the virtual source, WFS can also reproduce the sound field generated by the focused source, which refers to the source that is placed between the physical secondary sources and the receiver [32]. The driving signals of both types of sources are computed in the same way.

A benefit of the wave field synthesis is that it can reproduce the sound field correctly over a half space separated by a set of secondary sources [30]. Hence it can be applied to reproduce a spatially large sound field. Since it is impossible to implement an infinite continuous distribution of secondary sources, there will be truncation and spatial aliasing artifacts in practice [30]. The practical disadvantage of wave field synthesis is that it

requires many secondary sources, resulting in expensive computation costs. In [33] and [34], a local wave field synthesis method is proposed that results in improved accuracy in a smaller local region and allows more artifacts outside this region. The accuracy of the synthesized found field depends on the density of the secondary source distribution, i.e., increasing the density of the distribution improves the accuracy [33]. For local wave field synthesis, a limited harmonic order of sound field expansion is used to compute the driving signal in [34]. Multiactuator panels, as a special type of planar loudspeaker array, which can be used as an alternative to classical loudspeaker array, were proposed for wave field synthesis [35]. Multiactuator panels show some advantages, such as being able to integrate into rooms and diffuse radiation, but face structural and geometric issues [35].

In addition to mode matching methods and wave field synthesis, pressure matching methods can also be used for sound field reproduction. With pressure based methods, the signals, positions, and number of the secondary sources are optimized to minimize the sound pressure difference between the reproduced sound field and the ground truth measured by microphones at any number of observation points over the target region. The underlying principle is similar in that it is based on Kirchhoff–Helmholtz Integral and secondary sources are used to approximate the original sound field under a certain criterion. A least-squares criterion to measure the sound pressure difference is commonly adopted. To avoid the unrealizable solutions due to the ill-conditioned matrix of the least squares criterion, regularisation is often used to improve the sound field reproduction performance [36–39]. The methods based on the least-squares criterion are appropriate for low frequencies since in high frequencies, least-squares solutions can result in artifacts due to power leakage, which refers to more power being allocated to closer loudspeakers [40]. In addition, the least-squares solutions allocate powers to all secondary sources instead of a few active secondary sources. For better performance at high frequencies and sparser secondary sources, usage of the least-absolute shrinkage and selection operator (Lasso) was proposed in [40], where the loudspeaker weights are optimized in the frequency domain for each frequency. To avoid the inverse Fourier transform of loudspeaker weights using Lasso, a time-domain iterative mixed-norm constraint optimization algorithm based on group Lasso was proposed in [41]. The method can accurately reproduce the sound field with a few loudspeakers and significantly outperforms the least-square criterion. In addition to the above conventional signal processing based sound field reproduction, deep learning was proposed to solve this problem recently [42] and shows advantages over Lasso-based methods in noise sensitivity and computational speed.

The reverberation of the secondary sources in the physical reproducing environment degrades the sound field reproduction quality. To reduce the interference reverberation, several methods are proposed. Passive compensation refers to the method where the materials of the listening room are changed to compensate for reflections. However, the installation is expensive and impractical for use [43]. In the active compensation method, the room impulse responses between the secondary sources and the receivers

are measured and used to design the filters for the loudspeakers, which demands quite a few processing [43–46]. The active compensation methods require calibration of the loudspeakers [47]. Spherical arrays of fixed-directivity loudspeakers are used in [47] for sound field reproduction in a reverberant environment without compensation. The special setup of the loudspeaker reduces the reverberation within the array compared to the direct sound. It is simpler than the active compensation methods but with a lower reduction of reverberation effect [47].

1.2 Motivation and research objectives

As discussed, room acoustics describes how sounds interact with acoustical environments. We are interested in this interaction. Hence, we need room impulse responses which only change with the positions of sources and receivers in a room. We would like to investigate the underlying information contained in a room impulse response in an enclosed environment. The general goal of this dissertation is the analysis and development of processing methods of room impulse responses. Specifically, we address the following research questions in detail in this dissertation.

Question 1. *How can we extract room acoustic parameters from a room impulse response? Can we analyze it using an analytical method or a deep learning based method? What are the differences between these two kinds of methods?*

We refer to the room geometry, reflection coefficients, and source and receiver positions as room acoustic parameters in the context of this dissertation. When we know these room acoustic parameters, we can synthesize the corresponding room impulse responses using simulation methods. However, it is hard to derive these parameters directly from a room impulse response. Many algorithms [48–62] have been proposed to solve this problem, most are based on using multiple room impulse responses with prior information. However, it is impractical to assume the knowledge of prior information, such as the layout of sources and receivers, or the availability of multiple room impulse responses. Hence, we want to know whether we can extract these room acoustic parameters from a single room impulse response without prior information. We aim to solve this question via two different methods. The first method is to directly apply deep learning method to omnidirectional RIRs to estimate room geometry and reflection coefficients. For the second method, we divide it into a two-step process to investigate how these parameters are extracted. We hypothesize the TOAs are extracted first and the TOAs are used to estimate the room acoustic parameters. We apply a deep learning based method to estimate TOAs from RIRs. We then solve the room acoustic parameter estimation problem via an anlytical method, which helps us to better understand how these room acoustic parameters contain in the structure of RIRs.

Omnidirectional room impulse responses do not always provide enough spatial information explicitly for room acoustics. For a realistic listening experience in room acoustics,

we choose ambisonics as a representation of the sound field as discussed in Section 1.1. We choose ambisonics since it does not depend on the layout of secondary sources and facilitates storage and transmission. In addition, when ambisonics are used for an immersive audio-visual environment, the head rotations are easy to model in the spherical harmonics domain. The combination with ambisonics leads to our second and third research questions for this dissertation.

Question 2. *What attributes are required for a new virtual acoustic source to be consistent with a pre-defined physical context?*

To give listeners a realistic listening experience, a perceptually accurate description of the acoustic scenario is needed. It is useful to determine what precision is needed for a perceptually accurate description. A precise description may require significant latency. However, real-time reproduction also affects the listening experience, especially when visual cues are included. A mismatch is not pleasant for the user. As a consequence, there is a trade-off between the precise description and the real-time reproduction. We aim to determine the balance point, i.e., the necessary attributes to include a new virtual acoustic source in a predefined acoustical scenario for a realistic listening experience.

Question 3. *Is it possible to estimate an ambisonics room impulse response from a single omnidirectional room impulse response?*

Ambisonics room impulse responses contain directional information, which is implicitly contained in omnidirectional room impulse responses if certain conditions are met. As indicated by research question 1, the room acoustic parameters can be estimated from a single omnidirectional room impulse response. This indicates it is also possible to directly estimate ambisonics room impulse responses from omnidirectional room impulse responses using a deep learning based method.

1.3 Contributions and outline of the dissertation

In this section, we describe the outline of this dissertation and summarize the contribution of each chapter.

Chapter 2: Background

This chapter provides essential background information as the basis of this dissertation. To begin with, some basic theory of the sound field is introduced, which is the mathematical and physical fundamental knowledge used in this dissertation. Next, room impulse responses are described in detail since room impulse responses are the focus of the dissertation. Different kinds of room impulse response simulation methods are described. In addition, the state-of-art estimation of room acoustic parameters from room impulse responses is reviewed. After the description of room impulse responses, ambisonics is discussed in detail since ambisonics is another research focus of this dissertation. Higher order ambisonics and corresponding audio rendering methods are presented. In addition,

we define the ambisonics room impulse response and review the algorithms to estimate binaural room impulse responses, providing background information for the following chapters. At the end of Chapter 2, we provide a general overview of deep learning, which is an essential tool in this dissertation to solve our research questions. Specifically, we discuss multilayer perceptrons, convolutional neural networks, and the variational autoencoders.

Chapter 3: Room acoustic parameter estimation from room impulse responses using deep neural networks

This chapter answers another part of the research question 1, i.e., how to estimate room acoustic parameters using deep learning from a single room impulse response. The proposed method utilizes convolutional neural networks to estimate the room geometry and multilayer perceptrons to estimate the reflection coefficients. We do not require knowledge of the relative positions of sources and receivers in the room. The method can be used with only a single RIR between one source and one receiver. We show the new room geometry estimation model performs well with real-world measured RIRs. The deep learning based method is more robust to additive errors, irregular room shapes, and obstacles. In addition, it can be generalized to real-world measurements.

Chapter 4: Estimation of TOAs and Room Acoustic Parameters from an Omnidirectional Room Impulse Response

This chapter partially answers research question 1, i.e., estimating TOAs from a single room impulse response and using analytical methods to estimate room acoustic parameters. We utilize the transformer to estimate the time of arrivals of the direct path and specular reflections up to the second order. The image source method describes the TOAs of specular reflections, which might not correspond to peaks in real RIRs. As a result, we estimate TOAs described by the image source method. The estimated TOAs are used as inputs of the proposed analytical method. The proposed method is based on a symmetry analysis of the room impulse response. We show its robustness to erroneous pulses, non-specular reflections, and an unknown offset. The method can be applied to any room with parallel walls as long as the required arrival times of reflections are available and there exist adjacent walls at an angle of 90 degrees. In contrast to the state-of-the-art method, we do not restrict the location of the source and receiver and the number of room impulse responses. The proposed method achieves about the same accuracy as the method of Chapter 3 for real-measured data. Once the room acoustic parameters are estimated, we can also synthesize ambisonics room impulse responses with these parameters based on the image source method. This chapter also provides an analytical solution to research question 3.

Chapter 5: Necessary attributes for integrating a virtual source in an acoustic scenario

This chapter addresses the answer to research question 2. It investigates how one can integrate a new source into an existing immersive environment with finite information about the environment. We study what is required to integrate a new sound source into an acoustic scene so that people can perceive the new source as a natural component of the acoustic scene and in the correct direction. In this work, we do not consider the head rotation since it can be easily modeled using ambisonics. Through listening tests, we found at least third order ambisonics is required to integrate a new source. In addition, a finite number of early reflections can perform equally well to an entire room impulse response when a new source is added to an existing scenario. However, only a correct reverberation time is not sufficient.

Chapter 6: Ambisonics room impulse response generation from omnidirectional room impulse response using deep neural networks

This chapter provides a solution to research question 3, i.e., the ARR estimation from RIRs using deep neural networks. Generating an ambisonics representation from an omnidirectional signal is not always feasible. We show this mapping is possible in a room. The feasibility relies on the degeneracy of RIRs in a room. Our novel method only requires a single room impulse response without additional information if we only want to estimate ARR and reproduce the immersive environment. Suppose we want to apply the estimated ARR in an audiovisual environment, such as AR. In that case, we need additional information, for example, an image, to determine which mode it belongs to and the alignment between the coordinates of the image and the ARR. Our method is based on the image source method [13], which is sufficient for plausible augmented reality generation.

Chapter 7: Conclusion

In this chapter, we conclude the dissertation and summarise the contributions. In addition, we propose some open questions and give suggestions on solving these questions.

1.4 List of publications

This section lists all the papers during the PhD project, including submitted and published papers.

1

List of Journal Papers

1. **W. Yu** and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2021.

2. **W. Yu** and W. B. Kleijn, "Ambisonics room impulse response generation from omnidirectional room impulse response using deep neural networks," submitted to *Journal of the Audio Engineering Society*.

3. **W. Yu** and W. B. Kleijn, "Estimation of TOAs and Room Acoustic Parameters from an Omnidirectional Room Impulse Response," submitted to *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

List of Conference Papers

1. **W. Yu** and W. B. Kleijn, "Necessary attributes for integrating a virtual source in an acoustic scenario," *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, 2018, pp. 21-25.

2

Background

This chapter aims to provide essential background knowledge of this dissertation, which helps the understanding of the algorithms in the following chapters. This chapter introduces room impulse responses, ambisonics, and deep learning.

To start with, this chapter discusses the fundamental theory of the sound field in Section 2.1. The room impulse response (RIR) is the most important signal that we focus on in this dissertation. We analyze RIRs for estimating room acoustic parameters in Chapter 4, use deep learning based methods to estimate room geometry and reflection coefficients from RIRs in Chapter 3, examine the necessary information of RIRs to integrate a new acoustic source in Chapter 5, and estimate ambisonics room impulse response from RIRs in Chapter 6. Hence, it is of great importance that we describe RIRs as background information and we focus on omnidirectional RIRs in Section 2.2. Ambisonics is another crucial aspect in the dissertation since it is used in Chapter 5 as a rendering tool, and ambisonics room impulse response is our target signal in Chapter 6. We introduce ambisonics in Section 2.3. The estimation of room acoustic parameters in Chapter 3 and ambisonics room impulse response in Chapter 6 both use deep learning, which is discussed in Section 2.4.

2.1 Basic theory of Sound fields

A sound field is a composition of many plane waves of different phases, amplitude, and direction. To describe the sound field physically, let $p(\boldsymbol{x},t)$ denote the sound field where p is the pressure of the sound field and $\boldsymbol{x}$ and t denote the position and the time respectively. We first need to introduce wave equation [31] in a source free area, which is

$$\Delta p(\boldsymbol{x},t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{x},t)}{\partial t^2} = 0, \tag{2.1}$$

where Δ is the Laplacian operator, and c denotes the speed of sound. (2.1) is a linear, lossless wave equation that describes the sound propagation in a quiescent, homogeneous, inviscid, non-heat-conducting, isotropic, perfectly elastic medium [63, 64]. If we apply Fourier transform to (2.1) in time domain, then we obtain the homogeneous Helmholtz equation [10, 31], which is

$$\Delta p(\boldsymbol{x},k) + k^2 p(\boldsymbol{x},k) = 0, \tag{2.2}$$

where $k = \frac{\omega}{c}$ denotes the temporal frequency, ω is frequency in rad/s, and c is the speed of the sound.

For plane waves, the sound pressure is constant on any plane perpendicular to the propagation direction, where wave fronts refer to the planes with constant sound pressure and wave normal refers to the perpendicular lines of the wave fronts [10]. The expression of a plane wave propagating along the x direction at time t can be written as

$$p(x,t,\omega) = \hat{p}\exp\left[i(\omega t - kx)\right], \tag{2.3}$$

where $\hat{p}$ is a constant which denotes the amplitude of the plane wave. In addition to plane waves, the sound field can also be a composition of spherical waves. In contrast, the wave fronts of spherical waves are concentric spheres generated from a point source [10].

Let r denotes the distance, *phi* denotes the direction, and *theta* denotes the elevation, the expression for a spherical wave is

$$p(r,\phi,\theta,t,\omega) = \frac{A}{r}\Gamma(\phi,\theta)\cdot\exp\left[i(\omega t - kx)\right], \tag{2.4}$$

where $\Gamma(\phi,\theta)$ is a normalized directional factor to make its absolute maximum equal to 1, and A is a constant.

There exist several methods to represent the sound field, such as the Kirchhoff–Helmholtz integral and spherical harmonics. The Kirchhoff–Helmholtz integral represents the sound field via the sound pressure and velocity on the boundary. It is the basis of the wave field synthesis [27, 28, 30] and the boundary element method [65, 66], which will be discussed in Section 2.2. Spherical harmonics form a set of orthogonal and complete basis functions. It is the basis of ambisonics, which will be discussed in detail in Section 2.3.

2.2 Room impulse response

A room impulse response models the acoustic environment between a sound source and a receiver in a room. A RIR is composed of a direct signal, early reflections, and late reverberation. Reflections can be divided into specular reflections, which follow the law of reflection, and diffuse reflections, which can propagate in any direction from the surface [4]. RIRs are widely studied in various works, for example, speech dereverberation [67, 68]. In this section, we describe the RIR simulation methods and room acoustical parameter estimation from RIR.

2.2.1 Simulation methods

There exist many algorithms for RIR simulation. The two main categories of methods to simulate RIRs are the wave equation (or wave) based methods and the geometrical acoustics based methods. For each method, we give a brief tutorial to explain how it works and discuss the advantages and disadvantages in this subsection.

Wave based methods

Wave based methods [10, 69, 70] simulate RIRs numerically and accurately. The acoustic space needs to be discretized to solve wave equations using wave based methods. They include the Finite-Element Method (FEM) [7–9], the Boundary-Element Method (BEM) [65, 66], and the Finite-Difference Time Domain (FDTD) [71–74] based methods. Wave based methods can achieve high accuracy. However, these methods require a high computational effort, especially for high frequencies.

The Finite-Element Method The Finite-Element Method [7–9] is a numerical method for room impulse response calculation in an enclosed space [73]. FEM can model RIRs in

complicated room geometries with complex boundary conditions [75]. However, it requires that the size of each element must be much smaller than the size of the wavelength at the solved frequency, which is often set to six to ten elements per wavelength [75, 76]. The FEM method also faces some challenges, such as computational cost and the pollution effect [75]. To use FEM, the acoustic space is first divided into grids with non-overlapping and interconnected elements with nodes. Smaller elements result in smaller errors. We then have a mesh of finite elements with discrete nodes, and the physical quantities are described in each element [75]. The sound pressure can be approximated as

$$p(\boldsymbol{x}) \approx \sum_{j=1}^{n} N_j(\boldsymbol{x}) p_j, \tag{2.5}$$

where n is the number of nodes, N_j and p_j are the shape function and sound pressure of the j-th node respectively. Many element shapes, such as line, triangular, tetrahedral, and pyramidal, can be considered depending on the applications. To reduce the order of equations and include the boundary conditions, a weak variational formulation of the wave equation (2.1) is introduced as

$$\int_{\Omega} [\nabla\omega \cdot \nabla p + \omega(\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2})]\,\mathrm{d}\Omega + \int_{\Gamma_z} (\omega \frac{1}{c\zeta}\frac{\partial p}{\partial t})\,\mathrm{d}\Gamma_z = -\int_{\Gamma_s} (\omega\rho\frac{\partial v_n}{\partial t})\,\mathrm{d}\Gamma_s, \tag{2.6}$$

where Ω is the domain, ω is a continuous and differentiable weighting function referred to as a test function, Γ_s is the vibrating surface of the boundary, Γ_z is the boundary surface with an impedance condition, v_n is the velocity normal to the surface, and ζ is a normalized acoustic impedance. Then (2.6) can be discretized as

$$\mathbf{K}\mathbf{p} + \mathbf{C}\dot{\mathbf{p}} + \mathbf{M}\ddot{\mathbf{p}} = \mathbf{q}, \tag{2.7}$$

where $\dot{}$ and $\ddot{}$ correspond to the first-order and second-order time derivatives, respectively, $\mathbf{q}$ is a source term, writing the shape function N_j in vector form as $\mathbf{N}$, and then $\mathbf{K}$, $\mathbf{C}$, $\mathbf{M}$ are stiffness, damping, and mass matrices over each element and can be computed as

$$\mathbf{K} = \int \nabla\mathbf{N}^T \cdot \nabla\mathbf{N}\,\mathrm{d}V, \tag{2.8}$$

$$\mathbf{C} = \frac{1}{c\zeta}\int \mathbf{N}^T \cdot \mathbf{N}\,\mathrm{d}S, \tag{2.9}$$

$$\mathbf{M} = \frac{1}{c^2}\int \mathbf{N}^T \cdot \mathbf{N}\,\mathrm{d}V. \tag{2.10}$$

The discretized wave equation can be solved in the modal domain, the frequency domain, or the time domain.

The modal domain solutions of the FEM system neglect sources and use eigenvalue analysis [75]. Assuming the sound pressure pressure is time-harmonic, i.e., $p \sim e^{i\omega t}$ and applying the Fourier transform in time to (2.7), the quadratic eigenvalue problem can be formulated as

$$[\mathbf{K} + \lambda\mathbf{C} + \lambda^2\mathbf{M}]\boldsymbol{\phi} = \mathbf{0}, \tag{2.11}$$

where $\lambda = i\omega$ is an eigenvalue and $\boldsymbol{\phi}$ is an eigenmode. The methods to solve eigenvalue problems can be employed to solve (2.11) [75, 77].

The frequency domain solution also assumes time harmonic sound pressure [75, 78]. Combining with (2.7), we obtain

$$\mathbf{A}\mathbf{p}_f = \mathbf{q}_f, \tag{2.12}$$

where $\mathbf{p}_f$ is the sound pressure in the frequency domain, $\mathbf{A} = [\mathbf{K} + i\omega\mathbf{C} - \omega^2\mathbf{M}]$ denotes the global stiffness matrix and the source term in the frequency domain can be defined as

$$\mathbf{q}_f = -i\omega\rho \int \mathbf{N}^T \hat{u}(\omega)\, \mathrm{d}S, \tag{2.13}$$

with $\hat{u}(\omega)$ being the complex valued amplitude of the normal velocity. The frequency domain solution can be obtained by inverting $\mathbf{A}$ for each frequency using solvers, for example, Gaussian elimination [75, 79]. The inversion is computationally expensive at high frequencies for high accuracy. Taking the inverse Fourier transform to get time domain RIRs can result in non-causal RIRs [78].

The time-stepping approach, also known as Newmark-beta method, solves (2.7) in the time domain as [75, 78, 80]

$$[\mathbf{M} + \frac{\Delta t}{2}\mathbf{C} + [\beta(\Delta t)^2]\mathbf{K}]\ddot{\mathbf{p}}_{t+\Delta t} = \mathbf{q}_{t+\Delta t} - \mathbf{C}\mathbf{r}_t - \mathbf{K}\mathbf{s}_t, \tag{2.14}$$

where the sound pressure and its first order derivative with respect to time are

$$\mathbf{p}_{t+\Delta t} = \mathbf{p}_t + (\Delta t)\dot{\mathbf{p}}_t + (\Delta t)^2(\frac{1}{2} - \beta)\ddot{\mathbf{p}}_t + [(\Delta t)^2\beta]\ddot{\mathbf{p}}_{t+\Delta t}, \tag{2.15}$$

$$\dot{\mathbf{p}}_{t+\Delta t} = \dot{\mathbf{p}}_t + (\Delta t)(1-\gamma)\ddot{\mathbf{p}}_t + [(\Delta t)\gamma]\ddot{\mathbf{p}}_{t+\Delta t}, \tag{2.16}$$

with Δt being a time step, γ and β being hyperparameters which control the performance, and

$$\mathbf{r}_t = \dot{\mathbf{p}}_t + \frac{\Delta t}{2}\ddot{\mathbf{p}}_t, \tag{2.17}$$

$$\mathbf{s}_t = \mathbf{p}_t + (\Delta t)\dot{\mathbf{p}}_t + (\frac{1}{2} - \beta)(\Delta t)^2\ddot{\mathbf{p}}_t. \tag{2.18}$$

We can then solve (2.14) for $\ddot{\mathbf{p}}_{t+\Delta t}$ using a direct or iterative method [75, 80–82]. The time domain solution is difficult to use for frequency dependent boundary conditions, and the time step determines the stability [78].

The Boundary-Element Method The Boundary-Element Method [65, 66] is similar to the FEM. The difference lies in that the surface of the boundary of the acoustic space is divided into elements instead of the space, and the elements can either be continuous or discontinuous [73, 83]. BEM can work for both interior problems and exterior problems. Calculating on the boundary instead of the acoustic space changes the problem from 3D to 2D, which results in a smaller boundary element matrix and requires fewer elements for the same accuracy, but complex asymmetry and non-sparsity are introduced in BEM [64, 84, 85].

BEM is based on the Kirchhoff–Helmholtz integral equation [64, 86, 87], which is

$$c(\boldsymbol{x})p(\boldsymbol{x}) = \int_{S_0} [p(\boldsymbol{x}_0)\frac{\partial G(\boldsymbol{x},\boldsymbol{x}_0)}{\partial n} - \frac{\partial p(\boldsymbol{x}_0)}{\partial n}G(\boldsymbol{x},\boldsymbol{x}_0)]\,\mathrm{d}S + \int_V f(\boldsymbol{x}_0)G(\boldsymbol{x},\boldsymbol{x}_0)\,\mathrm{d}V, \tag{2.19}$$

where $\boldsymbol{x}$ is the point in acoustic space V, $\boldsymbol{x}_0$ is the point on the surface S_0, $G(\boldsymbol{x},\boldsymbol{x}_0) = \exp(-jkR)/R$ is the Green's function with $R = |\boldsymbol{x}-\boldsymbol{x}_0|$, and $c(\boldsymbol{x})$ is the solid angle defined as

$$c(\boldsymbol{x}) = \begin{cases} 4\pi & \boldsymbol{x} \in V, \boldsymbol{x} \notin S_0 \\ 4\pi + \int_{S_0} \frac{\partial}{\partial n}(\frac{1}{R})\,\mathrm{d}S & \boldsymbol{x} \in V, \boldsymbol{x} \in S_0 \\ 0 & \boldsymbol{x} \notin V, \boldsymbol{x} \notin S_0 \end{cases}. \tag{2.20}$$

Using the same discretization as (2.5), considering the case where no source is distributed in the acoustic space and the point of the sound field on the surface, (2.19) can be reformulated as a linear system as

$$\mathbf{D_s}\mathbf{p_s} = \mathbf{M_s}\mathbf{v_s}, \tag{2.21}$$

where $\mathbf{p_s}$ and $\mathbf{v_s}$ denote the pressure and normal velocity vectors of the nodes on the surface, $\mathbf{D_s}$ and $\mathbf{M_s}$ denote the dipole and monopole matrix on the surface [64, 88]. The solution can be derived by a non-linear eigenvalue problem [84], matrix inversion [64], and iterative solvers [89, 90] as FEM. BEM also faces a few challenges, such as reliability issues for exterior problems [84] and the existence of singular integrals [85].

The Finite-Difference Time Domain Method Differently from the FEM and BEM, which solve the problems in the spatial domain, the Finite-Difference Time Domain [70–74] based methods solve the locally discretized wave equation in time domain, and the derivatives in the equations are replaced by the finite difference approximation [73]. The main disadvantage of the FDTD method is the dispersion error, which results in a lower traveling speed for high frequency waves [91]. Although a phase error is introduced through discretization, it is an accurate and realizable method for low frequencies in small rooms [73]. The FDTD method can handle complex geometry and several different boundary conditions [92, 93]. The FDTD method is well suited for the parallel computation of the update equation which allows accelerating computation using GPUs [94–97].

There exist a few categories of schemes of FDTD. For example, Yee's staggered scheme [98] was the first proposed FDTD method. However, it requires oversampling for low dispersion error, which results in a high computational load [91]. Implicit schemes, which update the equations of an element by solving a linear equation system at the new time step, can reduce the computational load by using a lower sampling rate but face problems of implementation and boundary conditions in irregular rooms [91]. In the context of this dissertation, we only give a detailed description of one of the most efficient schemes, i.e., the non-staggered compact explicit scheme on a rectilinear stencil [91]. The stencil is defined as the number of neighboring nodes used in the update equations [99]. At a non-staggered grid, variables of an element are colocated at the same place while they are interleaved at a staggered grid [100]. In the explicit schemes, the equations of an element are updated by values of adjacent or several non-adjacent values of previous time steps. A compact scheme refers to the update by only adjacent nodes. Writing the wave equation (2.1) in a 3-D Cartesian coordinate system gives

$$\frac{\partial^2 p}{\partial x^2}+\frac{\partial^2 p}{\partial y^2}+\frac{\partial^2 p}{\partial z^2}=\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}. \tag{2.22}$$

Let l,m,i denote the spatial indexes in x, y, z axis respectively, n denotes the time index, we can define the update variable $p^n_{l,m,i}$ as

$$p^n_{l,m,i}\equiv p(x,y,z,t)|_{x=lX,y=mX,z=iX,t=nT}, \tag{2.23}$$

where X is the grid spacing and $T=1/f_s$ is the time step. Then the explicit compact scheme can be described by [91]

$$\delta t^2 p^n_{l,m,i}=\lambda^2[(\delta x^2+\delta y^2+\delta z^2)+a(\delta x^2\delta y^2+\delta y^2\delta z^2+\delta x^2\delta z^2)+b(\delta x^2\delta y^2\delta z^2)]p^n_{l,m,i}, \tag{2.24}$$

where a and b are two free parameters that determine the characters of the scheme, $\lambda=cT/X$ is the Courant number, and the second-order derivative centered finite-difference operators $\delta t^2,\delta x^2,\delta y^2,\delta z^2$ are defined by

$$\delta t^2 p^n_{l,m,i}\equiv p^{n+1}_{l,m,i}-2p^n_{l,m,i}+p^{n-1}_{l,m,i}, \tag{2.25}$$

$$\delta x^2 p^n_{l,m,i}\equiv p^n_{l+1,m,i}-2p^n_{l,m,i}+p^n_{l-1,m,i}, \tag{2.26}$$

$$\delta y^2 p^n_{l,m,i}\equiv p^n_{l,m+1,i}-2p^n_{l,m,i}+p^n_{l,m-1,i}, \tag{2.27}$$

$$\delta z^2 p^n_{l,m,i}\equiv p^n_{l,m,i+1}-2p^n_{l,m,i}+p^n_{l,m,i-1}. \tag{2.28}$$

To ensure the numerical stability of the compact explicit scheme, the Courant number and the free parameters need to satisfy [91]

$$\lambda^2\leq\min(1,\frac{1}{2-4a},\frac{1}{3-12a+16b}), \tag{2.29}$$

$$a \leq \frac{1}{2}, \quad b \geq \frac{1}{16}(12a - 3). \tag{2.30}$$

For further details of a list of schemes, coefficients, and properties can refer to [91].

Summary Among the wave-based methods mentioned above, FEM can model physical phenomena, for example, diffraction, but when applied to room acoustics modeling, the resolution is insufficient. BEM is preferred to model exterior problems but does not perform well for complex, inhomogeneous acoustic space [9]. Compared to FEM and BEM, FDTD is more widely applied to room acoustics due to its modeling resolution and computational efficiency.

Since wave-based methods are generally computationally expensive, they are not suitable to generate a large scale database that can be used to train a neural network. In Chapter 3, we use FDTD to generate a small database since it is more accurate than geometrical acoustics based methods. This small database is a good alternative between geometrical acoustics simulated RIRs and real RIRs, which can address the problem of limited available real RIRs. Consequently, we can apply transfer learning with the wave based simulated RIRs to our trained model.

Geometrical acoustics based methods

Geometrical acoustics based methods [4, 10] assume that the sound propagates in straight lines. Wave based methods can model accurately for low frequencies but suffer from high computational load. As a result, geometrical acoustics based methods are of great importance in modeling. The most commonly used geometrical acoustical methods can be classified into the image source method (ISM) [4, 10–13], the ray tracing method [101–105] and the beam tracing method [4, 106–115]. Unlike wave equation based methods, they are unable to simulate some low frequency effects such as diffraction. They need separate methods for diffraction simulation.

The Image Source Method Among the above mentioned methods, we highlight the image source method [10–13] since this is the technique we use most in the context of this thesis for simulated RIRs. It was first proposed by Allen and Berkley [13] in 1979. The image source method can model the TOAs of the direct path and specular reflections well. In addition, it is computationally efficient, making it suitable for generating a large scale database. However, RIRs simulated by the image source method differ from real measured RIRs in several aspects. Firstly, the image source method cannot model frequency dependent components, for example, frequency dependent reflection coefficients. Secondly, the image source method can not be used for curved and non-smooth reflective surfaces and can not model diffraction or scattering. It is only able to model specular reflections accurately. Lastly, empty rectangular rooms are always assumed, although several improved methods

exist that can deal with some irregular shapes. These assumptions make the simulated RIRs far from the real RIRs, indicating that the room acoustics modeling algorithms are still incomplete.

We explain the image source method now. In the image source method, an empty rectangular room is assumed, and non-specular reflections are not considered. In addition, it assumes that sound propagates along straight lines. Each reflection can be modeled as a pressure wave emitted from an image source in free space. We use $\mathbf{p}, \mathbf{m}$ to label each reflection where each element of $\mathbf{p} = (q, j, l)$ can take a value of 0 or 1, indicating the direction of the reflection, and each element of $\mathbf{m} = (m_x, m_y, m_z)$ can take an integer value, indicating the position of the virtual room where image sources locate. In three-dimensional (3D) space, we denote the position of the receiver as (x_r, y_r, z_r) and the position of the source as (x_s, y_s, z_s). Implementing the image source method [76], the image source position can be represented as $(qx_s + 2m_xL_x, jy_s + 2m_yL_y, kz_s + 2m_zL_z)$, where (L_x, L_y, L_z) are the length width and the height of the room. Let $d_{\mathbf{p},\mathbf{m}}$ denote the corresponding path length, then the time delay can be calculated as $\tau_{\mathbf{p},\mathbf{m}} = d_{\mathbf{p},\mathbf{m}}/c$. The amplitude of each reflection is determined by the reflection coefficients $\beta_{x_1}, \beta_{x_2}, \beta_{y_1}, \beta_{y_2}, \beta_{z_1}, \beta_{z_2}$, reflection order $O_{\mathbf{p},\mathbf{m}}$, and image source position. The reflection order $O_{\mathbf{p},\mathbf{m}}$ can be computed as

$$O_{\mathbf{p},\mathbf{m}} = |2m_x - q| + |2m_y - j| + |2m_z - l|. \tag{2.31}$$

If we assume the finite and constant reflection coefficients for each wall, then the RIR can be written as [13]

$$h(t) = \sum_{\mathbf{p},\mathbf{m}} \beta_{x_1}^{|m_x - q|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y - j|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z - l|} \beta_{z_2}^{|m_z|} \frac{\delta(t - \tau_{\mathbf{p},\mathbf{m}})}{4\pi d_{\mathbf{p},\mathbf{m}}}, \tag{2.32}$$

which we will use for ARR computation in Chapter 6.

The Ray Tracing Method The ray tracing method was extended from optical applications to room acoustics in [105]. The basic procedure uses similar principles as the image source method. With the ray tracing method [4, 101–104], the source emits the rays according to a predefined distribution or Monte Carlo simulation, and valid reflected paths are retained. The ray tracing methods face a detection problem and limited spatial resolution [116]. The detection problem originates from the fact that it is impossible for a ray to hit a point receiver. As a result, the ray tracing method assumes a finite-size receiver. Then it may suffer from misidentification of rays or duplicated registered rays [4]. The limited spatial resolution results from the limited number of traceable rays. However, it can handle not only specular reflections but also diffusion reflections.

The Beam Tracing Method There exist two kinds of beam tracing methods, one as an improvement of the ray tracing method and another as an improvement of the image source

method [4]. We discuss these two categories separately. As an improvement of the ray tracing method, the beam tracing method improves the detection problem faced by the ray tracing method. Instead of assuming a finite-size receiver, it assumes volumetric rays [106–112]. As an improvement of the image source method, the beam tracing method focuses on pruning out the invalid image sources in an early stage to minimize the number of beams [113–115] to reduce the computation complexity and facilitate real-time applications.

Summary Among the geometric acoustics based methods, we choose the image source method for a large RIR database generation, as mentioned above. The simulated RIRs might be perceptually acceptable for artificial scenarios in some applications. Still, they are generally far from the real measured RIRs due to the assumptions and approximations for these algorithms. Our primary goal in this dissertation is the estimation of room acoustic parameters and the ambisonics representation from RIRs, where the specular reflections are far more critical than non-linear effects, for example, diffraction [4, 10]. Consequently, we use the image source method to generate a large RIR database since it can accurately model specular reflections and, more importantly, is computationally efficient. We consider scattering and diffraction as distortions to RIRs, and our algorithms aim to eliminate the effect of these distortions.

Additional RIR simulation methods

Since wave based methods and geometrical acoustics methods have their advantages and disadvantages, different methods can also be combined as a hybrid method to simulate RIRs [117–122]. There also exist simulation methods that do not belong to these two categorises. A pole-zero model was used to fit RIRs is presented in [123], which can model the early reflections precisely. Deep learning can also be used for RIR simulation, for example, the generative adversarial network [124].

Scattering and Diffraction

Real-world surfaces are not always smooth, which results in scattering [125]. The scattering coefficient equals the energy ratio of the nonspecular reflection and the total reflection, where a complete diffuse reflection corresponds to scattering coefficient 1 and an ideal complete specular reflection corresponds to scattering coefficient 0 [4, 126–128]. The scattering coefficient increases with the order of reflections [129]. Lambert's cosine law [10] is used to model the diffusion energy distribution where the diffusion reflection energy of each angle is proportional to the cosine of the reflection angle. The diffuse rain algorithm [130] is commonly used to model scattering, where each diffuse reflection emits secondary radiation to the receiver. The acoustic radiance transfer method [131, 132] can also handle diffuse reflections. We include diffusion reflection in Chapter 3 to verify the robustness of our proposed algorithm.

Diffraction happens when the reflection surfaces are of finite size, and the sound reflects on the edge. The edge will become the source of the additional reflected sound, and the sound spreads in all directions [10]. Several methods exist to model diffraction [4]. A time domain model adds diffraction from the edges to specular reflections [133, 134]. High frequency asymptotic methods, such as the geometrical theory of diffraction [135, 136], and uniform theory of diffraction [136–138], were proposed to model diffraction. Another category of methods [139–141] solve the Kirchhoff-Helmholtz Integral Equation based on the Kirchhoff approximation [142] for diffraction. These methods are often combined with geometrical acoustics based methods to model room acoustics.

2.2.2 Room acoustical parameters determination from room impulse response

In this subsection, we first discuss the estimation of TOAs of specular reflections from RIRs. We then discuss the existing work on estimating the room geometry vector. After that, we review a closely related topic, room volume estimation. Next, we review the estimation of reflection coefficients and reverberation time. Finally, we discuss the methods to estimate the positions of sources and receivers.

TOAs of specular reflections estimation from RIRs

RIRs are composed of the pulse of direct path and reflective pulses. The direct path refers to the transfer between the source and the recevier without reflections. The room acoustic parameters estimation relies on the time of arrival (TOA) information of the pulses. Since it is difficult to distinguish individual peaks in late reverberation, we focus on the specular reflections in the early reflection part. Consequently, we review the existing work to estimate TOAs of specular reflections before the room acoustic parameter estimation. However, it is difficult to evaluate TOA estimation algorithms on their own since the ground truth TOAs are usually not available.

The TOAs of specular reflections can be estimated by finding the peaks [143], but with limited accuracy. Two greedy sparse approximation methods, i.e., matching pursuit (MP) and orthogonal matching pursuit (OMP), are compared in [144] in terms of estimating the arrival times and amplitude of reflections. To find reflections throughout RIR $h(t)$ without the effect of their power, the natural exponential decay of the power of RIR is compensated as amplitude compensated RIR $\hat{h}(t) = e^{\beta t} h(t)$ with $\beta \geq 0$. A dictionary $\boldsymbol{\mathcal{D}}$ is defined, which contains all translations of direct pulse, and the amplitude compensated RIR can be written as a linear combination of n elements from $\boldsymbol{\mathcal{D}}$. MP and OMP are two greedy iterative descent approaches to sparse approximation updating the RIRs. The correlation between the direct pulse and the RIR is calculated to estimate the reflections. The difference is that OMP guarantees that the n-th order residual is orthogonal to the n-th order approximation of RIRs. Experiments on simulated RIRs show that OMP outperforms MP in terms of detecting

less spurious and duplicated reflections [144]. A template matching filter technique is proposed in [55] to estimate the TOAs of reflections. The filter is equalized through sliding correlation or matched filter, which equalizes the direct pulse to a single peak. The proposed method can be applied to real measured RIRs and is robust to additive noise [55]. Dynamic time warping within a matching pursuit inspired algorithm upon the pulse of direct path is used in [145] to detect early reflections. The proposed method assumes the room impulse response is composed of a pulse of direct path and a series of time-shifted, warped, attenuated, and low pass filtered direct pulses. Direct sound is first extracted by a short Hann window centered around the sample with the highest energy. An iterative algorithm using dynamic time warping is applied to find reflections. The proposed algorithm is able to detect reflections of changed shape by bounded dynamic time warping to refine the location and duration of pulses. It can also detect overlapped reflections using concatenated direct pulses [145]. It can be applied to real measured RIRs and detect less spurious reflections than the orthogonal matching pursuit based method [144] and matched filtering based methods [55]. Continuous cross-wavelet transform (XWT) [146, 147] can also be applied to detect early reflections. After applying XWT to the RIRs, a watershed segmentation procedure, an algorithm to segment gray-scale images [148], is performed to locate the time and frequency of reflections. Finally, the wavelet transform reconstruction formula is used to reconstruct the time-domain reflections. It performs well with simulated RIRs but not with measured RIRs [146, 147]. A multifractal approach [149, 150] can be applied to detect early reflections, which is based on the fact that the early reflections show some similarity to the direct pulse. The distribution of Hölder's exponent is calculated, and the large values correspond to the presence of early reflections. It can detect early reflections in simulated and real RIRs accurately. However, the proposed method is computationally expensive [149, 150]. Early reflections can also be detected by estimating the excess coefficients within RIR short segments [151]. This is based on the idea that a segment that contains a burst and a segment that does not contain a burst differ significantly in terms of distribution. The coefficients are compared with the threshold, which is chosen based on the segment length and a predefined probability of error detection of the burst. The proposed method can be applied to both simulated and real RIRs, but the application to real measured RIRs needs further improvement [151].

Room geometry estimation

Room geometry is an important room acoustic parameter, and the knowledge of room geometry can benefit many room acoustical applications, for example, source separation. Existing algorithms to estimate room geometry from RIRs all require some prior information, for example, the locations of the sources and the microphones [48–58].

Room geometry estimation often requires multiple RIRs to avoid high order reflections and improve accuracy. Several room geometry estimation methods exist based on Euclidean distance matrices (EDM) formed by the squared distance between the source and receivers

[48, 50, 51]. The room geometry of a convex polyhedral room can be estimated with a few RIRs using the EDM-based method proposed in [48]. To avoid the usage of higher order reflections, it requires a single source and at least four receivers and the knowledge of their pairwise distances. It assumes the availability of all first order reflections. It solves the echo labeling problem, which prunes out the spurious pulses and assigns the echoes to the correct walls. The candidate echo combination is used to augment the matrix, and the valid combination corresponds to an image source if the rank is at most five. The positions of image sources can be calculated once the echoes are correctly labeled. Then the room geometry can be inferred. It can be applied to the real-measured RIRs. Since augmentation of all possible echo combinations to EDM is computationally expensive, an efficient method using a graph theoretical approach is proposed in [51], where the echo combinations are modeled as nodes. The task is to find the maximum independent set in the graph, which refers to a set of vertices without direct interconnection. It can achieve an average of 2.4 cm accuracy with at least two sources and five receivers on shoe-box shaped rooms with image-source method generated RIRs. A greedy subspace method based on EDMs is proposed to find the feasible echo combinations for room geometry localization [50]. It requires a single source and multiple receivers with known locations. It is more computationally efficient compared to [51].

In addition to EDM-based methods, alternative methods exist for room geometry estimation using multiple RIRs. A two-step geometrical method is proposed in [49] to estimate room geometry from simulated RIRs between one sound source and five receivers. It requires knowledge of the locations of the receivers. The time of arrivals (TOAs) of the detected peaks and the positions of the receivers are used to determine the positions of the real source and image sources. The walls are estimated by the positions of the real source and the image sources. It checks whether the reflective point and the source are on the same side of all confirmed walls to determine if the wall physically exists. The method can achieve approximately 1 cm accuracy on four simulated rooms, and the accuracy depends on the positions of the source and receivers. [52] obtains sets of TOAs from RIRs by detecting and labeling peaks in RIR stacks using an image processing method combined with graph theory. These TOAs are used to estimate the receiver position and image receiver positions with knowledge of the array geometry of sources. Finally, the room geometry can be inferred with estimated positions. It can be applied to the image-source method simulated RIRs and the real measured RIRs. A greedy iterative algorithm [53] is proposed to estimate room geometry where the possible wall positions are discretized in grids and iterated to match with TOAs. The method first only considers first-order reflections, and second order reflections are taken into account after the first order reflection position is determined. It assumes multiple sources and receivers. The room geometry and the source/receiver localization problem can be solved together using a general optimization problem [54]. It assumes multiple sources and multiple receivers. Given the source signal, a matched filter can be used to extract TOAs of direct path and reflections. An iterative algorithm

can estimate source/receiver position with estimated TOAs. Then the room geometry can be inferred using an exhaustive search over the discretized grids as [53]. Assuming a single source and an array of receivers with known geometry, a template matching method is used to detect TOAs which are used to compute TDOAs for source localization in [55]. Then each TOA is transformed into an elliptical constraint of the reflectors, and multiple constraints are combined to locate room geometry. The Hough transform is used to improve the robustness to noisy environments.

Room geometry can also be inferred from a single RIR where higher order reflections are necessary. A simulated single-channel RIR is used to estimate room geometry in a rectangular room based on the image source method [56]. It uses a set of time of arrival (TOA) measurements of reflections to estimate 2D room geometry. It assumes that the TOA measurements are labeled with image sources and that RIRs consist of direct sound and first and second-order reflections. Using the coordinates of the reflections, the distance between the source and receiver can be inferred by the TOAs from adjacent directions, and room geometry can be inferred by the TOAs from opposite directions. Another room geometry method based on a single RIR is proposed in [57]. Given the first and second order reflections, the room geometry can be uniquely inferred by matrix analysis. The method can also estimate the source position, assuming co-located omnidirectional source and receiver. Assuming the knowledge of receiver position, the room geometry and source position can be estimated using a single RIR [58]. Using a genetic algorithm, the proposed method minimizes the distance between the TOAs of echoes in simulated and measured RIRs.

A relaxation of room geometry estimation is the room volume estimation problem. Room volume is also important since it affects reverberation time. Room volume estimation is formulated as a classification problem in [152], where room volume is classified into six volume class values. It does not require source-to-receiver distance. It trains Gaussian mixture models, and the classification is based on the maximum likelihood criterion. Seven room acoustic parameters are first extracted from a given RIR and serve as the input of the model, for example, reverberation time. With these parameters, a statistical pattern recognition approach is used for room volume classification. This method can achieve a 0.1% equal error rate (EER) with simulated RIRs and a 19.1% EER with real-measured RIRs. However, this method does not account for the fact that room volume is continuously distributed. Room volume estimation can also be formulated as a regression problem [153]. Room volume is estimated with CNNs from noisy reverberant signal-channel speech signals that are split into frames with a 25% overlap. After training, the estimated volume is within approximately a factor of two to the true volume value.

Reflection coefficients estimation

Reflection coefficients characterize room reverberation effects. However, they are difficult to estimate directly since they are angle and frequency dependent in real acoustical envi-

ronments. Reflection coefficients are usually estimated using wave equation based methods. Assuming the knowledge of room geometry, sound source's information which includes position, strength, phase, and frequency, and a set of signals measured with microphones, [154] formulates the estimation of reflection coefficients as an inverse boundary problem based on the boundary element method. The inverse problem is solved using an iterative algorithm. Reflection coefficients are estimated by comparing the measured data and the finite difference time domain simulation in [155]. The estimation is formulated as an optimization problem, and the adjoint method is used to compute the gradient of the cost function efficiently. It assumes that the room geometry is known and there exist one source and multiple receivers.

Since reverberation time and characterizing room reverberation effects are closely related to reflection coefficients, we briefly discuss work on reverberation time estimation as a relaxation of the reflection estimation problem. The reverberation time, $RT60$, of a room is defined as the time it takes for sound to decay 60 dB. Sabine-Franklin's formula [156] is commonly used to estimate the reverberation time:

$$RT_{60} = \frac{24 \ln 10}{c_{20}} \frac{V}{Sa} \approx 0.1611 \mathrm{sm}^{-1} \frac{V}{Sa}, \tag{2.33}$$

where c_{20} is the speed of the sound in the room for 20 degrees Celsius, V is the room volume, S is the total surface area of the room and a is the average absorption coefficient of room surfaces. From (2.33), we can conclude that reverberation time is related to room geometry and reflection coefficients. Given a RIR, the reverberation time can be directly estimated from the calculated energy decay curve [157, 158].

Positions of sources and receivers estimation

The RIR can also be used to locate sources and receivers [54, 57, 59–62]. We already discussed some methods for localizing sources and receivers in the previous paragraphs [54, 57, 58]. By assuming that both the room geometry and the receiver position are known, [59] locates the source with a single RIR with Euclidean distance matrices. The principle is identical to the method in [48]. One source and ten receivers can be localized [60] using the same principle as [48]. It iterates over the echo combination to find the minimum error between measured data and the data generated from estimated positions. A source localization method is proposed using the parameters extracted from RIRs with an ad-hoc microphone array [61]. The room geometry is assumed to be known, and the positions of receivers can be estimated with [59, 60]. The extracted features are used to fit a TDOA surface and an amplitude surface across the room to locate the source. Then the center of two optimal fitted areas is the estimated source position. Source localization can also base on the FEM with a single receiver. It assumes the knowledge of room geometry and no sparsity in the temporal domain. The spatial sparsity of sources is exploited for source localization. A cross-correlation based iterative sensor position algorithm is proposed in

[62] to estimate the positions of an array of receivers. With the estimated receiver position, the proposed method estimate TOAs and DOAs from RIRs using dynamic programming projected phase slope algorithm [159] and multiple signal classification [160], respectively. The source position can then be estimated via triangulation. With the estimated source and receiver position, room geometry can be estimated via random sample consensus [161].

2.3 Ambisonics

A single omnidirectional RIR is not always enough for room acoustic applications since the directional information is only implicitly contained. Ambisonics contain directional information explicitly. Hence it is important that we can create and manipulate ambisonics data. Ambisonics [16–18] has become the de-facto standard representation for AR systems and is particularly suitable for AR systems as head rotations are easily modeled as the rotation of sound fields in the spherical harmonics domain. It describes the sound field by means of a small set of temporal signals. Ambisonics room impulse responses (ARRs) can be used to generate ambisonics signals by convolving with source signals [162, 163]. Recent work on ambisonics often uses higher order ambisonics (HOA), which is an extension of the original first-order ambisonics system developed by Gerzon [16]. HOA is used for spatial audio encoding, transmission and as a basis for rendering. With an ambisonics representation of sufficient order, a high quality audio rendering system can give listeners a realistic spatial audio experience. To start, we introduce the basic information about ambisonics and the generation of ambisonics data of arbitrary order from a set of mono sound signals. This is the basis of our ARR generation method in Chapter 6. Since the availability of higher order ambisonics is not always possible, being able to upscale ambisonics to higher order ambisonics is important. The proposed method in Chapter 6 can also be viewed as an upscaling method of ambisonics signals. As a reference, we overview the existing algorithms that upscale ambisonics to higher order ambisonics. We then briefly discuss audio rendering systems, which allows us to demonstrate our work in Chapter 5. Next, we describe ambisonics room impulse response, which is our target signal in Chapter 6. Finally, we review the existing work on multi-channel room impulse response estimation from omnidirectional RIR, which is a close topic to Chapter 6.

2.3.1 Introduction to higher order ambisonics

Ambisonics [16–18] describes the 3-D sound field at a receiver's position instead of depending on the description of specific sound sources. Ambisonics represents the sound field for the so-called interior case, where all sources lie outside the region of interest. Thus, ambisonics is a particular representation of the interior-case solution to the acoustic wave equation as (2.1), or, equivalently, the Helmholtz equation as (2.2).

Spherical harmonics [164, 165] form a complete set of orthogonal basis functions defined on the surface of a sphere. We adopt full three dimensional normalization (N3D)

scheme for the relative amplitudes of channels such that the sum of squares of values in each degree equals the number of values in that degree. It is widely used in ambisonics software packages and is characterized in [166] as the most logical normalization scheme for a natural sound field. The corresponding spherical harmonics formulation is

$$Y_n^m(\theta,\phi) = N_n^{|m|} P_n^{|m|}(\sin(\theta)) \begin{cases} \sin(|m|\phi), & \text{for } m < 0 \\ \cos(|m|\phi), & \text{for } m \geq 0 \end{cases}, \tag{2.34}$$

where $Y_n^m(\theta,\phi)$ is the spherical harmonic of order n and degree m with $-n \leq m \leq n$, $P_n^{|m|}$ is the associated Legendre function, and $N_n^{|m|}$ is the normalization term. With N3D, we have

$$N_n^m = \sqrt{2n+1}\sqrt{\frac{2-\delta_m}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}, \delta_m = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{if } m \neq 0. \end{cases} \tag{2.35}$$

We now discuss a complete solution to the Helmholtz equation. Then the sound signal p measured at the spherical coordinates $\mathbf{r} = (r,\theta,\phi)$ can be represented as [167]

$$p(\mathbf{r},k) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} i^n j_n(kr) Y_n^m(\theta,\phi) B_n^m(k), \tag{2.36}$$

where $j_n(kr)$ is the spherical Bessel function of the first kind and $B_n^m(k)$ are the *ambisonics coefficients*.

When (2.36) is truncated to a particular N, then the sound field will be accurate within a spherical region near the origin, which is commonly called *sweet zone*. If we truncate equation (2.36) at $n = N$, we can represent the sound field in the sweet zone with $(N+1)^2$ temporal signals. The sweet zone increases in size with N. Its size is inversely proportional to frequency. We denote by D_R^{3D} the dimensionality of three-dimensional ambisonics signals after truncation. The dimensionality is related to N as $D_R^{3D} = (N+1)^2$. Furthermore, let R denote the radius of the sweet zone and f denote the frequency of the signal. Then we have [168]:

$$D_R^{3D} = (\lceil \frac{4\pi R}{\lambda} \rceil + 1)^2 \approx 73 R^2 \frac{f^2}{c^2}. \tag{2.37}$$

We consider far-field sound sources and model the sound pressure $S_q(k)$ from sound source q as a plane wave arriving from an incidence angle (θ_q,ϕ_q). The spherical harmonic expansion of the plane wave transfer function is described as [18]

$$G(\mathbf{r}) = e^{i<\mathbf{k}_q,\mathbf{r}>} = 4\pi \sum_{n=0}^{\infty}\sum_{m=-n}^{n} i^n j_n(kr) Y_n^{m*}(\theta_q,\phi_q) Y_n^m(\theta,\phi). \tag{2.38}$$

Using (2.36) and (2.38), we find the *ambisonics coefficients* from a plane waves

$$B_n^m(k) = \sum_q 4\pi Y_n^{m*}(\theta_q, \phi_q) S_q(k). \tag{2.39}$$

2

In addition to the plane wave, the q-th source can be modeled as a point source at position (r_q, θ_q, ϕ_q) with sound pressure $S_q(k)$. Then the spherical harmonic expansion of the point source transfer function is [31]

$$G(\mathbf{r}|\mathbf{r}_q) = \frac{4^{ik|\mathbf{r}-\mathbf{r}_q|}}{4\pi|\mathbf{r}-\mathbf{r}_q|} = (-i)kh_n^{(2)}(kr_q)j_n(kr)Y_n^{m*}(\theta_q, \phi_q)Y_n^m(\theta, \phi), \text{ for } r_q > r, \tag{2.40}$$

where $h_n^{(2)}(kr_q)$ is the n-th spherical Hankel function of the second kind,

$$h_n^{(2)}(kr_q) = j_n(kr_q) - iy_n(kr_q), \tag{2.41}$$

with $y_n(kr_q)$ the n-th spherical Bessel function of the second kind. It is noteworthy that $h_n^{(1)}(kr) \propto e^{ikr}$ represents an outgoing wave and $h_n^{(2)}(kr) \propto e^{-ikr}$ represents an incoming wave [169]. From the view of the listener, it is an incoming wave, so we should use $h_n^{(2)}(kr_q)$ in the spherical harmonics expansion of the transfer function of the wavefield generated by a point source. Similarly, using (2.36) and (2.40), the *ambisonics coefficients* from the point source can be derived as

$$B_n^m(k) = \sum_q -ikh_n^{(2)}(kr_q)Y_n^{m*}(\theta_q, \phi_q)S_q(k). \tag{2.42}$$

When the sources are modeled as point sources, infinite bass boost problems need to be taken into account, and near field control filters should be applied [17, 24] in the encoding stage. The near-field compensated higher order ambisonics $B_n^m(k)^{\mathrm{NFC}(r_q/c)}$ can be written as

$$B_n^m(k)^{\mathrm{NFC}(r_q/c)} = \frac{1}{F_n^{\mathrm{NFC}(r_q/c)(\omega)}} B_n^m(k), \tag{2.43}$$

where

$$F_n^{\mathrm{NFC}(r_q/c)}(\omega) = i^{-(n+1)} \frac{h_n^{(2)}(kr_q)}{h_0^{(2)}(kr_q)}. \tag{2.44}$$

Temporal ambisonics signals can be obtained by taking the inverse temporal Fourier transform of each $B_n^m(k)$ signal. The temporal first-order ambisonics signals are the well-known B-format signals [16].

2.3.2 Upscaling to higher order ambisonics

The spatial resolution of the represented sound fields depends on the order of ambisonics. Hence, it is of great importance to have higher order ambisonics signals. However, measuring higher order ambisonics is not always feasible, and it has a high demand on the specific hardware, for example, the MH Acoustics Eigenmike spherical microphone array [170] can measure up to third order ambisonics. Upscaling lower order ambisonics to higher order signals is an alternative solution. As shown in (2.37), higher order ambisonics also results in a larger sweet zone.

There exist a number of upscaling methods that utilize the sparsity of the source signal [171–177]. These methods use a sparse plane wave decomposition and an overcomplete basis matrix. The decomposed low order ambisonics is then used to reproduce the sound field with high resolution. This can be implemented either in the frequency domain or time domain. To solve the underdetermined problem in the frequency domain, a short-time Fourier transform is applied to ambisonics, and an iteratively-reweighted least-square (IRLS) algorithm is used to solve a multiple-measurement vector (MMV) convex optimization problem for each frequency bin [171]. The proposed method is robust for multiple sources in reverberant environments. For the time-domain method, the sub-band filtering is applied to lower order ambisonics and solves the problem for each sub-band [172–175]. To solve the MMV convex optimization problem, an IRLS algorithm is used in [172, 173], an order recursive matching pursuit (ORMP) algorithm is used in [174], a sequential matching pursuit (MP) is used in [175]. Non-uniform spatial dictionaries are applied to ambisonics in [173] to increase the spatial resolution in the region of interest. Compared to the frequency domain algorithms [171], the time-domain methods [172, 174, 175] are more computationally efficient and more robust. Generally, the matching pursuit based algorithm [174, 175] outperforms the other compressed sensing techniques [171, 172] in terms of computational efficiency and upscaled ambisonics. To improve the robustness in the presence of noise or reverberation, pre-processing methods based on eigen-value decomposition are proposed in [176, 177] to separate the directional and diffuse parts of the sound field. The diffuse component is extracted by projecting the ambisonics signals into an orthogonal sub-space with respect to the directional component [176, 177].

In addition to the upscaling to higher order ambisonics based on the sparsity assumption, the emphasis operator [178] applied to ambisonics can also increase spatial resolution. It uses Clebsch-Gordan coefficients to emphasize the directionality of the sound field. The emphasis operator can be applied to both time and frequency domain ambisonics. In contrast to the sparsity based methods, which aim at adaptive emphasis, the proposed operator can not only adaptively but also statically emphasize the sound field. In addition, the emphasis operator is computationally efficient [178]. The spatial decomposition method (SDM) [179] can also be employed to increase the spatial resolution and the size of the sweet zone. The proposed method uses the intensity vector to estimate the direction of every sample in the first order impulse response and re-encode each sample at the

estimated direction with higher order ambisonics. Since the SDM-based method makes a simplified assumption on a single direction each time, a dual directional vector and a $2+2$ directional signal estimator are adopted in [180] for a better representation of the transition part between early reflections and late reverberation of RIRs. The directional signal estimator is adapted from the high angular plane wave expansion algorithm in a short time Fourier transform domain to make it work in the time domain, and then four directional signals can be detected each time. The directional signals can then be encoded into higher order ambisonics as SDM based method. Besides the conventional signal processing methods mentioned above, a sequential multi-stage deep neural network is trained to upscale ambisonics [181]. It consists of sequentially stacked DNNs, and each stacked DNN upscales ambisonics by one order. Experimental results prove its ability for improved spatial resolution.

2.3.3 Audio rendering system

A high-quality audio rendering system is a fundamental tool for research in ambisonics. Audio rendering systems aim to give listeners a realistic spatial audio experience and allow us to evaluate and demonstrate our work on spatial audio. Audio rendering can be divided into loudspeaker array rendering and binaural rendering. The term binaural indicates that the rendering system is aimed at headphones.

Audio rendering with loudspeakers aims to reproduce the sound field within a spatial region. A conventional audio rendering system, referred to as mode matching decoding (MMAD), takes the pseudo-inverse of the encoding matrix [182, 183]. However, it requires a regular loudspeaker array covering the full sphere and is suitable for low frequencies. Decoding a N-th order ambisonics, MMAD requires at least $(N+1)^2$ loudspeakers. For an irregular loudspeaker array, the MMAD is unstable and faces the problems such as localization error, energy, altered loudness, and source width fluctuation because of varied decoding energy [182, 184, 185]. Let g_m denote the gain of the m-th loudspeaker with $1 \leq m \leq M$ and $\hat{\boldsymbol{u}}_m$ denote the unitary vector representing the incoming wave direction of the m-th loudspeaker, the energy vector r_E [186] can be defined as

$$r_E \cdot \hat{\boldsymbol{u}}_E = \frac{\sum_{m=1}^{M} g_m^2 \hat{\boldsymbol{u}}_m}{\sum_{m=1}^{M} g_m^2}. \tag{2.45}$$

Max-rE decoders [186, 187] optimize the energy vector of the target sound field, which aims at a larger sweet zone at high frequency. As MMAD, max-rE works well with regular loudspeaker arrays and faces some artifacts with irregular loudspeaker arrays. Since a regular loudspeaker array is not always possible in real applications, a number of decoders

are proposed for irregular loudspeaker arrays. To overcome the energy problem of the irregular loudspeaker array, energy preserving decoding [185] has been proposed, which preserves the decoded energy by removing the singular values of the encoding matrix. It can achieve the same localization accuracy as the basic MMAD but removes the uniform layout constraint of the loudspeaker array with the same number of loudspeakers. All-Round Ambisonic Decoding (AllRAD) [188] decodes ambisonics to an optimal virtual t-design loudspeaker array, which is then mapped to real loudspeakers using vector-base amplitude panning (VBAP). It is more stable, and does not require a uniform loudspeaker array layout or the minimum number of loudspeakers. However, the ambisonics order and loudspeaker array do affect the rendering quality [188]. The decoding is divided into three steps for irregular loudspeaker array in [189]. The regular structure of the array is used to decode the lower order ambisonics. The remaining part is further divided into a symmetric part and an asymmetric part where the symmetric layout is used for decoding based on mixed order ambisonics, and the asymmetric layout is responsible for a larger sweet zone and a stable reconstruction error which depends on the radius. Subjective experiments show it is better than the MMAD since it has a larger sweet zone. A matching projection decoder is proposed for irregular loudspeaker arrays in [190]. This greedy algorithm calculates the projection value of the ambisonics signal and then assigns the maximum projection value to the corresponding loudspeaker until all loudspeakers are assigned with a gain. This method performs better than the MMAD in terms of objective and subjective experiments.

Binaural rendering aims to mimic the listener's auditory system with two ears. A number of techniques can be used for binaural rendering of ambisonics. Perhaps the most common technique is to simulate playback over a given loudspeaker array [191–194], where each virtual loudspeaker signal is filtered with appropriately adjusted head related transfer functions (HRTFs) [195]. The head rotations can be realized either through sound field rotation or continuous-azimuth HRTF format based representation [192]. Diffuse-field equalization, which removes direction-independent components of RIRs in the frequency domain, is applied in [194] to improve the high frequency rendering without additional computational cost. In addition, by transforming HRTFs into the spherical harmonics domain, ambisonics can also be decoded binaurally, where the sound pressure at each ear can be calculated as the sum of the product of the HRTFs and plane waves from all directions in the spherical harmonics domain [196–198]. However, this method requires a large number of HRTFs, which is impractical in many scenarios [194]. An order-dependent compensation filter is proposed in [199] in combination with a tapering window to reduce the coloration due to the truncated order of HRTFs in the spherical harmonics domain.

2.3.4 Ambisonics room impulse response

The ambisonics room impulse response (also referred to as spatial room impulse response) refers to the transfer function between a source and a receiver in a room, the spatial aspects

of which can be captured and measured by spherical microphone arrays. It differs from RIR since it contains directional information. ARRs are essential for sound field analysis and spatial sound reproduction. ARRs can be convolved with signals to generate ambisonics signals and be rendered with various approaches as described above, which are commonly used in immersive audiovisual environments, such as AR.

Definition 1. *When the source signal is an excitation signal, i.e., delta function, the set of $B_n^m(k)$ becomes the ambisonics room response in the frequency domain. Multiplying a frequency-domain source signal with the $B_n^m(k)$ results in the ambisonics representation of the sound field around the receiver.*

2.3.5 Multi-channel room impulse response estimation

Because we are not aware of existing work of ARR estimation, we review the algorithms that estimate multi-channel RIRs from an omnidirectional RIR. These algorithms are similar to the ARR estimation since the underlying spatial information of the input RIR is used and both problems need knowledge of reflections, i.e., positions of image sources from omnidirectional RIRs. The multi-channel RIRs contain the spatial information explicitly and are perceptually important for acoustic environment auralization, although their measurement is time consuming and not always realistic in practice.

An algorithm to estimate an arbitrary number of RIRs from one or two RIRs is proposed in [200]. We only introduce the methods for one input RIR case since it is similar to our problem. It is assumed the direct sound always comes from positive x direction. The first peak is identified as the direct path. The specular reflections and diffuse reflections are separated and processed by different models. The first step is the estimation of the source-receiver distance and the room volume using the diffuse field acoustics and reverberation time using the method in [201]. Then the room geometry is determined using a pre-defined fixed ratio. As the second step, up to four strong peaks are identified as first order specular reflections from floor, ceiling and walls. These specular reflections are used to determine the source and receiver positions using predetermined rules but some values can be set arbitrarily as long as the direct path distance is correct. The detailed description of the rules can be found in [200]. The image source method is then applied to calculate the image source positions and the reflection coefficients. The diffuse reflections are divided into time sections. In each time section, the RIR is modeled as a few point sources scattered around the receiver. An arbitrary number of RIRs can then be calculated using the image source method and the scattered point sources [200]. Experimental results proved the generated RIRs resemble the desired ones. It performs well in the early specular reflections but less efficient or accurate for the complete RIRs. Several approximations exist in the proposed method, such as the ratio of the room edges and the estimation of source receiver position.

As a contrast, binaural room impulse response (BRIR) estimation requires head related transfer functions (HRTFs) of each reflection direction. BRIRs can be estimated from an

omnidirectional RIR and a set of HRTFs [202, 203]. Their algorithm assumes knowledge of geometric information of the room volume, the direction of the direct path, and a pre-processed binaural noise. The RIR is divided into three segments by pre-assigned time slots, i.e., direct sound, early reflections, and diffuse reverberation. The direct sound is filtered by the HRTF of that direction. The early reflections are filtered with HRTFs of the predefined reflection pattern. Binaural diffuse reverberation is estimated in each frequency bin by shaping the envelope of binaural noise. This method is mathematically and conceptually simplified but contains several approximations. For example, the early reflection pattern is predefined and thus approximated. Although the algorithm can produce plausible BRIRs, these approximations result in some perceptual differences, for example, timbre and tone color change [202, 203]. [204, 205] improve the method in [202, 203] to allow changes in different aspects. Using a parameter based description of RIRs, the listener's positions can be changed by modifying the direct path and the early reflections based on an estimated geometric model [204]. Room modifications can be realized in [205] by adjusting diffuse reverberation according to the frequency-dependent reverberation time. Although approximations still exist, these methods can result in plausible BRIRs with low computational effort.

Spatial room impulse responses can be estimated from one monaural RIR using the method in [206], which extends and improves the method in [202]. The monaural and spatial parameters are derived from the input RIR. Firstly, the proposed method detects the amplitude and the TOA of the direct path and early reflections. Six to ten early reflections with highest amplitude are selected. The direction of arrivals (DOAs) of early reflections can be determined by a pseudo-randomized directional distribution or a previously determined DOA pattern or by using the image source method with approximated room geometry. In addition, the standard room acoustic parameters are also calculated, such as reverberation time and clarity. The reflection filters are also derived to adjust the magnitude spectra of early reflections. Next, the reverberation level, describing the level of diffuse field in the early reflection part of RIR, is estimated. The reverberation level is used to ensure the preservation of RIR energy when synthesizing BRIRs with directional and diffuse reflections. Finally, combining the detected early reflections and the image source method, the parameters of an arbitrary position in the room can be calculated, which can then be used to calculate the corresponding SRIR or BRIR. Similarly to the previous mentioned methods, the proposed methods consist of several approximations, for example, the DOAs of early reflections.

As mentioned, ARR estimation is not only a similar problem to BRIR estimation; ARRs can also be used to estimate BRIRs. First-order ARRs (also referred to as B-format RIRs) can be used to model BRIRs with a set of HRTFs [207, 208]. Since ARRs contain the direction of arrival (DOA) information, the DOAs can be estimated, and an appropriate HRTF can be chosen to filter the direct sound. The early reflections are linearly combined to match the spectral and frequency dependent interaural coherence cues of real BRIRs [207]. Due

to the limited spatial resolution of B-format RIRs, directional sharpening is applied in [208] to improve resolution. For directional sharpening, the proposed method estimates and assigns directions for each sample using the pseudo-intensity vector (PIV), which is calculated using the zero-th order and first-order ambisonics signals. The direction with the highest PIV value indicates the source position [209]. A parametric model is proposed in [210] to estimate BRIRs from spherical microphone array measurements, which can be considered as an arbitrary order of ARRs. The measured RIRs are divided into a directional part and a diffuse part. The corresponding descriptive acoustic parameters are stored separately, for example, the time of arrival and the energy decay curve. This method estimates BRIRs based on a parametric description only. Directional parameters are used to describe early reflections, which are modeled by sound field decomposition techniques. Diffuse parameters are used to characterize diffuse components and interaural coherence of late reverberation. The main advantage of the parametric model is that the modification of room acoustic parameters is easy to simulate [210].

2.4 Deep learning

Conventional signal processing uses mathematics and physics to analyze and process the signals. A number of processing techniques are commonly used to deal with distortions in signals, such as filtering and Fourier transform. In acoustics, some effects are hard to be modeled by mathematics or physics, for example, the changed phase of pulses upon reflections in RIRs. Deep learning solves the task from a different perspective. Deep learning learns the underlying common patterns from a large amount of input data and applies the learned pattern to the unlearned data. We have already discussed a few conventional signal processing based room acoustic algorithms in this chapter. We also want to solve our problem using deep learning and compare the deep learning based methods with conventional signal processing methods.

Deep Learning shows good modeling properties for many applications. In general, it requires high computational capacity and the availability of large databases. Different from conventional modeling methods, deep learning uses neural networks to learn a function between the input and output from a large amount of data. Each layer of a neural network can be viewed as a simple function with unknown parameters. Combining multiple layers forms a nonlinear modeling function whose parameters are learned by training with the available dataset.

Before applying deep learning to room acoustic problems, we discuss a few commonly used neural network models in this section. Understanding these neural networks helps us choose the proper ones for our tasks. Specifically, we discuss multilayer perceptrons, convolutional neural networks, variational autoencoders, and transformers, which will be used in the following chapters.

2.4.1 Multilayer Perceptron

Multilayer perceptrons (MLPs), also known as feed-forward neural networks, refer to neural networks that are composed of multiple layers (perceptrons), where each unit in one layer is connected to all units in the previous layer. The perceptron concept was first proposed by Rosenblatt in 1958 [211]. With each layer, an intermediate result is computed as the dot product of the input and the weights and an added bias, which is forwarded to the non-linear activation function. Each perceptron can be written mathematically as

$$y = \varphi(w^T x + b), \tag{2.46}$$

where φ denotes the non-linear activation function, w and b are the weights and bias, and x and y are the input and the output of the perceptron.

Universal approximation theory [212] demonstrates that an MLP with only one hidden layer and an arbitrary continuous sigmoidal nonlinearity can uniformly approximate any continuous function. Although an MLP with only one hidden layer can uniformly approximate any continuous function, the number of neurons has to be exponentially large. It has been proved that considering the expressiveness of an MLP with ReLU activation, depth is more important than width [213]. This motivates us to use MLPs with more hidden layers instead of a wide shallow network. MLPs are relatively straightforward to implement and widely used in a variety of classification and regression problems, e.g., [214–218].

In this dissertation, we use MLPs to estimate reflection coefficients from RIRs in Chapter 3. Reflection coefficients only exist in the amplitudes of reflective pulses. We hypothesize MLPs treat these pulse as features and be able to estimate reflection coefficients.

2.4.2 Convolutional Neural Network

CNNs have been used for various applications and show good modeling ability. CNNs were first proposed by [219] for visual pattern recognition. CNNs are primarily used in computer vision, such as image classification [220–222], and image recognition [223–225]. In addition to image data, CNNs can also analyze videos [226–228]. Until recently, CNNs were not widely used in acoustic signal processing. Recent applications confirm that CNNs have good modeling ability for acoustic problems and can outperform state-of-the-art algorithms in this context. Such applications include audio classification [229–231], speech dereverberation [232–234], speech enhancement [235–237].

Many variations of CNN architectures have been developed, such as LeNet, AlexNet, and VGGNet. LeNet, a classical CNN, was first proposed in the 1990s for handwritten and machine printed character recognition [238]. In 2012, AlexNet was proposed for image classification problems and obtained a considerably lower error rate than the previous state-of-art [239]. This error rate was further reduced with VGGNet by addressing the importance of depth [240]. From these classical CNN architectures, we can learn how to build a convolutional neural network. A CNN commonly consists of several convolutional

layers, each followed by a pooling layer for downsampling, a few dropout layers to prevent overfitting, and several fully connected layers at the end.

The convolution operation is multiplying a filter (also called a kernel) and part of the input data, which is of the same size as the filter, element-wise and adding a bias to the sum of the multiplication result. The filters slide spatially over the entire input data and perform the convolution operation at each position as the output. Each convolution layer has a set of independent filters that take the feature maps (or the actual input signal) as input and produce another set of feature maps as output. Each feature map corresponds to a channel with the number of output channels set by the network designer. Different channels view different aspects of the input feature maps [241]. Each neuron in each feature map only connects to one part of the input feature map and shares the same connection weights as other neurons in the same feature map. CNNs capture the spatial relationships within the input by parameter sharing, i.e., sharing the same connection weights, and sparse connection, i.e., connecting to one part of the input. The feature maps can be down-sampled using pooling operations. Upsampling is achieved with so-called transposed convolutions (also called fractionally strided convolutions) [241], which effectively insert zeros in the feature maps before a convolution operation. The transposed convolution operation can be viewed as exchanging the forward and backward pass of convolution operation.

CNNs can capture and preserve the implicit structure of the input signal. As discussed, the information of room acoustic parameters lies implicitly in the TOA sof reflective pulses in RIRs. As a result, we choose CNNs to explore how room acoustic parameters, specifically, room geometry, contain in RIRs in Chapter 3.

2.4.3 Recurrent Neural Networks

Recurrent neural networks (RNNs) [242–247] are mainly applied to sequential data, for example, numerical time series data of stock. RNNs are widely used in applications such as text generation [248–250] and language modeling [251–253].

Different from MLPs and CNNs, RNNs have cycles and send feedback information to itself [246]. This means they can consider both the input of current step and the input of the previous time step. The process can be formulated mathematically [246, 254]. Let $\boldsymbol{H}_t$, $\boldsymbol{X}_t$, and $\boldsymbol{O}_t$ denote the hidden state, the input, and the output at time t, respectively. The hidden state and the output can be computed as

$$\boldsymbol{H}_t = \phi_h(\boldsymbol{X}_t\boldsymbol{W}_{xh} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hh} + \boldsymbol{b}_h), \tag{2.47}$$

$$\boldsymbol{O}_t = \phi_o(\boldsymbol{H}_t\boldsymbol{W}_{ho} + \boldsymbol{b}_o), \tag{2.48}$$

where $\boldsymbol{W}_{xh}$, $\boldsymbol{W}_{hh}$, $\boldsymbol{W}_{ho}$ denote the weight matrix between the input and the hidden state, the hidden-state-to-hidden-state matrix, and the weight matrix between the hidden state and the output, $\boldsymbol{b}_h$ and $\boldsymbol{b}_o$ denote the bias vector, and $\phi_h(\cdot)$ and $\phi_o(\cdot)$ denote the activation

function of the hidden state and the output [246]. RNNs face the problems of vanishing and exploding gradients [246, 255].

To handle the vanishing gradients problem, the long short term memory units (LSTMs) are proposed, which allow RNN to learn over much more time steps [246]. LSTM incorporates non-linear and data dependent controls, i.e., gate cells, into RNN cells that ensure the gradients do not vanish [255]. For a gate cell in LSTMs, there is an output gate $\boldsymbol{O}_t$ that reads the entries of the cell, an input gate $\boldsymbol{I}_t$ that reads data into the cell, and a forget gate $\boldsymbol{F}_t$ that resets the content of the cell. These gates can be computed as [246]

$$\boldsymbol{O}_t = \sigma((\boldsymbol{X}_t\boldsymbol{W}_{xo} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{ho} + \boldsymbol{b}_o), \tag{2.49}$$

$$\boldsymbol{I}_t = \sigma((\boldsymbol{X}_t\boldsymbol{W}_{xi} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hi} + \boldsymbol{b}_i), \tag{2.50}$$

$$\boldsymbol{F}_t = \sigma((\boldsymbol{X}_t\boldsymbol{W}_{xf} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hf} + \boldsymbol{b}_f), \tag{2.51}$$

where $\boldsymbol{W}$ denotes the weight matrix, $\boldsymbol{b}$ denotes the bias, and $\sigma(\cdot)$ denotes the sigmoid activation function. In addition, there is a candidate memory cell $\tilde{\boldsymbol{C}}_t$ with a tanh activation function, which is defined as

$$\tilde{\boldsymbol{C}}_t = \tanh((\boldsymbol{X}_t\boldsymbol{W}_{xc} + \boldsymbol{H}_{t-1}\boldsymbol{W}_{hc} + \boldsymbol{b}_c). \tag{2.52}$$

The gates together with an old memory content $\boldsymbol{C}_{t-1}$ can control the amount of preserved old memory content for the new memory content $\boldsymbol{C}_t$ as

$$\boldsymbol{C}_t = \boldsymbol{F}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{I}_t \odot \tilde{\boldsymbol{C}}_t, \tag{2.53}$$

where $\odot$ denotes the element-wise matrix multiplication [246]. We then can compute the hidden state as

$$\boldsymbol{H}_t = \boldsymbol{O}_t \odot \tanh(\boldsymbol{C}_t). \tag{2.54}$$

There exist some other varieties of RNN, such as deep recurrent neural networks [256–258] and bidirectional recurrent neural networks [259–261]. However, since this dissertation does not deal with sequential data, we do not review them in detail. Further details can be found in [246, 247, 255].

2.4.4 Residual Network

Residual Networks (ResNets) [262–265] introduce identity mapping layers to solve the problems faced by the deep neural networks such as vanishing and exploding gradients and degrading training accuracy. The output of the identity mapping layer is added to the output of the stacked layers. Let $\mathcal{H}(x)$ denote the desired mapping function by a few stacked layers, and x denotes the input to this set of stacked layers. Assuming the same size of input and output layers, instead of training to fit $\mathcal{H}(x)$, the stacked layers are trained to

fit the residual function $\mathcal{F}(x) := \mathcal{H}(x) - x$ [262]. The residual function is applied to every few stacked layers. If the input and output layers are of different sizes, a projection matrix W_s can be applied to the input x to match dimensions. ResNets make it possible to train a very deep neural network. It is easier to train a residual mapping and gain accuracy with increased depth [262]. ResNets can benefit many computer vision tasks, for example, object detection [266–268].

Some variations of ResNet are proposed for improved performance, and we list two of them here. Stable ResNets [269] is proposed to prevent exploding gradients and ensure the expressivity with increased depth. It is achieved by multiplying layer/depth depending scaling factors with the residual function. The scaling factors include uniform scaling factors with similar magnitudes for all layers and decreasing scaling factors. Experimental results prove that Stable ResNets outperform ResNets, but the selection of scaling factors remains an open problem [269]. ResNeXt is proposed in [270] that combines ResNets and inception models. It follows the split-transform-merge paradigm as the inception model. The outputs of different paths are added together and different paths share the same topology. The residual functions are applied to the inception modules. Increasing the cardinality, i.e., the number of different paths, gains accuracy more efficiently compared to increasing depth or width [270].

The residual connection is used in the transformers, which will be discussed in Section 2.4.6. Since adding residual connections can at least perform the same as the original neural network, we plan to add residual connections to our room acoustic parameter estimation model as future work.

2.4.5 Variational Autoencoder

VAEs [271–274] can be used as generative models or as methods to remove redundancy from an input representation. VAEs can be used for speech enhancement [275, 276], image classification [277, 278], and so on. An autoencoder is a neural network that consists of an encoder that maps the input to a latent representation and a decoder that maps the latent information to an approximation of the input data. It is assumed that the high dimensional data can be embedded in a low dimensional manifold. Ideally, the bottleneck layer (latent space) of an autoencoder describes the data within the manifold and corresponds to an abstract description of the input data without redundancy. A VAE adds noise in the latent layer and assumes the latent distribution approximates normal. Then sampling from the noisy latent distribution can be used to generate new data using the decoder only. Thus, VAEs can be used either to remove redundancy or to generate new data.

There exist several varieties of VAEs [279–281]. β-VAEs [279] introduce a hyperparameter β to allow users to set the trade-off between generative power and reconstruction power. β-VAEs outperform the baseline models in terms of the degree of disentanglement but with sacrificed reconstruction quality. Here, disentangled representation refers to the

latent layer where each neuron only changes with one generative factor [279]. Factor VAEs [282] improve β-VAEs with a better trade-off between generative power and reconstruction power. Factor VAEs encourage the latent representation distribution to be factorial and independent across dimensions. It improves the disentanglement and maintains the reconstruction quality but requires a low total correlation. Based on factor VAEs, relevant factor VAEs [283] are proposed without the requirement of total correlation, which adjusts the weights of disentanglement during training instead of hyperparameter tuning. By a L_0-regularisation which prunes the dimensionality of the latent layer, pruning VAEs [284] promotes disentanglement of the latent representation and figures out the intrinsic dimensionality at the same time. Bounded Information Rate Variational Autoencoders (BIR-VAEs) [285] treat the latent layer as a communication channel and bound its information rate with a pre-defined SNR, which are computationally less expensive and provide a meaningful latent space. Variance constrained VAEs [286] only constrain the variance of the latent layer, which allows a more natural representation of the data. Introspective VAEs (IntroVAEs) [287] differ from the above mentioned VAEs since the encoders are also trained to distinguish between the generated data and real data like generative adversarial networks (GANs). It combines the advantages of VAEs and GANs but does not require a separate discriminator as a hybrid model [288].

We focus on variance constrained autoencoders (VCAEs) [286] to implement our ARR estimation task in Chapter 6 as they are easy to implement and achieve good performance. Although sampling from the latent layer is difficult with VCAEs, we aim to analyze RIRs rather than generate new data from the latent space. We use X, an $\mathbb{R}^d$-valued random variable, to represent the signal where d denotes the length of each signal and $X \sim P_D(x)$, whose distribution is determined by the data. A VCAE [286] is composed of an encoder $Q_{Z|X;\psi}$ and a decoder $P_{X|Z;\eta}$ that are implemented by neural networks with parameters ψ and η respectively. Let Z, an $\mathbb{R}^{d_z}$-valued random variable, represent latent space of dimensionality d_z. The distribution of Z is unknown. VCAEs do not constrain the distribution of z but do constrain the variance of z. The latent space z follows that $z = \mu_\psi(x) + \epsilon$, where $\epsilon \sim P_\epsilon$ is defined by the system designers. The loss function can be written as [286]

$$\max_{\eta,\psi} E_{X\sim P_D} E_{Z\sim Q_{Z|X;\psi}}[\log p_\eta(X|Z)] - \lambda |E_{Z\sim Q_{Z;\psi}}[\|Z - E_{Z\sim Q_{Z;\psi}}[Z]\|_2^2] - v|, \tag{2.55}$$

where v denotes the target total variance, and λ controls the trade-off between the reconstruction performance and the variance of the latent space. A VCAE is similar to a regular autoencoder but with the ability to control the information rate traveling through each neuron in the latent layer. A VCAE can be viewed as a communication channel [289] where the code is given by $\mu_\psi(x)$, the channel is defined by $p_\psi(\epsilon)$, and the output is given by $z = \mu_\psi(x) + \epsilon$. Choosing $p_\psi(\epsilon) = \mathcal{N}(0, \sigma_\epsilon^2 \cdot I_{d_z})$, the upper bound of the information rate can be computed as $I_{\text{bits}} = \frac{d_z}{2} \log_2(\frac{1}{\sigma_\epsilon^2})$ [289].

2.4.6 Transformer

The transformer model is based on a parallel multi-head attention mechanism, dispensing with recurrence and convolutions [290]. It was first proposed for sequence transduction. Transformers can be applied to not only sequences but also other applications, such as image generations [291], image recognition [292–294], and audio classification [295]. The transformer model is good at modeling long input sequences since it models the dependency between elements of different positions. The transformer model allows for more parallelization and performs better than other models for transduction tasks [290]. Transformers are relatively computationally expensive to train.

Self-attention, i.e., intra-attention, is an attention mechanism representing sequences that relate different inputs from a set of inputs [296]. In contrast, the cross attention mechanism [297] combines different input sets where one set provides query and the other set provides key and value [298]. To compute the attention vector, a query vector, a key vector, and a value vector are created. In sequence processing, the computation of the elements, as well as their embedding vectors, depends on their position. The encoded position information is added to the input sequence as an input embedding vector in the transformer [290, 299]. Thus, the attention mechanism can capture the relationship among different positions in sequences without the restriction of sequential position as recurrence.

Attention [296] is a function that maps from the query and key-value pairs to the output. An attention vector computed based on key and query describes the importance of each element of the value to the current output element. One widely used attention mechanism is referred to as scaled dot product attention. To compute a set of queries simultaneously, the queries, keys, and values are packed into matrices as Q, K, V, and the outputs can be computed as [290]

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}V), \tag{2.56}$$

where d_k denotes the dimensions of queries and keys. The matrix computation makes it fast and space-efficient [290].

The transformer utilizes multi-head attention, which learns different representations from different positions in parallel. It outperforms the single attention function and facilitates the exploitation of different relations between elements. The multi-head attention mechanism can be computed as [290]

$$\text{Multihead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_h)W^O, \tag{2.57}$$

where head_i is computed using (2.56) as $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, W denotes the trainable parameter matrices, and h denotes that we perform the attention mechanism h times.

A standard transformer employs the encoder-decoder structure as conventional sequence transduction but uses stacked fully-connected layers [290]. Both encoder and decoder adopt a residual connection and a layer normalization. Each layer of the encoder is composed of two sublayers. One is the multi-head attention mechanism and the other is a fully connected layer. The keys, values, and queries of the attention in the encoder come from the output of the previous encoder layer. The decoder has a similar structure but has an additional sublayer using cross-attention between the encoder and the decoder. The keys, values, and queries of the attention in the decoder come from the output of the previous decoder layer. For the encoder-decoder attention layer, the queries come from the previous decoder layer, while the keys and values come from the output of the encoder [290]. An encoder only transformer, a variety of the standard transformer, is bidirectional encoder representations from transformers (BERT) [300]. Similarly, a multi-layer transformer decoder only can also perform natural language processing tasks, for example, Generative Pre-trained Transformer (GPT) [301]. The transformer model allows for more parallelization for improved efficiency and outperforms other models for transduction tasks [290].

Transformers can model the relationships between all points of a signal in spite of their positions. Different from transformers that treat all points equally, CNNs focus on the local spatial structures of various degrees [302]. TOAs of specular reflections can be considered as a complex function of room acoustic parameters in RIRs. The TOAs of specular reflections are not just local maximal since scattering and phase change can hide the true specular reflections. We hypothesize this information spreads along the entire RIRs. Hence transformers are more appropriate than CNNs. As a consequence, we use transformers to estimate TOAs of specular reflections up to second order in Chapter 4. Since transformers are relatively new comparing to CNNs, it is also interesting to re-formulate the problem in Chapter 3 using transformers and compare the results with CNNs as a future work.

3

Room Acoustical Parameter Estimation from Room Impulse Responses Using Deep Neural Networks

We describe a new method to estimate the geometry of a room and reflection coefficients given room impulse responses. The method utilizes convolutional neural networks to estimate the room geometry and multilayer perceptrons to estimate the reflection coefficients. The mean square error is used as the loss function. In contrast to existing methods, we do not require the knowledge of the relative positions of sources and receivers in the room. The method can be used with only a single RIR between one source and one receiver. For simulated environments, the proposed estimation method can achieve an average of 0.04 *m accuracy for each dimension in room geometry estimation and* 0.09 *accuracy in reflection coefficients. For real-world environments, the room geometry estimation method achieves an accuracy of an average of* 0.065 *m for each dimension.*

This chapter is published as "Room acoustical parameter estimation from room impulse responses using deep neural networks," by W. Yu and W. B. Kleijn, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2021.

3.1 Introduction

Augmented reality (AR) is an immersive audio-visual environment where artificial objects are added to a real-world scenario, providing the user with an enhanced and interactive experience [15]. Augmented reality will play an increasingly important role in numerous contexts, such as education, manufacturing, and archaeology. An accurate description of acoustic environments is essential for generating perceptually acceptable sound in an AR system. Estimating room acoustical parameters forms an important aspect of modeling an acoustic environment accurately. In this chapter, we consider the estimation of the room geometry and reflection coefficients from room impulse responses.

The room impulse response (RIR), the transfer function between the sound source and the listener, characterizes the acoustic environment of a room. It is composed of direct-direction sound, early reflections, and late reverberation. An RIR is affected by the position of the sound source and the receiver, the room geometry, and the reflection coefficients. In the context of this chapter, we consider rectangular rooms and define room geometry to be a three-dimensional vector, which contains the length, width, and height of a room. The room geometry and the reflection coefficients can be used to model and analyze acoustic behavior inside a room via RIRs. We are interested in the estimation of the room acoustical parameters from RIRs.

In this chapter, we use deep learning to solve this estimation problem. In recent years, deep learning has seen a rapid increase in usage as a result of the increased computational power and the availability of large databases. Relevant deep neural networks (DNNs) to our work are feedforward multilayer perceptrons (MLPs) and convolutional neural networks (CNNs). MLPs [303] are composed of fully connected layers and can approximate most mapping functions. This property makes them applicable in various areas, such as ecology [214], chemistry [215], and climate change [216]. CNNs contain a set of generalized filters of different levels to extract features from the signals. CNNs have been used for various applications such as image classification [220–222], and speech recognition [304–306].

We use CNNs for room geometry estimation and MLPs for the estimation of reflection coefficients. CNNs can analyze data with salient spatial structures [307] and we hypothesize that the room geometry defines patterns in RIR signals. Reflection coefficients influence the strength of reflective pulses, which we hypothesize MLPs are able to learn from RIR signals. Due to the limited amount of real-world measured RIRs, we first train the neural network with artificial data. After that, we use transfer learning to make the model work with real-world measured RIRs.

The main contribution of this chapter is the usage of deep neural networks to estimate room acoustical parameters. In contrast to state-of-the-art methods for estimating room acoustical parameters, our method only requires a random RIR between a single sound source and a single receiver in the room without any additional information. The new room geometry estimation model performs well with real-world measured RIRs.

This chapter is organized as follows. In section 3.2, we formulate the estimation problem of the room acoustical parameters. We then describe the solutions of the room geometry estimation problem and the reflection coefficient estimation problem separately in section 3.3 and section 3.4. The experimental results are discussed and analyzed in detail in section 3.5. Finally, we conclude our paper in section 3.6.

3.2 Problem Formulation

In this section, we formulate our problem, i.e., room acoustical parameter estimation from RIRs, and discuss the motivation for using deep neural networks to solve it.

We aim to use deep neural networks to estimate room acoustic parameters separately and blindly from a *single* RIR. Since the room acoustical parameters are described by continuous variables, we formulate the room acoustical parameter estimation problem as a regression problem. We define the input and output pair of the neural network with a random variable pair (X,Y). Specifically, in our problem, X is an $\mathbb{R}^{d_X}$-valued random variable that represents RIRs where d_X denotes the length of each RIR signal vector, and Y is an $\mathbb{R}^{d_Y}$-valued random variable that represents the room acoustical parameters where d_Y denotes the length of each room acoustical parameter vector.

We aim to learn a continuous deterministic function h to predict y from x, where (x,y) is a realisation of the random variable pair (X,Y). Hence, we have $\hat{y} = h(x)$ where $\hat{}$ labels an estimate. To measure the generalisation ability of the learned function h, we use a loss function $l : \hat{y} \times y \to \mathbb{R}_+$. The risk R of the predictor can then be defined as:

$$R = \mathbb{E}[l(h(x), y)], \tag{3.1}$$

where the expectation $\mathbb{E}$ is calculated with respect to the distribution $f_X(x)$ (recall y is a deterministic function of x). As the neural network does not know the distribution $f_X(x)$ of the input data during learning, we approximate the risk R of the predictor with the empirical risk R_{emp} on the training set:

$$R_{\text{emp}} = \frac{1}{m} \sum_{i=1}^{m} l(h(x_i), y_i), \tag{3.2}$$

where m denotes the size of training dataset and each (x_i, y_i) pair is one copy of the realisation $(x,y) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}$ in the training dataset.

As we have mentioned above, the RIR is affected by both room geometry and reflection coefficients. For a given room geometry, reflection coefficients, and source and microphone position, the corresponding RIR can be computed for an empty box-shaped room. However, given an RIR in the real world, we might be not able to determine a set of parameters due to the existence of obstacles, a non-regular room shape, changes in temperature, and measurement noise. As a result, we conclude the relationship between the RIR and

the room acoustical parameter is probabilistic. It is difficult to use conventional signal processing techniques to estimate room geometry and the reflection coefficients since the RIR can not be formulated as an analytical function of the room acoustical parameters. This motivates us to use deep neural networks as a non-linear mapping function to estimate room geometry and reflection coefficients from RIRs.

When we consider the effect of room geometry on RIRs, each geometry corresponds to a characteristic set of arrival times for the pulses. We hypothesize that the kernels of CNNs can extract the arrival-time patterns, where the room geometry information lies. Hence we use CNNs to estimate the room geometry from RIRs.

The effect of the reflection coefficients on RIRs is encoded in the strength of each pulse in the RIRs. It is independent of the time of arrival (TOA) of each pulse. With a multilayer perceptron, these pulses can be treated as features. This motivates us to use MLPs when we estimate reflection coefficients since we assume this information is mainly related to the feature values.

3.3 Room geometry estimation

In this section, we describe room geometry estimation based on convolutional neural networks. We solve the problem first for simulated data and then use transfer learning to solve the problem for real-world data.

In convolutional neural networks (CNNs), the receptive field of each neuron is processed with a set of kernels that do not vary across the input data. For our geometry-estimation problem, this corresponds to assuming that the RIR contains similar structures with respect to room geometry across all delays. In this section, we describe how we use convolutional neural networks to estimate room acoustical parameters. We first describe our base method and how we evaluate the precision of our model. We then propose two methods to improve the accuracy of the base method. Finally, we generalize our method to real-world RIRs.

3.3.1 Baseline method

As our base method, we use CNNs to estimate the room geometry vector from RIRs blindly. We hypothesize room geometry vectors can be estimated from a single random RIR of a room without any additional information. To solve the problem, our neural network has three output nodes for the length, width, and height of a room. We use the time-domain RIR as the input of our regression model without any pre-processing. Since the ordering of the three lengths of the geometry is arbitrary, we re-order the geometry vector in ascending order as a pre-processing step.

We adopt a commonly used CNN architecture as a basis. In this architecture, each convolutional layer is followed by a batch normalization layer [308] and an activation function. Since our input signal is a time-domain signal, we use one-dimensional convolutional layers and one-dimensional batch normalization layers. To keep a balance between the

number of parameters and the modeling ability of neural networks, the neural network consists of eight one-dimensional convolutional layers and three fully connected layers. The number of channels (filters) in the convolutional layers increases with depth while the output dimensionality of the convolutional layers decreases.

In a regression problem, a quadratic loss is commonly used to track the training process and measure the generalization ability. Using this quadratic loss in (6.6), we define the mean square error (MSE) as the empirical risk, which is used as the objective function to train our CNN in order to minimize the squared distance between the estimated room geometry and the true room geometry. We chose the MSE loss since it is relatively sensitive to outliers. The loss function is then defined as

$$l(g, \hat{g}) = \frac{1}{m} \sum_{i=1}^{m} \| g_i - \hat{g}_i \|_2^2, \tag{3.3}$$

where $\| \cdot \|_2^2$ is the l^2-norm, m denotes the size of training dataset, $g \in \mathbb{R}^{m \times 3}$ denotes the true room geometry and $\hat{g} \in \mathbb{R}^{m \times 3}$ denotes the corresponding estimated room geometry.

To characterize the estimation performance of our method, we evaluate bias and variance on the test data. Bias measures the mean deviation of our estimates from the true value and variance measures how much our estimates vary from the mean estimated value. Minimizing the MSE results in a balance between bias and variance since the relationship between MSE, bias and variance can be described as

$$\text{MSE} = \text{Bias}^2 + \text{Variance}. \tag{3.4}$$

Since bias is also a parameter that a neural network tries to learn during the training process, our CNN model should in principle result in an unbiased estimator. For an unbiased estimator, we can increase the precision by averaging over the estimates.

3.3.2 Improved methods

Two methods can be used to improve the accuracy of our baseline method, i.e., the averaging method and the semi-blind estimation method. We describe both methods separately in this subsection.

Multiple RIRs can be used to increase estimation precision by averaging estimates. For each room, we select N random independent RIRs. The method is to average over the N estimates to calculate the final estimate for the room. The variance of the estimator will decrease by averaging over N independent estimates. Although the accuracy is limited by the bias, the estimation precision can be increased.

In addition to the above mentioned averaging method, we can also increase accuracy by adding restrictions when we generate RIRs. When we estimate room geometry from RIRs, the source/receiver position, and reflection coefficients can be considered as nuisance

factors. We want to reduce the effect of nuisance factors in our problem to increase estimation accuracy. It requires more effort and more information to assume knowledge of reflection coefficients or exact source/receiver position. However, we can consider a setup where the relative position between the source and the receiver is fixed without the system knowing the distance or absolute position. We then remove one nuisance factor in RIR generation. By adding such a restriction, we hypothesize the estimation accuracy can be increased compared to blind room geometry estimation.

3.3.3 Generalization to real-world room impulse responses

Our goal is to generalize our method to real-world RIRs. On the one hand, since the amount of available real-world data is insufficient for training, we augment our data by processing our simulated RIRs to make our simulated RIRs close to real-world data. On the other hand, due to the imbalanced amount of simulated database and real database, transfer learning can be applied to improve generalization performance. In this subsection, we will first discuss how we use transfer learning. After that, the data augmentation technique will be covered. Finally, we describe how we apply our method to real-world RIRs.

Transfer learning [309] was proposed to improve the performance of a new task based on prior knowledge from a related trained task. Since we are able to generate a simulated RIR database of sufficient size to cover a wide range of room geometriesfor training, we can first train a neural network with an RIR database generated with the image source method. Then this trained neural network can be used as initialization when we train the neural network with a real RIR database of small size.

Instead of directly using transfer learning for real RIR database from the pre-trained model, which is trained on the ISM generated RIRs, we augment data as a transition stage. Compared to real-world measured RIRs, RIRs that are generated by the ISM lack some distortions, for example, additive environmental noises. Consequently, the neural network, which is trained by simulated RIRs, may adapt to certain features that are obscured to a real-world database and may fail to generalize well to a real RIR database. [310] proposed a simple and computationally cheap method to augment data for speech recognition, where they warp the features, mask blocks of frequency channels, and blocks of time steps. With this simple augmentation method, they could outperform prior work and achieve state-of-art performance. Inspired by this work, we can add some distortions to our simulated RIR as a data augmentation policy. In the following several paragraphs, we will introduce how we augment our data.

In the real world, it is almost impossible to obtain clean RIRs. In rooms and concert halls, a signal to noise ratio (SNR) of an RIR is commonly between 30 and 50 dB [158]. Hence, it is reasonable to include additive noise with an SNR between 30 and 50 dB in the RIR.

Obstacles are quite common in the real world, but we are not aware of an efficient

method to simulate the effect of obstacles. Since we want to apply our model to real-world data, we have to mimic the effect of obstacles in our simulated RIR database. In the context of this paper, we discuss two artificial distortion types and one analytical method to simulate RIRs with obstacles in rectangular rooms. We will discuss these three methods separately.

The first type of artificial distortion to simulate the effect of obstacles is computationally inexpensive although rudimentary. The existence of obstacles will block some reflection paths and add some extra reflection paths. As a consequence, the first method is to randomly add and delete a random number of pulses in each RIR generated by the ISM.

As the second method, we add patterns to the blocked pulses due to the existence of obstacles. This method is also computationally feasible for simulations. Since each RIR can be viewed as a composition of a direct path between each image source and the receiver, the reflective pulse is blocked when the corresponding image source is blocked by the obstacle. This method is not physically correct since it only considers the blocked reflective pulses when their last reflection segment is blocked by the obstacle. Our derived pattern covers a subset of true blocked reflective patterns. To avoid the occlusion effect, we consider 2D non-reflective obstacles to simplify the problem. The blocked area, which is extended to infinity, can be then be defined with the receiver as the vertex and the obstacle as the base. When the shape of the obstacle is a quadrilateral, the blocked area can be considered as a pyramid that extends to infinity. Our task is to determine whether the image source lies inside this extended pyramid. To determine the position of the image source, we calculate the dot product between the normal of each face and the vector between the receiver and the image source position. If the dot products are negative with respect to each face, then the image source is inside this pyramid. The method can be generalized to determine whether the reflective pulse is blocked when the obstacle is any polygon.

As the third method of modeling obstacles, we use a method based on adaptive rectangular decomposition (ARD) to simulate the sound propagation in 3D space with obstacles, which was proposed to model sound propagation in 3D complex environments [70]. This method utilizes the analytical solution of the wave equation in a rectangular domains and an efficient implementation of the discrete cosine transform (DCT) to facilitate computation on a desktop computer. However, it remains a challenge to generate an RIR database of sufficient size to train a neural network with this ARD-based method. As a result, this method is only used as a data augmentation method in the context of this paper. The procedure can be summarised as follows. We approximate each obstacle as a cuboid. Adaptive rectangular decomposition is then utilized to decompose the scene into rectangular partitions. After that, sound propagation can be simulated in each partition with the analytical solution to the wave equation on rectangular domains based on the DCT [10]. For the absorbing boundary, a perfectly matched layer absorber is employed [311]. A finite-difference approximation is used for sound propagation between two neighboring rectangular partitions. The RIRs that are generated with this method provide a useful transitional RIR between

the RIRs generated with the image source method and real measured RIRs.

Our ultimate goal is to make the model work with a real-world RIR database. We first use transfer learning from the ISM generated RIRs to the transitional RIR database, which includes RIRs with noise, RIRs with obstacles generated with the three different methods. We then use transfer learning again from this transitional model with a real RIR database. To make efficient use of the small number of real world RIRs for our experiments, we use cross-validation [312] to train and test room geometry estimation. That is, we first divide the database into distinct parts. Each time, we select one subset as the test dataset and mix the remaining subsets as the train dataset. Finally, we average the test results over the folds of the cross-validation method.

3.4 Room reflection coefficients estimation

We now describe room reflection coefficients estimation. Since databases that contain both RIRs and reflection coefficients are not available, the method will be applied to simulated data only. RIRs are composed of reflective pulses. The strength of reflective pulses depends on reflection coefficients and propagation path length. We hypothesize MLPs are able to learn reflection coefficients from a RIR without any additional information.

We first describe the general estimation procedure and discuss the effect of re-ordered reflection coefficients on estimation accuracy. After that, we discuss the frequency dependency of the reflection coefficients. Finally, we describe how we link the reflection coefficients with the room geometry.

3.4.1 General reflection coefficients estimation

The reflection coefficient is a factor determining the RIR and this factor is encoded in the strength of reflective pulses in an RIR. We hypothesize there exists a continuous mapping function from the RIR signal to the reflection coefficient. Since MLPs can uniformly approximate any continuous function, we use MLPs to estimate reflection coefficients from a random RIR blindly. We use the time-domain RIR as the input of our regression model without any transformation. Similarly to our reflection coefficient estimation problem

In a real-world room, reflection coefficients are different on different walls and can even be different in different areas of a single wall. We will not cover different reflection coefficients on a single wall. Thus, In a rectangular room, we assume there are six reflection coefficients corresponding to the six walls. We re-order the six reflection coefficients in ascending order as a pre-processing step.

Similarly to the room geometry estimation problem, we use the MSE as our objective function to train the model, which is defined as

$$l(c,\hat{c}) = \frac{1}{m}\sum_{i=1}^{m} \parallel c_i - \hat{c}_i \parallel_2^2, \tag{3.5}$$

where $c \in \mathbb{R}^{m\times 6}$ is the true reflection coefficient matrix and the $\hat{c} \in \mathbb{R}^{m\times 6}$ is the estimated output.

We then discuss the effect of ordered reflection coefficients. We aim to verify that our neural network does learn the reflection coefficients from the RIRs and does not just correspond to an ordering of random outputs unrelated to the reflection coefficients. We use $X = [X_1, ..., X_6]$ to denote the six reflection coefficients and $Y = [Y_1, ..., Y_6]$ to denote the target of our neural network, i.e., the six ordered reflection coefficients. The real output of our neural network is denoted by $\hat{Y} = [\hat{Y}_1, ..., \hat{Y}_6]$. In the following we assume that the coefficients each have a uniform distribution, which we will impose in our simulation experiments.

We use $\tilde{Y} = [\tilde{Y}_1, ..., \tilde{Y}_6]$ to denote a set of ordered but unrelated random variables. Thus, distance measures between Y and $\tilde{Y}$ form an upper bound on the expected error of our neural network output: $E[|Y_i - \hat{Y}_i|^2] < E[|Y_i - \tilde{Y}_i|^2]$. $E[|Y_i - \hat{Y}_i|^2]$ will be computed experimentally for each i, which corresponds to the MSE. Our objective here is to compute $E[|Y_i - \tilde{Y}_i|^2]$ theoretically for each i.

We first need to compute the probability density function of Y_i and $\tilde{Y}_i$. Since Y_i and $\tilde{Y}_i$ are the i-th order statistic of $X_1 \cdots, X_6$ respectively, they are identically independent distributed for each i. We assume $X_1, \cdots, X_6$ are iid random variables that follow a standard uniform distribution. We can then compute the probability density function of Y_i and $\tilde{Y}_i$ respectively according to the order statistic [313]. That is, $Y_i \sim \text{Beta}(i, 7-i)$ and $\tilde{Y}_i \sim \text{Beta}(i, 7-i)$, where Beta(.,.) denotes the beta distribution. The Beta distribution is a continuous distribution defined on the range $(0, 1)$ with density

$$f_Y(y) = \frac{1}{\text{B}(i, 7-i)} y^{r-1}(1-y)^{s-1}, \tag{3.6}$$

where B(.,.) is the Beta function. The pdf of $\tilde{Y}_i$, $f_{\tilde{Y}}(y)$, is identical to that of $f_Y(y)$.

With the probability density function of Y_i and $\tilde{Y}_i$, our next step is to compute the probability density function of $Y_i - \tilde{Y}_i$, which is denoted as D_i. Following Theorem 2.1 in [314], if Y_i and $\tilde{Y}_i$ are two independent random variables having support in $(0, 1)$, the pdf of $D_i = Y_i - \tilde{Y}_i$ is defined as

$$f_{D_i}(d) = \begin{cases} \int_0^{1+d} f_Y(t) f_{\tilde{Y}}(t-d) \mathrm{d}t & -1 < d < 0 \\ \int_0^{1-d} f_Y(d+t) f_{\tilde{Y}}(t) \mathrm{d}t & 0 < d < 1 \end{cases}. \tag{3.7}$$

With this pdf, we can compute the second moment of D_i, which corresponds to the expected value of $|Y_i - \tilde{Y}_i|^2$, as

$$E[D_i^2] = \int_{-1}^{1} d^2 f_{D_i}(d) \mathrm{d}d. \tag{3.8}$$

With the above derivation, we are able to calculate the expected value of $|Y_i - \tilde{Y}_i|^2$ for each i. Taking the square root of the expected values, we can compute the expected upper

bound of the root mean square error (RMSE), $\sqrt{E[|Y_i - \tilde{Y}_i|^2]}$, which for the six dimensions is $[0.1750, 0.2259, 0.2474, 0.2474, 0.2259, 0.1750]$.

3.4.2 Frequency dependent reflection coefficients estimation

In this subsection, we discuss the frequency dependency of the reflection coefficients. To define an appropriate model for estimating frequency-dependent reflection coefficients, we must know how reflection coefficients vary with frequency. [315] lists several absorption coefficients in different frequencies. For example, the absorption coefficients of a painted concrete block change from 250 Hz (0.05) to 4000 Hz (0.08), the absorption coefficients of a lightweight drapery change from 125 Hz (0.03) to 250 Hz (0.04), and the absorption coefficients of plaster on lath change 500 Hz (0.06) to 4000 Hz (0.03). As all these examples change only moderately over frequency, we assume a simple model with piecewise constant reflection coefficients.

With the piecewise constant reflection coefficient assumption, we add a preprocessing step to divide the full-band RIR into several frequency bands with bandpass filters so that we can estimate reflection coefficients in different frequency bands. Among different kinds of bandpass filters, Chebyshev filters show a good computational speed although they are not perfect on stop-band attenuation [316]. Consequently, we choose Chebyshev type I filters [317] as our lowpass filter, which can be transformed into a bandpass filter or highpass filter as needed. With this pre-processing process, we will get access to RIRs in different frequency bands. We can then apply the previously discussed estimation methods for each frequency band separately.

3.4.3 Linking reflection coefficients with room geometry

Knowledge of six reflection coefficients only is generally insufficient. In this subsection, we focus on how to link the reflection coefficients with the room geometry. We assume that we already know the room geometry that can be estimated as described in Section 3.3. This linking problem can be solved by two methods, a machine learning based method and a conventional signal processing method.

With the machine learning based method, we build a CNN that takes an RIR signal conditioned on the room geometry as the input. The choice of CNN architecture is based on the logic in Section 3.3, where the conditioning is the only difference. The conditioning is fed into the network twice, at the input layer and at a middle layer. The output is a combination of the room geometry and the corresponding pairs of reflection coefficients. Within each pair, since there does not exist an order between two reflection coefficients, we re-order the two reflection coefficients in ascending order.

With the conventional signal processing method, we use $RT60$ as a bridge. On the one hand, ISO 3382 [318] shows how to measure $RT60$ from the reverberation time $T20$ or $T30$.

We first need to calculate the energy decay curve from the RIR signal. The energy decay curve EDC at time t is defined as [157]

$$EDC(t) = \int_t^{\infty} h^2(\tau)\mathrm{d}\tau, \tag{3.9}$$

where $h(\tau)$ is the room impulse response. The reverberation time $T20$ ($T30$) is defined as the time that the energy decays from -5 dB to -25 (-35) dB, which can be calculated from the energy decay curve. With this, $RT60$ is three times $T20$ or twice $T30$. On the other hand, we can compute $RT60$ with Sabine-Franklin's formula as in (2.33). As what we have mentioned, we can estimate room geometry as in Section 3.3 and estimate reflection coefficients as in Section 3.4.1. Different combinations of room geometry and reflection coefficients result in a different $RT60$. By performning an exhaustive search, we are able to find a combination of room geometry and reflection coefficients that is closest to the correct $RT60$.

3.5 Experimental results and analysis

In this section, we present our experiments. In the first subsection, we describe the setup of our experiments. We describe experiments on room geometry estimation in the second subsection. Finally, we present our experiments on the estimaton of the reflection coefficients.

3.5.1 Experimental setup

In the following, we first discuss the database we used to train and test our model. After that, we describe the configuration of our neural networks and how we train and test them. Finally, we introduce how we use bandpass filters for sub-band RIRs in the frequency-dependent reflection coefficient estimation problem.

Database

As is discussed in Section 3.3.3, a large-scale dataset of good quality is needed to train neural networks. An overview of the database we use is shown in Table 3.1.

Table 3.1: Database description

Dataset	# rooms	# sources	# receivers
Real-world RIRs	9	5	31
Clean RIRs of empty room	400000	1	1
RIRs with noises	200000	1	1
RIRs with the 1st artificial distortion type	200000	1	1
RIRs with the 2nd artificial distortion type	50000	1	1
RIRs generated with the ARD-based analytical method	144	1	1000

We used [319] as our real-world RIR database because it contains a relatively large number of real RIRs, several room types are covered, and the room geometry was measured in each room. This database contains nine distinct rectangular rooms that are not empty. Since we aimed our work at moderate or small rooms, we did not include three large rooms of the database, i.e., one conference room (with geometry $28 \times 11 \times 3$ m) and two lecture rooms (with geometry $20 \times 12 \times 5$ m and $23 \times 17 \times 7$ m). The selected six rooms include one hotel room, one meeting room, three office rooms, and one enclosed staircase. The geometry of these selected rooms varies between $4.4 \times 2.8 \times 2.2$ m and $14.2 \times 6.9 \times 3.6$ m. The corresponding RT30, the time that it takes to decay 30 dB, varies between 0.59 s and 1.85 s. Within each room, an average of 155 RIRs is given between five sources and 31 receivers.

To build an RIR dataset, we used the ISM to simulate RIRs [76]. We refer to this dataset as a clean RIR dataset of empty rooms. The shape of the rooms is rectangular and the rooms are empty. The speed of sound was set to $c = 340$ m/s. The sampling frequency was set to 8000 Hz. The length of each RIR was 4096 because an approximate 0.5 s RIR contains at least the direct path signal and early reflections in an indoor environment. Each dimension of the room geometry, i.e., length × width × height, was assumed to be iid between $6 \times 5 \times 4$ m and $10 \times 8 \times 6$ m. The room geometry range covers moderate and small rooms and is close to the real-world RIR database described above. The reflection coefficients of the walls were simulated as iid between 0 and 1. We randomly placed one source and one receiver in each room and generated the corresponding RIR. We labeled each RIR with room geometry and reflection coefficients. In our experiments, the number of the image-source method simulated RIRs was 400000, which was divided into a training dataset, a validation dataset, and a test dataset with the ratio $7 : 2 : 1$ for the baseline method.

The clean RIR training dataset of empty rooms was randomly divided into two equal parts for RIRs with noise and the first artificial distortion type. With one part, an additive Gaussian white noise was added to each RIR with an SNR uniformly distributed between 30 dB and 50 dB.

With the first artificial distortion of the RIR as defined in Section 3.3.3, a random number (this number was set to be uniformly distributed between 10 and 100) of pulses was added or deleted from the first 0.1 s of the clean RIRs. This choice was motivated by the hypothesis that the early reflection part of RIR provides more information for room geometry estimation than late reverberation.

With the second artificial obstacle pattern as defined in Section 3.3.3, we generated an RIR database of 50000 rooms. For each room, we randomly placed one rectangular obstacle of an arbitrary size inside the room and generated the corresponding RIR. This process was repeated nine times, i.e., there were nine distinct distorted RIRs for each room in this database.

For the RIRs generated with the analytical method based on ARD, due to the restriction of computational cost, we simulated a scenario with one source and 1000 receivers in each

of 144 rooms. We randomly placed one to three obstacles of a random size in each room. We changed the reflection coefficients and geometry of the room. Each combination was denoted as one configuration.

Neural network description

In this subsection, we describe how we train and test our neural networks. In addition, we describe the configuration of our neural networks for different objective functions. We did an ablation study on network architecture and hyperparameter tuning with a grid search as a preliminary experiment for each neural network. The network architecture and hyperparameters below were chosen based on this preliminary experiment with our target database. If some properties of the target database change, we always performed an ablation study on network architecture and hyperparameter tuning with grid search.

We used a GPU node to train our neural network. The output node is the room acoustical parameter of the given room. The network was trained with the Adam optimizer [320], to minimize the training loss. The learning rate of the Adam optimizer was 0.001 and the coefficients used for computing running averages of the gradient and its square were set to be $(0.9, 0.999)$. We iterated for 2000 epochs and recorded the MSE loss for each epoch. To prevent overfitting, early stopping is used as regularisation in our model [321]. Early stopping is performed when the validation performance degrades in 100 successive epochs to guarantee the training performance without overfitting and keep a balance on the computational effort. In each epoch, we set the model on evaluation mode and computed the validation error for early stopping. In addition, mini-batch based training is used to increase computational efficiency [322]. The batch size was set to be 50. After training, we set the model to evaluation mode and computed the RMSE per dimension in the test set.

For geometry estimation, our network architecture and the corresponding parameters are shown in Table 6.1, where b denotes the batch size. First the layer size decreases as the number of channels (feature maps) increases. The features are finally mapped to the geometry with fully connected layers. We use a leaky rectified linear unit (Leaky ReLU) [323] as the activation function. After each convolutional layer, there are always a batch normalization layer and a Leaky ReLU layer [323], which we do not list in the Table 6.1 since the output size does not change. The network contains 4577763 trainable parameters in total.

To estimate six frequency-dependent reflection coefficients, we use a multilayer perceptron regressor with nine hidden layers. The size of each layer was halved with each layer, from 2048 to 8. A rectified linear unit (ReLU) [324] was used as an activation function after each hidden layer.

To link the reflection coefficients to the room geometry, the network is described in Table 3.3, where b denotes the batch size and we omit the batch normalization layer and the Leaky ReLU layer in the table. The conditioning, i.e., the room geometry vector, is concatenated to the RIR at the input layer and to the reshaped output vector before the fully

Table 3.2: Network architecture of room geometry estimation

Operation	Kernel Size	Stride	# Channels	Output Size
Input				$(b, 4096)$
Reshape				$(b, 1, 4096)$
Conv1D	4	4	32	$(b, 32, 1024)$
Conv1D	2	2	32	$(b, 32, 512)$
Conv1D	8	8	128	$(b, 128, 64)$
Conv1D	2	2	128	$(b, 128, 32)$
Conv1D	2	2	512	$(b, 512, 16)$
Conv1D	4	4	512	$(b, 512, 4)$
Conv1D	4	4	1024	$(b, 1024, 1)$
Conv1D	1	1	1024	$(b, 1024, 1)$
Reshape				$(b, 1024)$
Fully connected				$(b, 160)$
Fully connected				$(b, 64)$
Fully connected				$(b, 3)$

connect layers. Each output vector is reshaped to a 3×3 matrix, where the first column is the room geometry vector, each row of the second and the third columns is a pair of reflection coefficients corresponding to that edge.

Sub-band RIRs

When we take frequency dependency into account, we assumed the reflection coefficients are piecewise constant. The order of the Chebyshev type I filter was set to be 10 for a relatively short transition band. The maximum ripple factor was set to be 1 dB. Each full-band RIR was transformed into three signals, a lowpass RIR ($0 - 1000$ Hz), a bandpass RIR ($1000 - 2000$ Hz), and a highpass RIR ($2000 - 4000$ Hz). With this transformation, we were available to four sets of sub-band RIR data. The training and test process, and the network configuration are the same as for the full band RIRs.

3.5.2 Experiments on room geometry estimation

In this subsection, we present experiments on room geometry estimation. We first compare the baseline method and the proposed semi-blind estimation method for simulated data. After that, we discuss experiments for the proposed averaging method. We then compare our proposed method with a reference signal processing based method. Finally, we describe how we generalize our method to real-world RIRs.

As the first experiment of room geometry estimation, we set up the experiments of our baseline method and the proposed semi-blind estimation method for simulated data. For

Table 3.3: Network architecture of linking reflection coefficients to room geometry

Operation	Kernel Size	Stride	# filters	Output Size
Input				$(b, 4099)$
Reshape				$(b, 1, 4099)$
Conv1D	3	3	32	$(b, 32, 1366)$
Conv1D	5	5	32	$(b, 32, 273)$
Conv1D	3	3	128	$(b, 128, 91)$
Conv1D	5	5	128	$(b, 128, 18)$
Conv1D	4	4	512	$(b, 512, 4)$
Conv1D	4	4	512	$(b, 512, 1)$
Conv1D	1	1	1024	$(b, 1024, 1)$
Conv1D	1	1	1024	$(b, 1024, 1)$
Reshape				$(b, 1024)$
Fully connected				$(b, 160)$
Fully connected				$(b, 64)$
Fully connected				$(b, 9)$

Table 3.4: Comparison of base room geometry estimation method and semi-blind room geometry estimation.

Method	Baseline method	Semi-blind method
RMSE (m)	$[0.0497, 0.0398, 0.0249]$	$[0.0180, 0.0181, 0.0167]$
Bias (m)	$[0.0048, -0.0032, -0.0013]$	$[0.0012, -0.0003, -0.0014]$
Variance (m^2)	$[0.0024, 0.0016, 0.0006]$	$[0.0003, 0.0003, 0.0003]$

the semi-blind room geometry estimation, we pre-set a random source-receiver relative position relationship and generated the corresponding RIR dataset, whose only difference with respect to our original RIR dataset was the receiver-source relative position. We compared the performance of these two cases in terms of RMSE, bias, median, and variance per dimension in the test set. We used the mean estimation error to approximate bias. In addition, we plot the error distribution of both methods in Figure 3.1, where the error here refers to the MSE of each room geometry estimation.

We list the RMSE, bias, variance, and median of the base method and the semi-blind method in Table 3.4. A positive sign indicates our prediction is larger than the true geometry value. The RMSE, bias, and variance show different values with respect to length, width, and height because the range on these three elements is different and they are independent of each other. We also performed an experiment with our baseline method to compare the estimation accuracy between rectangular rooms and cube rooms. The RMSE of cube rooms is $[0.0534, 0.0374, 0.0243]$ m, which does not show a difference from rectangular rooms.

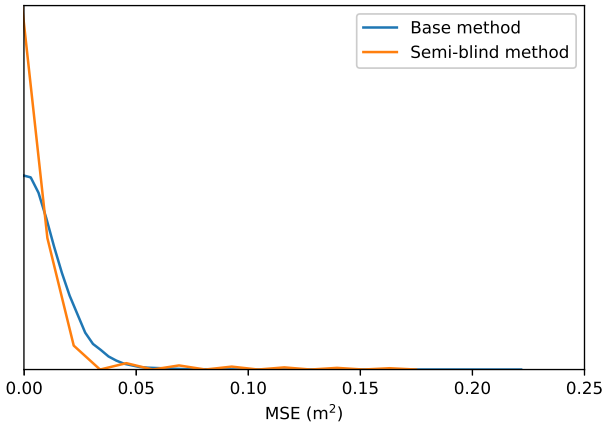


Figure 3.1: MSE distribution of room geometry estimation.

This confirms that the estimation of length, width, and height are independent of each other. As shown in Table 3.4, the small bias vector confirms that our CNN model is not significantly biased after training and the small variance confirms that most estimation errors are relatively small and they do not vary much. The error distribution in the test set of both methods is shown in Figure 3.1. Observing the error distribution in Figure 3.1, the error follows a long-tailed distribution, which confirms that most estimation errors are relatively small, which is consistent with the small variance in the test set. Comparing the experimental results of the baseline method and the semi-blind method, the semi-blind method outperforms the baseline method in terms of accuracy. To conclude, by the addition of a restriction on the relative source-receiver position relationship, the estimation accuracy of room geometry estimation is increased.

The second experiment of room geometry estimation was related to the proposed averaging method to increase the estimation accuracy. We aim to investigate the effect of the number of available RIRs in each room. For this experiment only, we generated a dataset with 16 RIRs per room to do the experiments and the RIRs in this dataset were distinct from those in the training dataset. In each room, 16 RIRs were generated independently, i.e., they correspond to 16 different randomly placed sources and 16 different randomly placed receivers. These RIRs were then use for inference with averaging. We ordered the estimates by the true room geometry and grouped the estimates to one, four, eight, and 16 estimates per room to perform the averaging method. Finally, we computed the RMSE, bias, and variance of the average method.

Table 3.5: Root mean squared error and variance of averaging method.

# RIRs	RMSE (m)	Variance (m^2)
1	$[0.049, 0.039, 0.045]$	$[0.0024, 0.0015, 0.0020]$
4	$[0.027, 0.033, 0.042]$	$[0.0007, 0.0011, 0.0018]$
8	$[0.022, 0.032, 0.040]$	$[0.0005, 0.0010, 0.0016]$
16	$[0.018, 0.031, 0.025]$	$[0.0003, 0.0009, 0.0006]$

Table 3.6: Comparison of proposed method and state-of-art method.

	Proposed method	Method in [51]
Average error (m)	0.0247	0.0235
Average run time (s)	3.22×10^{-4}	2.43

Next we describe the experimental result for the averaging method. The bias of the estimate is $[0.0045, -0.0027, -0.0015]$ m, which does not change by averaging over N estimates. The RMSE, median, and variance under different numbers of RIRs are listed in Table 3.5. The method with one RIR corresponds to our baseline method. The RMSE, bias, and variance are slightly different from the results in Table 3.4 because the test database is not the same. From Table 3.5, we can conclude that, as expected, averaging leads to improved performance. The variance decreases with averaging but does not decrease by a factor of N since there exist nuisance factors, reflection coefficients, and source/receiver positions, which imply that the RIRs in each room are not independently conditioned on room geometry. To conclude, the performance is better when more RIRs are used for averaging although our estimation is still biased.

As the third experiment, we compared our proposed method with the signal processing method proposed in [51] in terms of system requirements, estimation error, and average run time. The experiments are both based on the RIRs generated by the ISM method. For calculating the run time, the experiments were averaged over 600 experiments. The result is shown in Table 3.6. The method in [51] uses five sources and five receivers and a 96000 Hz sampling frequency while the proposed method only requires sixteen random RIRs and an 8000 Hz sampling frequency. From the experimental results, our proposed method achieves approximately the same accuracy while requiring approximate 10^4 less computational effort after training. To conclude, our CNN based room geometry estimation method is computationally efficient with approximately the same estimation error and, in contrast to the conventional signal processing based method, does not require prior knowledge or knowledge of the measurement configuration. Moreover, if lower accuracy is required, our method allows the usage of fewer measurements.

Our last experiment on room geometry estimation was the generalization to real-world

Table 3.7: Room geometry estimation with real-world measured RIRs.

Room	RMSE (m)	RMSE after averaging (m)
Hotel room	[0.1516, 0.1276, 0.2615]	[0.1046, 0.0505, 0.1169]
Meeting room	[0.1083, 0.0639, 0.1508]	[0.0916, 0.0220, 0.0440]
Office 1	[0.0508, 0.0532, 0.1023]	[0.0056, 0.0249, 0.0384]
Office 2	[0.0803, 0.0757, 0.2240]	[0.0390, 0.0207, 0.0938]
Enclosed staircase	[0.1790, 0.0998, 0.0970]	[0.1696, 0.0923, 0.0825]
Office 3	[0.1516, 0.0365, 0.1305]	[0.1432, 0.0081, 0.0112]

RIRs with transfer learning. Before feeding the real-world RIRs into the neural network, we first resampled the real-world RIRs to 8000 Hz and then used the first 4096 samples of the resampled RIR as the input. With transfer learning, the base method model was adopted as initialization and the learning rate of the optimizer was set to be one-tenth of the original learning rate. This generalization was split into two steps. We trained 500 epochs for each step to prevent overfitting. We describe the two steps in detail in the next two paragraphs.

The first step was the transfer learning from the base model with additive noise, randomly deleted and added pulses, derived approximate distorted RIRs due to obstacles, and the RIR genereated with the RD-based analytical method for obstacles. These distorted RIRs were mixed as the training dataset for transfer learning in the first step. The model after the first step was saved as an initialization for the second step.

In the second step, we used transfer learning with real-world RIRs [319]. Cross-validation was used for the six selected rooms in the database. In each test set, we computed the RMSE per dimension to evaluate the generalization performance. Since there were multiple RIRs per room, the proposed averaging method was performed in each test set to increase accuracy.

The experimental results for room geometry estimation with real-world measured RIRs are shown in Table 3.7. Before averaging over multiple estimates from multiple RIRs, the minimal RMSE on a single dimension is 0.05 m and the maximum error is 0.26 m. The 0.26 m RMSE appears in the hotel room with two beds and other furniture inside, which is a room with relative many obstacles, but this error reduces to 0.12 m after averaging. After averaging, the minimal RMSE is 0.01 m and the maximal is 0.17 m. The 0.17 m RMSE after averaging method appears in the enclosed staircase, which is relatively difficult to handle because of the stairs. The difference between RMSE with and without averaging method does not consistently follow the results shown in Table 3.5. This is because the real measured 151 RIRs in each room are from five sources and 31 receivers, which indicates the measurements are not independent from each other.

We did an additional experiment to evaluate the importance of these four augmentation methods, where we left one data augmentation method out each time and repeated the two

Table 3.8: Evaluation of the importance of four data augmentation methods.

The left out data augmentation method	Average RMSE difference (m)
RIRs with noises	0.0310
RIRs with the 1st artificial distortion type	0.0570
RIRs with the 2nd artificial distortion type	0.0648
RIRs generated with the ARD-based analytical method	0.1210

steps in the previous experiment. We computed the RMSE after averaging and compared it with Table 3.7. We computed the average RMSE difference, where the positive sign indicates an increase in the RMSE when one data augmentation method is left out.

The average RMSE in Table 3.7 after averaging is 0.0644m. The leave-one-out experimental result is shown in Table 3.8. Observing the result, when one data augmentation method is left out, the corresponding RMSE increases. This shows all four data augmentation methods are all necessary and make a contribution to the estimation accuracy. In addition, comparing the increased RMSE (m), we can conclude that RIRs generated with the ARD-based analytical method is the most important among these four methods. This is likely because this method simulates the effect of obstacles on real-world RIRs most accurately.

3.5.3 Experiments on the estimation of reflection coefficients

In this subsection, we describe our experiments that relate to the reflection coefficients. We first describe the experiments on estimating only reflection coefficients from RIRs, where we cover the frequency-independent case and the frequency-dependent case. Next, we describe the experiment on linking the reflection coefficients to room geometry.

We performed the reflection coefficient estimation experiments under the assumption of six distinct reflection coefficients, one for each wall. We divide this into two cases according to their frequency dependency. For the frequency-independent reflection coefficients, we estimate the reflection coefficients from the corresponding full-band RIR. With respect to the frequency-dependent reflection coefficients, we estimate the reflection coefficients from the sub-band RIRs independently. We compared the estimation error of the sub-band RIRs and the full-band RIRs to explore the effect of frequency bands on reflection coefficient estimation accuracy.

The experimental results of estimating six distinct reflection coefficients in a rectangular room are shown in Table 3.9. With the full band RIRs, the average RMSE per dimension is 0.09. With the sub-band RIRs, part of the information of the RIRs is lost. Consequently, the RMSE of the sub-band RIRs is larger. In addition, the RMSE of the low pass RIR is smaller than that of the bandpass RIR and the high pass RIR. This is likely because the relation between the RIR and the coefficients is smoother for low pass signals and it is easier to

Table 3.9: RMSE of multiple reflection coefficients estimation.

Signals	RMSE
Full band RIRs	$[0.0872, 0.0954, 0.0984, 0.0929, 0.0826, 0.0837]$
Low pass RIRs	$[0.0904, 0.0979, 0.1001, 0.0971, 0.0903, 0.0873]$
Band pass RIRs	$[0.1098, 0.1213, 0.1124, 0.0978, 0.0906, 0.0884]$
High pass RIRs	$[0.1108, 0.1241, 0.1146, 0.0981, 0.0927, 0.0923]$

learn a smoother function by a neural network. In addition, when observing the RMSE for each reflection coefficient, the RMSE in the middle position is relatively large. This is consistent with the upper bound in Section 3.4.1 and results from having ordered reflection coefficients in the interval $[0, 1]$.

Comparing the experimental results in Table 3.9 and the upper bound derived in Section 3.4.1, each RMSE in Table 3.9 are substantially smaller than the upper bound derived in Section 3.4.1. This indicates our neural network does learn reflection coefficients from RIRs instead of simply generating a set of ordered random numbers.

In the remainder of this subsection, we describe the experiments on linking the reflection coefficients to the room geometry as outlined in Section 3.4.3. We start with the machine learning based method. With the machine learning based method, we computed the RMSE for the reflection coefficients to evaluate the estimation accuracy. Since the room geometry serves as conditioning, the RMSE for the room geometry is negligible and not recorded here. Based the estimated reflection coefficients, which are linked to the room geometry, we computed the $RT60$ with the Sabine-Franklin formula, which is compared with the $RT60$ calculated from the energy decay curve to compute the RMSE. After that, we took the six reflection coefficients from each output, re-ordered them, and computed the RMSE for each reflection coefficient again to compare the accuracy with the previous reflection coefficients only estimation experiment.

The experimental result of linking reflection coefficients to room geometry using machine learning based method is shown in Table 3.10. Each row of the second and the third columns is the RMSE for the pair of reflection coefficients corresponding to that edge. The RMSE for the paired reflection coefficients is slightly worse than for the previous experiment but the model can still link a pair of reflection coefficients to the room geometry. The corresponding RMSE for the $RT60$ based on these estimates is 0.0220 s. When we reordered the six estimated reflection coefficients, the RMSE is $[0.0795, 0.0742, 0.0809, 0.0854, 0.0854, 0.0915]$, which is approximately the same as the result in Table 3.9. This result proves that the estimation accuracy of the reflection coefficients does not decrease but the linking operation decreases the accuracy a little.

In addition to the machine learning based method, we can also link the reflection coefficients to the room geometry using the conventional signal processing method. Since

Table 3.10: RMSE of linking reflection coefficients to room geometry.

Room geometry	Reflection coefficients	
Edge 1	0.1017	0.1391
Edge 2	0.1058	0.1435
Edge 3	0.1117	0.1427

we use estimated room geometry and reflection coefficients, we only recorded the RMSE for $RT60$. We computed $RT60$ with the estimated room acoustical parameters using Sabine-Franklin's formula. We then compared it with the $RT60$ calculated from the energy decay curve, and recorded the RMSE.

Computing the $RT60$ using the conventional signal processing method, the corresponding RMSE is 0.0083 s, which is smaller compared to the machine learning based method. Since the difference in the RMSEs for estimates of the room geometry is negligible, the difference in the RMSEs for the $RT60$ is due to the linking process of the reflection coefficients.

3.6 Conclusion

We showed that it is possible to estimate the geometry of a shoebox-shaped room and also the reflection coefficients of its walls from RIRs using deep neural networks. We formulated the problem as a regression problem with the MSE as a loss function. In contrast to conventional methods, the proposed methods only requires a single RIR between a source and a receiver and do not require knowledge of their positions or relative distance. For the room geometry estimation task, we used convolutional neural networks. We first trained the neural network with artificial data. Then transfer learning was used to make the method work for real-world RIRs. We achieved an average of 0.065 m testing accuracy for real-world data. We used multilayer perceptrons to estimate the wall reflection coefficients from simulated RIRs. We obtained an RMSE of approximately 0.09 for each reflection coefficient when the reflection coefficients are different for the six walls. This value increased slightly if we require pairs of reflection coefficients to be associated with an estimated room geometry. In addition, we were able to estimate frequency-dependent reflection coefficients and achieved similar accuracy.

4

Estimation of TOAs and Room Acoustic Parameters from an Omnidirectional Room Impulse Response

We describe a new method for room acoustic parameter estimation given room impulse responses (RIRs). The method is composed of two parts. The first part utilizes the transformer to estimate the time of arrivals (TOAs) of the direct path and specular reflections up to the second order. The image source method describes the TOAs of specular reflections, which might not correspond to peaks in real RIRs. For this reason, we estimate TOAs described by the image source method. The estimated TOAs are used as inputs of the analytical method to estimate room acoustic parameters, i.e., room geometry and source/receiver positions. The analytical method is based on a symmetry analysis of room impulse responses. In contrast to the state-of-the-art methods, the proposed method only requires a single room impulse response. For real-world environments, the proposed method can achieve an accuracy of an average of 0.0597 *m,* 0.0650 *m, and* 0.0760 *m on each dimension of room geometry, source position, and receiver position, respectively, with a failure rate of* 18.5%*.*

4.1 Introduction

Accurate acoustic environment modeling forms an important aspect of room acoustics. It has a variety of applications such as speech enhancement [325–327], speech recognition [328–330], and sound rendering [331–333]. The room impulse response (RIR) characterizes the room acoustic environment. It is affected by a set of room acoustic attributes, including room geometry, the positions of the source and receiver, and the reflection coefficients. In this chapter, we aim to extract time of arrivals (TOAs) of reflections, and derive the room geometry and source/receiver positions given a single RIR.

A number of methods exist to model RIRs. They can be categorized into wave based methods and geometrical acoustics based methods. Wave based methods [10, 69, 70] simulate RIRs numerically and accurately. The acoustic space needs to be discretized to solve wave equations. Methods include the Finite-Element Method (FEM) [7–9], the Boundary-Element Method (BEM) [65, 66], and the Finite-Difference Time Domain (FDTD) [71–74] based methods. Wave based methods can achieve high accuracy. However, these methods have a high computational load, especially for high frequencies. Geometrical acoustics based methods [4, 10] assume that the sound propagates in straight lines. The most commonly used geometrical acoustic methods can be classified into the image source method (ISM) [4, 10–12, 334], the ray tracing method [101–105] and the beam tracing method [4, 106–115]. Unlike wave equation based methods, they are unable to simulate some low frequency effects such as diffraction.

Among the above mentioned methods, we highlight the image source method [10–12, 334] and the ray tracing method [101–105] since we use these two methods to simulate RIRs in this chapter and the image source method is the basis of the proposed room acoustic parameter estimation method. The ISM was first proposed by Allen and Berkley [334] in 1979. The ISM can model the TOAs of the direct path and specular reflections accurately. In addition, it is computationally efficient, making it suitable for generating a large scale database. However, RIRs simulated by the image source method differ from real measured RIRs in several aspects. Firstly, the image source method cannot model frequency dependent components, such as, frequency dependent reflection coefficients. Secondly, the image source method can not be used for curved and non-smooth reflective surfaces and can not model diffraction or scattering. Lastly, empty rectangular rooms are always assumed, although several improved methods exist to deal with irregular shapes. These assumptions make the simulated RIRs far from real-world RIRs. The ray tracing method was extended from optical applications to room acoustics in [105]. The basic procedure uses similar principles as the image source method. With the ray tracing method [4, 101–104], the source emits the rays according to a predefined distribution or Monte Carlo simulation, and valid reflected paths are retained. The ray tracing method can handle not only specular reflections but also diffuse reflections. The ray tracing methods face a detection problem and limited spatial resolution [116]. The detection problem originates from the fact that

it is impossible for a ray to hit a point receiver. To prevent this, the ray tracing method assumes a finite-size receiver. As a result, it may suffer from misidentification of rays or duplicated registered rays [4]. The limited spatial resolution results from the limited number of traceable rays.

Identifying early reflections in a room impulse response is of great importance in many room acoustics applications, such as room geometry estimation and speech dereverberation [145]. Conventional signal processing methods solved this problem with limited accuracy because of phase distortion and non-linear effects. Due to the increased computational power and the availability of the large scale database, deep learning has seen a rapid increase in usage. Conventional deep learning models include multilayer perceptrons (MLPs) [211, 213], convolutional neural networks (CNNs) [219, 238, 240], recurrent neural networks (RNNs) [247, 335, 336], and so on. Transformers [290, 337] recently have become increasingly popular. They model the relationships between all nodes on each layer independently of their positions. Transformers do not rely on recurrence or convolution and show a good modeling ability due to the multi-head attention mechanism. Hence, we use transformers to estimate the TOAs of early reflections from RIRs.

The main contribution of this chapter is the estimation of TOAs of specular reflections and room acoustic parameters from a single RIR. In the context of this chapter, we refer to the reflection as the entire ray from the source to the receiver, which can contain multiple elementary reflections. Since the phase distortion can blur or bias peaks in the RIR and the peaks might not result from the specular reflections, we aim to estimate the TOAs of specular reflections described by the image source method. Given a single omnidirectional room impulse response, we first use a deep learning based method to estimate the TOAs of specular reflections up to second order. We then use an analytical method to simultaneously estimate the room geometry and source/receiver position without prior information based on estimated TOAs. The resulting room acoustic parameter estimation method applies to rooms with parallel wall pairs. The experimental results confirm the validity of our proposed method.

This chapter is organized as follows. Section 4.2 discusses the estimation of TOAs of specular reflections using transformers. We then describe the degeneracy of room impulse responses and the room acoustic parameter estimation in Section 4.3. The experimental results are discussed and analyzed in detail in Section 4.4. Finally, we conclude our chapter in Section 4.5.

4.2 TOAs of specular reflection estimation

In this section, we describe the estimation of TOAs of the direct path and specular reflections up to second order using transformers. First, we discuss the motivation to use deep neural networks to solve the problem. Next, we formulate the estimation problem for the TOAs of the specular reflections estimation problem. Finally, we describe how we solve this

problem.

As discussed, TOA estimation is of great importance in room acoustic applications, and conventional signal processing based methods either find the local maximum or compare the similarity between a pulse and the direct pulse. However, the pulse shape of the reflections will change due to phase distortion, and it cannot recognize whether a detected pulse corresponds to a specular reflection. Our previous work proves that deep learning based methods can handle real-measured RIRs [338]. This motivates us to use a deep learning based method to estimate TOAs of specular reflections. Specifically, since our room acoustic parameter estimation algorithm is based on the image source method, we aim to find the TOAs of specular reflections described by the image source method, which may not correspond to the peaks in real measured RIRs. Since room acoustic parameter estimation requires up to second order specular reflections and detecting specular reflections of higher order is more difficult, we aim to estimate TOAs of specular reflections up to second order only.

Transformers show good modeling ability in various applications. The multi-head attention mechanism allows the transformer to learn different relationships among different positions in RIRs. TOAs of specular reflections can be considered as a complex function of room acoustic parameters in RIRs. Consequently, we use transformers to blindly estimate TOAs of specular reflections described by the image source method from a single RIR. Our approach has the time-domain RIRs as input and a vector of TOAs as output. We adopt the transformer architecture of [290]. The TOA estimation problem can be formulated as a regression problem with omnidirectional RIR as input. A random variable pair (X,Y) can be used to define the input-output pair of neural networks, where X is an $\mathbb{R}^{d_X}$-valued random variable that represents RIRs with d_X denoting the length of each RIR signal vector, and Y is an $\mathbb{R}^{d_Y}$-valued random variable that represents the TOAs of specular reflections with d_Y denoting the length of each TOA vector. Then the problem can be formulated as learning a continuous deterministic function h with $y = h(x)$ where (x,y) is a realization of random variable pair (X,Y). To measure the performance of the regressor h, we define a loss function as l. We can then define the risk R as

$$R = \mathbb{E}[l(h(x),y)], \tag{4.1}$$

where the expectation $\mathbb{E}$ is calculated with respect to the distribution $f_X(x)$. For the deep learning based problem, the input distribution is unknown. Hence, an empirical risk based on the training dataset is used to approximate the risk R as

$$R_{\text{emp}} = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i),y_i), \tag{4.2}$$

where m is the size of training dataset and each (x_i,y_i) is a particular realisation of (X,Y). In the context of this chapter, we use the mean squared error (MSE) to measure the distance between the estimated TOAs and the ground truth.

Due to the limited amount and variety of real world data, we trained our TOA estimation model using simulated data and then evaluated the model using real-world data. Hence we required the simulated database to cover a wide range of room acoustic parameters and be of sufficient size. The geometrical acoustics based methods are more appropriate for generating simulated databases than wave-based methods due to their computational efficiency. The image source method can only simulate the specular reflections accurately. The effect of frequency dependent reflections, scattering, and others can not be simulated by the image source method. To make our training data close to real-measured RIRs and facilitate generalisation to real-world data, when we trained the neural network, we used a hybrid model that combines the image source method [12, 334] and ray tracing [101, 103, 105] to simulate the specular reflections and scattering, and also considered in the RIRs the frequency dependency of reflections, air attenuation and additive noise.

4.3 Room acoustic parameters estimation

This section describes our method to estimate room acoustic parameters. To start with, we formulate the image source method to simulate RIRs, which is the basis of our derivation. In the second subsection, we propose a method to determine a particular valid configuration of the room acoustic parameters from a single room impulse response. In Section 6.2.1, we will show how to map a valid configuration to any other valid configuration. Our methods apply to rooms with sets of parallel walls with adjacent walls at an angle of 90 degrees. We assume the pulses of the direct path, first and second order reflections are available in the observed RIR.

4.3.1 Image source method

In the image source method, an empty rectangular room is assumed, and non-specular reflections are not considered. In addition, it assumes that sound propagates along straight lines. Each reflection can be modeled as a pressure wave emitted from an image source in free space. We use $\mathbf{p},\mathbf{m}$ to label each reflection where each element of $\mathbf{p} = (q, j, l)$ can take a value of 0 or 1, indicating the direction of the reflection, and each element of $\mathbf{m} = (m_x, m_y, m_z)$ can take an integer value, indicating the position of the virtual room where the image source is located. In three-dimensional (3D) space, we denote the position of the receiver as (x_r, y_r, z_r) and the position of the source as (x_s, y_s, z_s). Implementing the image source method [76], the image source position can be represented as $(2m_xL_x + (1-2q)x_s, 2m_yL_y + (1-2j)y_s, 2m_zL_z + (1-2k)z_s)$, where (L_x, L_y, L_z) are the length width and the height of the room. Let $d_{\mathbf{p},\mathbf{m}}$ denote the corresponding path length, then the time delay can be calculated as $\tau_{\mathbf{p},\mathbf{m}} = d_{\mathbf{p},\mathbf{m}}/c$. The amplitude of each reflection is determined by the reflection coefficients $\beta_{x_1}, \beta_{x_2}, \beta_{y_1}, \beta_{y_2}, \beta_{z_1}, \beta_{z_2}$, reflection order $O_{\mathbf{p},\mathbf{m}}$, and image source position. The reflection order $O_{\mathbf{p},\mathbf{m}}$ is the number of reflections in a path and can be

computed as

$$O_{\mathbf{p},\mathbf{m}} = |2m_x - q| + |2m_y - j| + |2m_z - l|. \tag{4.3}$$

If we assume the finite and constant reflection coefficients for each wall, then the RIR can be written as [334]

$$h(t) = \sum_{\mathbf{p},\mathbf{m}} \beta_{x_1}^{|m_x-q|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y-j|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z-l|} \beta_{z_2}^{|m_z|} \frac{\delta(t-\tau_{\mathbf{p},\mathbf{m}})}{4\pi d_{\mathbf{p},\mathbf{m}}}. \tag{4.4}$$

4.3.2 Room acoustic parameters estimation

In this subsection, we describe how to estimate the room acoustic parameter from RIRs. Based on the image source method in Section 4.3.1, the path length of a reflection is calculated as the product of the TOA and the speed of sound. If a sound reflects on one wall, we define the direction of this reflection as belonging to this parallel wall pair. The coordinates originate at one corner of the room, and the three axes are assumed to be parallel to the walls. The proposed method identifies the reflections for each direction and then computes the wall-pair distance with the path lengths of reflections in this direction. We first describe our theorem to identify the directions of reflections. We then propose an algorithm to identify reflections. Finally, we show how to compute room geometry and source/receiver positions.

Identifying directions of reflections

We introduce a theorem to identify the directions of reflections given a set of unlabelled path lengths of the reflections. This theorem is used to classify higher order reflections into two sets: a multi-direction set and a single-direction set.

Theorem 1. *Let d_{ij}, d_i, d_j, and d_0 denote the path lengths of the (O_i+O_j)-th order reflection that reflects on two directions i and j, with O_i denoting the number of relection in drection i and O_j denoting the number of relection in drection j, and the direct path. Then $d_{ij}^2 + d_0^2 = d_i^2 + d_j^2$ holds. Vice versa, if there exists a path length of (O_i+O_j)-th order reflection d_{ij} that satisfies the equation, the reflections corresponding to path lengths d_i and d_j belong to different directions.*

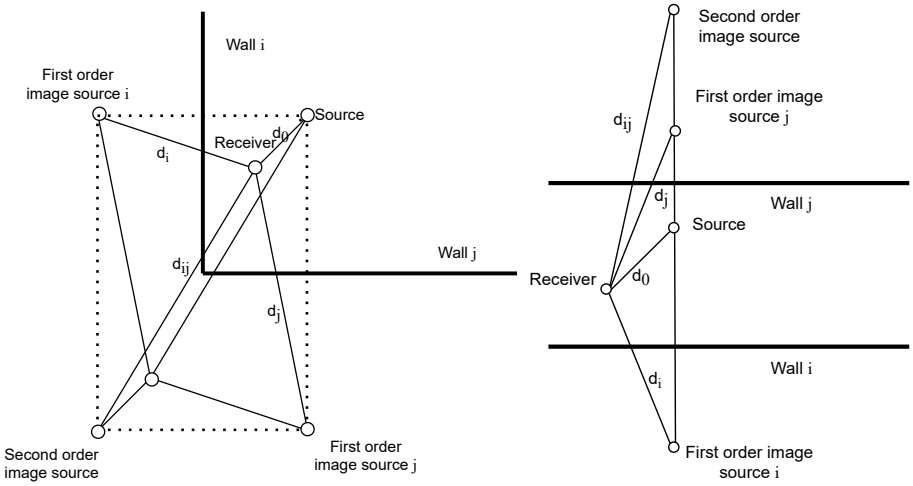
Proof. We assume the image source of path length d_i to be in the x direction and d_j to be in the y direction. The coordinates of the corresponding image sources are $(2m_xL_x + (1-2q)x_s, y_s, z_s)$, and $(x_s, 2m_yL_y + (1-2j)y_s, z_s)$ respectively. The coordinates of the (O_i+O_j)-th order image source are $(2m_xL_x + (1-2q)x_s, 2m_yL_y + (1-2j)y_s, z_s)$. We can then compute the path length d_i between the image source and the receiver as

$$d_i^2 = (2m_xL_x + (1-2q)x_s - x_r)^2 + (y_s - y_r)^2 + (z_s - z_r)^2.$$

The same formulation holds for the other cases. Formulating d_i, d_j, d_{ij} and d_0 in the same form, it is seen that $d_{ij}^2 + d_0^2 = d_i^2 + d_j^2$ always holds when d_i and d_j belong to path lengths of

reflections in different directions. When they belong to the same direction, the equation is not valid. □

Theorem 1 can also be proved with the parallelogram law as Fig. 4.1a. Fig. 4.1 shows the 2D case for better understanding but is also valid for the 3D case. From Fig. 4.1a, we know the distance between the source and the second order image source equals the distance between two first order image sources, and these two parallelograms share one diagonal. Following the parallelogram law, the sum of the squared side lengths of these two parallelograms are equal, i.e., $d_{ij}^2 + d_0^2 = d_i^2 + d_j^2$. Fig. 4.1b is an example where the higher order path length cannot be written as a function of lower order reflections.



(a) More than one direction (b) One direction

Figure 4.1: Second order reflections for the 2D case.

The *multi-direction* set of second (or higher) reflections refers to those that reflect in more than one direction. The multi-direction set does not provide independent equations since the squared path length can be written as a combination of lower order reflections. The reflections in this set are useless for our purpose and can be pruned out using lower order reflections since they are sure to arrive after the lower order reflections. An example of this is the second order reflection from the image source $(-x_s, -y_s, z_s)$, which reflects on both x and y directions; the squared path length can be written as a combination of the direct path (x_s, y_s, z_s) and two first order reflections from $(-x_s, y_s, z_s)$ and $(x_s, -y_s, z_s)$.

The *single-direction* set of second (or higher) order reflections refers to those that reflect along one direction. They can be used to determine the room acoustic parameters. The

N-th order reflections that reflect along one direction arrive later than the last arrived $(N-1)$-th order reflection in that direction. For example, the second order reflections from $(x_s \pm 2L_x, y_s, z_s)$ arrive after the first order reflection from $(2L_x - x_s, y_s, z_s)$. Since the source and receiver are exchangeable, which we will assume below without loss of generality, we assume $x_r < x_s$, $y_r < y_s$, and $z_r < z_s$ to analyze the sequence of these reflections. As an example, the reflection from $(x_s - 2L_x, y_s, z_s)$ arrives earlier than that from $(x_s + 2L_x, y_s, z_s)$.

Algorithm of classification of reflections

We describe our algorithm to classify reflections in detail. We assume we have a set of unlabelled path lengths of reflections in a RIR signal which contain the direct path, the first and second order reflections. We assume the first order reflections are distinguishable. The input set is sorted in ascending order and denoted as d_k, $k \in \mathbb{N}$. We aim to classify the path lengths of reflections into five sets, i.e., the first order reflections in the x direction S_x^1, the first order reflections in the y direction S_y^1, the first order reflections in the z direction S_z^1, the second order reflections in a single-direction S_{single}^2, and the multi-direction second order reflections S_{multi}^2. We have $|S_x^1| = 2$, $|S_y^1| = 2$, $|S_z^1| = 2$, $|S_{\text{single}}^2| = 6$, and $|S_{\text{multi}}^2| = 12$. We introduce a hyperparameter δ as an error threshold in the path length equation since a difference might exist between the detected peak position and the theoretical path length. The error threshold depends on the data property and path length.

The first arrived pulse d_0 always corresponds to the path length of the direct path. Without loss of generality, we assume the path length of the second arrived pulse (first arrived first order reflection) d_1 corresponds to $(-x_s, y_s, z_s)$. Let us label d_1 as d_x. The correctness of this assumption will be explained in Section 6.2.1. We iterate over the input set, and once a path length of reflection is classified into one set, it will be deleted from the input set. We first find all d_i and d_j that satisfy $d_j \in [\sqrt{d_i^2 + d_x^2 - d_0^2} - \delta, \sqrt{d_i^2 + d_x^2 - d_0^2} + \delta]$ where $1 < i < j$. These d_j belong to S_{multi}^2, and without loss of generality, we assume the smallest d_i belongs to S_y^1. Let us label this path length as d_y. We iterate over the remaining found d_i and the remaining path lengths d_k in the input set. For all d_i and d_k that satisfy $d_k \in [\sqrt{d_i^2 + d_y^2 - d_0^2} - \delta, \sqrt{d_i^2 + d_y^2 - d_0^2} + \delta]$, we have $d_k \in S_{\text{multi}}^2$ and $d_i \in S_z^1$. Then the remaining d_i belongs to S_y^1. Till now, we have already found all possible path lengths of reflections in S_y^1 and S_z^1. If $|S_y^1| > 2$ or $|S_z^1| > 2$, we cross validate these two sets, i.e., we iterate over a combination of two elements d_{yi} from S_y^1 and a combination of two elements d_{zj} from S_z^1 to find the combination that satisfies $\exists d_k, d_k \in [\sqrt{d_{zj}^2 + d_{yi}^2 - d_0^2} - \delta, \sqrt{d_{zj}^2 + d_{yi}^2 - d_0^2} + \delta]$. Next, we iterate over the input set again with d_y (this can also be replaced by one of the path lengths in S_z^1) to find d_j and d_i that satisfy $d_j \in [\sqrt{d_i^2 + d_y^2 - d_0^2} - \delta, \sqrt{d_i^2 + d_y^2 - d_0^2} + \delta]$. We then have $d_i \in S_x^1$ and $d_j \in S_{\text{multi}}^2$. Lastly, the remaining path lengths of reflections in the input set are allocated to S_{single}^2.

The algorithm is robust to mislabelled first or second pulses. If the first two arrived pulses do not correspond to the direct path and the first order reflection in x direction, Theorem 1 does not work for second order reflections, and the cardinality of the sets does not match. In that case we can conclude there exist erroneous pulses in the first two arrived pulses. We can then use the path length of the next arrived pulse and select two from these pulses and repeat the process until the cardinality is correct.

Estimation of room geometry and source/receiver positions

After classifying the reflections, we describe how to compute a valid configuration for the room geometry and source and receiver positions. As discussed, S^2_{multi} is not useful for this computation. Thus, we only use S^1_x, S^1_y, S^1_z, and S^2_{single}. The arrival sequence of reflections will be explained in section 6.2.1.

Our method is based on an iteration of S^2_{single}. This set has six second order reflections in this set, and the coordinates of the image sources are $(x_s \pm 2L_x, y_s, z_s)$, $(x_s, y_s \pm 2L_y, z_s)$, $(x_s, y_s, z_s \pm 2L_z)$. For each of these six elements, we perform the following computation. The reflection with the smallest path length in S^2_{single} has three possible directions. We use each second order reflection candidate, together with the first order reflections in this direction and the direct path, to compute the room geometry and the source and receiver position in this direction. For each second order reflection candidate, we derive the coordinate of another second order reflection in this direction and calculate the corresponding path length, which should be an element of S^2_{single} for the correct second order reflection candidate. This combination can also be verified with the reflection coefficients in the next subsection.

Let us determine if the hypothesis is correct that a particular distance in S^2_{single} corresponds to the x direction, i.e., the image source is $(x_s - 2L_x, y_s, z_s)$. Together with the path lengths of two first order directions in this direction, with image sources $(-x_s, y_s, z_s)$ and $(2L_x - x_s, y_s, z_s)$, and the path length of direct path from (x_s, y_s, z_s), we can compute the three unknowns, i.e., L_x, x_s, and x_r, from three linear independent equations. We can then compute the path length of the second order reflection from $(x_s + 2L_x, y_s, z_s)$ in this set, and if this is consistent with the initial hypothesis, then we have verified it. If so, we also computed the second second order reflection in this direction. This procedure allows us to find the second-order pulses in each direction.

The proposed method is relatively robust in four aspects. Firstly, generalizing this algorithm to include additional higher order reflections can improve the algorithm's robustness. Theorem 1 can also be applied to higher order reflections to classify directions. We can apply the proposed method to the higher order reflections that reflect along one direction if the corresponding TOA informaiton is available. As a result, the higher order reflections can be used to verify the solution of room acoustic parameters to improve robustness. Secondly, the proposed method is robust to additional peaks that do not belong to any reflection. With the proposed algorithm, these pulses will be misclassified into the set containing higher-order reflections that reflect along the same direction. Since

another second order reflection that also reflects along this direction does not exist, we know this peak is erroneous. Thirdly, the proposed method is robust to missing peaks to some extent. If the direct path is missing, it does not work. Hence, we assume the direct path is always available. If a first order reflection is missing, two cases can happen. One case is that the proposed algorithm can not find enough first order reflections. The second case is that another pulse is mislabelled as a first order reflection. However, as discussed in Section 4.3.2, the algorithm is robust to mislabelled first or second pulses. As a consequence, for both cases, the room acoustic parameters in the corresponding direction can not be derived. Still, it does not affect the room acoustic parameter estimation in the remaining directions. Finally, the proposed method is robust to the offset of RIR. Since the room acoustic parameter calculation uses the difference between TOAs of reflections and direct path, the offset is eliminated in this process.

The independent calculation in each direction is an advantage of our proposed method. Since the calculation for each direction is separable, the method can also be applied to some special cases. We can estimate the distance between parallel walls and the source/receiver position along this pair of parallel walls, whether the walls in other directions are very distant (such as in a hallway) or affected by furniture. The independent calculation for each direction makes our method work for non-shoebox shaped rooms with sets of parallel wall pairs as long as the required reflections are available. An example is a room with a sloped ceiling. Such a room has two pairs of parallel walls and a pair of non-parallel walls. For the second order reflections between the vertical wall and the floor, since they form a right angle, the second order reflections will be pruned out by the first order reflections. For the second order reflections between the ceiling and the remaining walls, since they do not follow Theorem 1, they will not be pruned out and will be classified into S_2^{single}. However, they will be recognized as erroneous pulses since no valid solution exists for room acoustic parameters.

4.3.3 Room impulse response degeneracy analysis

In Section 4.3.2, we discussed how to compute room acoustic parameters from a room impulse response. However, the room configuration, including room geometry, positions of source and receiver, reflection coefficients, and the coordinate system, is not unique for a particular RIR. Fig. 4.2 is a 2D example where eight different configurations result in an identical RIR. In addition, if we exchange the source and receiver positions, the RIR will also not change. Hence, for a 2D room, 16 configurations can result in an identical RIR. Except for this degeneracy, the TOA of reflections of a RIR is unique with respect to a room configuration.

Next, we analyze the degeneracy of a 3D room and determine the coordinate system based on the first order reflections. Let (L, W, H) denote the room geometry, and (x_s, y_s, z_s) and (x_r, y_r, z_r) denote the coordinates of source and receiver. Let us assume the first

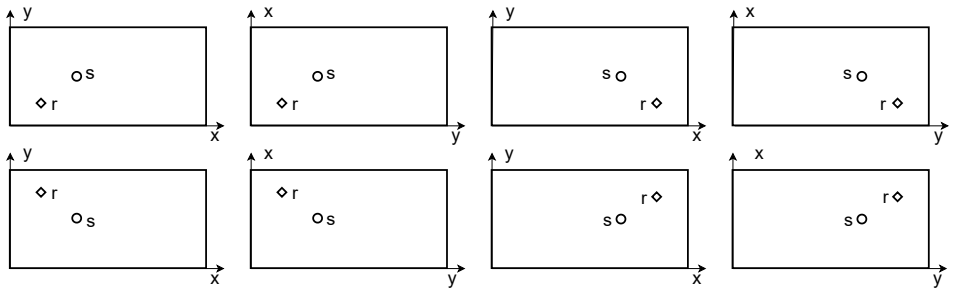


Figure 4.2: Different 2D configurations resulting in an identical RIR where s and r denote source and receiver, respectively.

reflection is the pulse that reflects on the wall $x = 0$. Since the first pulse can reflect on any wall, this gives us six-fold degeneracy. This is equivalent to the pulse that reflects on $x = 0$ has the shortest path length, which is

$$\begin{cases} \|-x_s - x_r\| < \|2L - x_s - x_r\| \\ (-x_s - x_r)^2 + (y_s - y_r)^2 < (x_s - x_r)^2 + (-y_s - y_r)^2 \\ (-x_s - x_r)^2 + (y_s - y_r)^2 < (x_s - x_r)^2 + (2W - y_s - y_r)^2 \\ (-x_s - x_r)^2 + (z_s - z_r)^2 < (x_s - x_r)^2 + (-z_s - z_r)^2 \\ (-x_s - x_r)^2 + (z_s - z_r)^2 < (x_s - x_r)^2 + (2H - z_s - z_r)^2 \end{cases} . \tag{4.5}$$

Similarly, the next arriving first order reflection not in the x direction introduces a four-fold degeneracy by assuming it reflects on the wall $y = 0$, which implies

$$\begin{cases} \|-y_s - y_r\| < \|2W - y_s - y_r\| \\ (-y_s - y_r)^2 + (z_s - z_r)^2 < (y_s - y_r)^2 + (-z_s - z_r)^2 \\ (-y_s - y_r)^2 + (z_s - z_r)^2 < (y_s - y_r)^2 + (2H - z_s - z_r)^2 \end{cases} . \tag{4.6}$$

Finally, the third arriving first-order reflection not in the x and y directions has two-fold degeneracy by assuming it reflects on the wall $z = 0$, which is

$$\|-z_s - z_r\| < \|2H - z_s - z_r\|. \tag{4.7}$$

This results in an overall 48-fold degeneracy. In addition, the coordinates of the source and receiver are exchangeable, which results in a total of 96-fold degeneracy. The conditions

for this example mode can be summarised as

$$\begin{cases} 0 < x_r < x_s \\ 0 < y_r < y_s \\ 0 < z_r < z_s \\ x_r + x_s < L \\ y_r + y_s < W \\ z_r + z_s < H \\ x_s x_r < y_s y_r < z_s z_r \\ x_s x_r < (W - y_s)(W - y_r) \\ y_s y_r < (H - z_s)(H - z_r) \end{cases}, \tag{4.8}$$

4

where the first three lines correspond to the exchangeable source and receiver coordinates, and the remaining lines are simplified versions of (4.5), (4.6), and (4.7), for example, the sixth line corresponds to (4.7).

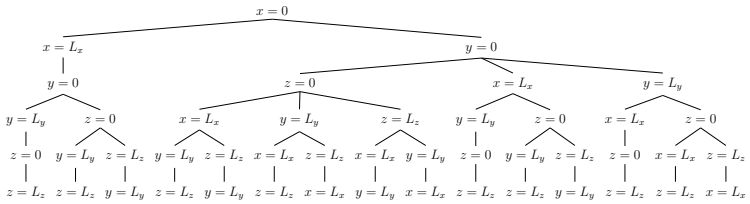


Figure 4.3: The sequence of first order reflections for the 3D case.

We discuss one case as an example of how the degeneracies appear in a practical setup since other cases are similar. The case corresponds to one particular branching pathway in Fig. 4.3, which shows the possible orderings of first-order reflections. As discussed, the first arrived first order reflection reflects on $x = 0$, which introduces six-fold degeneracy. We note that the second arrived first order reflection can reflect on $x = L_x$ or in the y or z direction. We consider the case where the second arrived first order reflection reflects on $x = L_x$. We then define the third arrived first order reflection to reflect on $y = 0$, which introduces four-fold degeneracy. Then the fourth arrived first order reflection will reflect on $y = L_y$ or in the z direction. We consider that the fourth arrived first order reflection reflects on $y = L_y$. We then define the fifth arrived first order reflection to reflect on $z = 0$, and the sixth arrived first order reflection to reflect on $z = L_z$. Assuming the fifth reflection to reflect on $z = 0$ introduces two-fold degeneracy. The exchange of the coordinates of source and receiver results in an overall 96-fold degeneracy for this branching pathway. For other pathways, we find the same result.

Mapping from one of the degenerate solutions to another is straightforward. The methods consist of coordinates exchange of source and receiver, exchange of x, y, and z coordinates, and symmetry with respect to $x = \frac{L_x}{2}$, $y = \frac{L_y}{2}$, or $z = \frac{L_z}{2}$. For an audio-only environment, it does not matter which case to choose. However, for an audio-visual environment, it is necessary to match one of the degeneracies with the visual scene for an acceptable experience. We need to use the available visual cues to determine which case and what combination of methods to use, which is out of the scope of this chapter.

4.4 Experiments

In this section, we present our experiments. We describe the setup of our experiments in the first subsection. In the second subsection, we describe the experiments on the TOA estimation and the room acoustic parameter estimation. The performance on room acoustic parameter estimation is compared with the CNN-based method proposed in [338].

4.4.1 Experimental Setup

In the following, we first describe the database we used to train and test our TOA estimation model. After that, we describe the configuration of our neural networks.

Database

A large-scale dataset of good quality is necessary to train neural networks. Simulated data was used to train and monitor the training process of our TOA estimation model. We used Pyroomacoustics [339] to generate our simulated database. We used a hybrid simulation method, which combined the image source method and the ray tracing method, to simulate RIRs in shoebox-shaped rooms. The speed of sound was set to be $c = 340$ m/s. The sampling frequency was set to 8000 Hz. The length of each RIR was 1024, which includes the specular reflections up to the second order. We assumed the rooms were rectangular and empty. Each dimension of the room geometry, i.e., length × width × height, was assumed to be iid between $15 \times 11 \times 4$ and $7 \times 5 \times 2$, which can cover a variety of rooms. One source and one receiver were randomly placed in the room, and they were assumed to be omnidirectional. To facilitate our room acoustic parameter estimation method, the room geometry and source/receiver position were assumed to follow one default mode as (6.3). The reflection coefficients on each wall were assumed to be frequency-dependent and were set uniformly distributed between 0 and 1 in each octave band of each wall. The signal to noise ratio (SNR) was set to be between 30 dB and 50 dB following [158]. To mimic the real measured RIRs, the air attenuation was also included. For the ray tracing method, the radius of the sphere of the receiver was set to be 0.5 m. The maximum TOA of rays was set to be 10 s, and the time granularity of bins in the energy histogram was set to be 4 ms. The total number of simulated RIRs was 400000, which was divided into a training dataset and a validation dataset with a ratio of $8 : 2$.

We used [319] as our real-world RIR database because it contains a relatively large number of real RIRs, and the room acoustic parameters were measured in each room. This database contains nine distinct rectangular rooms that are not empty. It covers several types of rooms, such as a meeting room, a lecture room, and an enclosed staircase. Within each room, an average of 155 RIRs are given between five sources and 31 receivers. Among these RIRs, we select the RIRs where the microphones can receive the direct path from the sources since our room acoustic parameter estimation method requires the availability of the direct path.

Neural Network description

We describe the neural network structure and how we train our neural network below. We performed a preliminary experiment with our target database to tune the hyperparameters using a grid search. The architecture of the transformer followed [290]. We set the number of heads in the multi attention mechanism to be 8. The number of layers in the encoder and decoder was set to be 6.

We used a GPU node to train the transformer. We trained the transformer using the Adam optimizer [320] to minimize the training loss. The learning rate of the Adam optimizer was 0.001, and the coefficients used for computing running averages of the gradient and its square were set to be $(0.9, 0.999)$. We adopted early stopping [321] as regularisation. Early stopping was performed when the validation performance degrades in 100 successive epochs to guarantee the training performance without overfitting and limit the computational effort. In addition, we set the maximum iteration epochs to be 5000. For computational efficiency, mini-batch training [322] was used with a batch size of 50. We evaluated the trained transformer using the real database of [319].

4.4.2 Experiments on TOA estimation and room acoustic parameter estimation

Since we trained the transformer to estimate TOAs matching the image source method, the TOAs may not correspond to peaks in the RIRs. In addition, the ground truth TOAs of specular reflections are not available in real-measured RIRs. Hence, we do not evaluate the TOA estimation directly. The estimated TOAs were used as inputs of the room acoustic parameter estimation algorithm, and we evaluated the room acoustic parameter estimation algorithm.

For the room acoustic parameter estimation algorithm, we assume the bandwidth of the RIR equals the Nyquist bandwidth. The error threshold was set to $\delta = 5/\mathrm{fs}$, which is a balance between the accuracy and successful rate. We recorded the RMSE of the estimated room geometry and source/receiver position. We also recorded the failed cases when the method outputs an empty set with the detected reflections. The failed cases were excluded from the RMSE calculation. It should be noted that when we calculated the RMSE and

counted the failed cases, we treated three coordinates separately since the estimation on three edges is independent. We recorded the average running time to evaluate the computational efficiency.

The RMSE of room acoustic parameter estimation on each edge and the failure rate is shown in Table 4.1. When the analytical method failed on one edge, the estimation of the

Table 4.1: Experimental results of room acoustic parameter estimation

RMSE of room geometry	# RMSE of source position	# RMSE of receiver position	# Failure rate
0.0597 m	0.0650 m	0.0760	18.5%

other edges is not affected, which shows that the estimation on three edges is independent. In addition, we compare the proposed method with the CNN based method in [338] in terms of the room geometry estimation accuracy, the failure rate and the computational efficiency, which is shown in Table 4.2. The proposed method use the same real measured

Table 4.2: Comparison of two room acoustic parameter estimation methods

Methods	# The proposed method	# The method in [338]
RMSE of room geometry	0.060 m	0.065 m
Failure rate	18.5%	0
Average wall clock time	13.47 s	3.22×10^{-4} s

RIRs as the method in [338]. Both methods run on a MacBook Pro, Mid 2014, 2.6 GHz Dual-Core Intel Core i5. From Table 4.2, we observe that these two methods show similar estimation accuracy. The RMSE of the CNN based method in [338] is slightly larger, but the failure rate is 0. We hypothesize that a larger RMSE of the CNN based method is because the CNN based method also finds it difficult to estimate the room geometry of some RIRs. For the RIRs that are difficult for estimation, the analytical method fails to give a valid solution, while the CNN based method can assign a random but reasonable value, which results in a slightly large RMSE. The proposed method is less efficient than the CNN based method in [338] because iteration is performed for the analytical method to estimate room acoustic parameters.

4.5 Conclusion

In this chapter, we proposed a method to derive the room geometry, the positions of source and receiver, and reflection coefficients simultaneously. The proposed method was divided into a transformer part and an analytical part. To start with, given a RIR, we used the transformer to estimate the TOAs of the direct path and the specular reflections up to the second order. The transformer only requires a single RIR without additional

prior information. The analytical method takes the estimated TOAs as input for room acoustic parameter estimation. It is based on the symmetry analysis of RIRs. The proposed analytical method is robust to erroneous pulses, non-specular reflections, and an unknown offset. The estimation on different dimensions is independent. For room geometry, source position, and receiver position, we achieved the RMSE of 0.0597 m, 0.0650 m, and 0.0760 m, respectively, with a failed portion of 18.5% of a real-world measured RIR database.

5

Necessary attributes for integrating a virtual source in an acoustic scenario

We investigate what information about a room is necessary to integrate a new source into an existing scenario. In particular, we consider the effects of the reflection order, the order of ambisonics signals and reverberation time. We conducted a series of listening tests and used the control variates method to determine the quantitative relevance of the selected attributes. In terms of integration and accurate localisation, at least third order ambisonics description of a source, is required for integration of that source. In addition, a finite number of early reflections can perform equally well to a full room impulse response when a new source is integrated into an existing scenario. However, the room impulse response with only the correct reverberation time is not sufficient.

This chapter is published as "Necessary attributes for integrating a virtual source in an acoustic scenario," by W. Yu and W. B. Kleijn, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, 2018, pp. 21-25.

5.1 Introduction

Head-set based virtual reality (VR) is a specific immersive audio-visual environment that simulates a user's physical presence in an artificial scenario with corresponding VR headsets. Virtual reality will play an increasingly important role in numerous aspects of daily life, such as entertainment, education and health care. Spatial audio aims to create a 3D audio experience, which is an important component for a believable VR system.

Our goal is to examine what information about a room is necessary to integrate a new source into an existing acoustic scene. This knowledge will allow us to synthesize a realistic, convincing audio component. We are not aware of existing work on the problem. To understand the integration problem better, we first review the composition of a head-set based VR system.

An accurate environment simulation is essential for perceptually acceptable sound in a VR system. To model the acoustics environment, we need to consider several physical attributes of sounds in a room, such as reflections and reverberation time. The image-source method is used to model reflections in a room [10, 13]. However, the computational load increases with an increasing number of reflection walls and it can only handle convex room shapes [340]. The high complexity of modelling reflections in acoustics environments makes efficient methods important [129, 341, 342]. Reverberation time, RT_{60}, is the time that the sound drops 60 dB below the original level [10]. Reverberation time is considered to be an important attribute in acoustic environment simulation. Several methods [343–345] exist to estimate the reverberation time.

Besides accurate environment simulation, a high quality soundfield reproduction system is of great importance. Ambisonics [16–18] has become the de-facto standard representation for VR systems. Ambisonics is particularly suitable for VR systems as head rotations are easily modelled as the rotation of sound fields in the spherical harmonics domain. With an ambisonics representation of sufficient order, a high quality binaural audio rendering system can give listeners a realistic spatial audio experience. Hence it allows us to demonstrate our work on spatial audio. A number of techniques can be used for binaural rendering of ambisonics [171, 346].

The main contribution of this chapter is that we investigate how one can integrate a new source into an existing immersive environment with finite information of the environment. We study what is required to make a new sound source integrate into an acoustic scene so that people can perceive the new source as a natural component of the acoustic scene and in the correct direction. In this chapter we assume the head is in a fixed location. Through listening tests, we found at least third order ambisonics is required to integrate a new source. In addition, a finite number of early reflections can perform equally well to a full room impulse response when a new source is added to an existing scenario. However, only correct reverberation time is not sufficient.

The chapter is organised as follows. In section 5.2 we describe our hypothesis of the

integration of virtual objects in an acoustic scene. In section 5.3, we discuss our experiments in detail and analyse the results. Finally, we conclude this chapter in section 5.4.

5.2 Integrating a virtual source

When we describe a soundfield, what is the necessary information of a room to make the sound natural and believable? We focus on the acoustics-only scenario, which implies that we omit the visual part of VR systems. The specific problem that we study here is the integration of a new sound source into an existing acoustic scenario. In a VR system, we already have an immersive environment. When we want to add a new source, like a virtual cat, we want to know what is required to make the new source perceptually plausible.

We study what aspects we can hear when we make specific modifications to a given acoustic scene. There exists a set of possibly relevant perceived attributes of a sound source in a room, such as room geometry and direct path direction to the source. In this chapter, we focus on the order of ambisonics signals, reflection order, and reverberation time. When we consider the reverberation time, we also take the direct path distance, direct path direction and room size into account. We discuss these selected attributes for integration separately below.

An important question is what order of ambisonics signals is necessary to make the integration of a new sound object believable. The most commonly used ambisonics signals are first-order ambisonics signals and third-order ambisonics signals. For head related transfer functions, [347] shows that an ambisonics order as low as four is sufficient, which indicates people do not perceive fine details during listening. Does this suggest that we do not need ambisonics signals of high order, such as order seven, to reproduce the soundfield? However, it is reasonable to explore the accuracy of the commonly used first order ambisonics and third order ambisonics. As discussed, when (2.36) is truncated to a particular N, the sound field will be accurate within a spherical region near the origin, which is commonly called the *sweet zone*, the size of which relates to the diameter, frequency and speed of sound.

Consequently, for third order ambisonics signals, if we assume the diameter of our head is 0.1m, the sound is correctly rendered at our ears up to 1600 Hz, which is too small comparing with the human hearing range. In addition, lower order ambisonics signals results in low angular resolution of soundfield reproduction. Is first-order or third-order enough for a believable VR system? Our hypothesis is that ambisonics signals of lower than order three are not sufficient for a believable VR system.

An important question with respect to reflections is whether we can use direct sound and a finite number of early reflections to replace the room impulse response to make a new sound source integrate into an existing acoustic scene. With an increasing number of reflections, the computational load of room impulse response increases [340]. Since real-time soundfield reproduction is required for a VR system, the computational load is a

significant problem although efficient algorithms exist [129, 341, 342]. The room impulse response is composed of direct-direction sound, early reflections, and late reverberation. Early reflections are relatively sparse first echoes and influence the spatial impression [348, 349]. Late reverberation is a dense decayed succession of echoes [350] and can degrade automatic speech recognition [351]. It is unclear if the late reverberation makes a difference when we integrate a new sound source into an existing scenario. Our hypothesis is that we can use direct sound and a finite number of reflections to replace the room impulse response and still obtain perceptually acceptable integration.

Reverberation time is considered to be one of the important attributes in acoustic environment simulation. We study the question if this measure is sufficient for the integration. It is commonly quantified in the form of Sabine's formula as (2.33). From (2.33), reverberation time is related to the room volume, surface area, and surface absorption. However, it does not vary with the positions of the sources and listeners.

If we only have correct reverberation time when we integrate a new source, is it sufficient? We divide the problem into two categories to examine the room volume, surface area, and direct path length. Firstly, for a fixed reverberation time and fixed room geometry, we want to know if different positions affect the integration. We assume we have one room impulse response of a room, which is generated with a fixed reverberation time. If we use this room impulse response, we only replace the direct path with the true direct path and keep other pathways fixed, is it perceptually acceptable for a VR system? Moreover, is the distance or the direction of the direct path important? Our hypothesis is that a room impulse response with a correct reverberation time and a correct direct path is sufficient to integrate a new sound source. Secondly, for a fixed reverberation time, we are interested if listenerscan hear the effect of different room sizes. We hypothesise that listeners can hear the difference in the different room sizes.

5.3 Experiments

We conducted listening tests to answer the questions asked in section 5.2. We used the control variates method to determine the quantitative relevance of the above selected attributes and used statistical analysis to analyse the experimental results. We first describe our experimental setup in the first subsection. We then present our experimental results and finally discuss these results.

5.3.1 Experimental setup

In this subsection, we give a general description of our experiments. Each artificial scenario lasts for ten seconds. In each scenario, there was one woman speaking in an empty rectangular room for four seconds. Then we added another woman as a new source to speak in this scenario, which lasts for six seconds and whose location is chosen randomly.

We choose the room size to be $6 \times 4 \times 3$m and the acoustic environment was modelled

by the image-source method [13]. We used the room impulse generator of [76] for our experiment. The speed of sound was set to $c = 342$ m/s. The reverberation time RT_{60} was set to be 0.4 s. We used HRTFs from MIT Media Lab [352]. The headphone used for the listening test was Beyerdynamic^{TM} DT 990 pro.

Ambisonics signals of order nine were used to reproduce the soundfield as a reference. We first resampled the input wav file with 16 kHz. After resampling, we constructed a four times oversampled Gabor frame and applied square-root Hann windows to satisfy the condition of perfect reconstruction. Based on the stationarity of the source signal and the length of room impulse, we chose a window support of 32 ms, which corresponds to 512 samples.

We used the commonly used audio rendering technique. We simulated playback over a given physical loudspeaker array, where each virtual loudspeaker signal is filtered with appropriately adjusted head related transfer functions (HRTFs) [195]. In our experiments, 598 secondary sources were used, the layout of which was same as that used for the HRTF database . We assumed the radius of a human head is 0.1 m and the center of the listener's head was located at $(3, 2, 1.7)$.

There were twelve participants for the experiments, which included two women and ten men. The subjects were not experts in spatial audio. Test subjects were allowed to listen to each scenario multiple times and change the volume in between. The experiments lasted approximately 30 minutes overall. The subjects answered questions for 16 scenarios. For each scenario they were required to answer if the new source is in the same scenario in the reference scenario and point out the azimuth and elevation of the new source with our user interface (the angular resolution is 10 degrees).

5.3.2 Description of Experiments

We conducted three sets of experiments to examine the three selected attributes, i.e., reflection order, the order of ambisonics signals, and reverberation time. We describe these three sets of experiments in detail below.

Our first experiment aimed to examine the relationship between the integration quality and the order of ambisonics signals. The reference scenario was reproduced with ninth-order ambisonics signals. The new source to be added to the scenario was reproduced with ambisonics signals of order one, three, five, seven, and nine respectively.

Our second experiment examined the influence of reflection order. In the reference scenario, the length of the room impulse response was set to be 340 ms, which included the early reflections and the late reverberation. To simplify the notation, we refer to the 340 ms room impulse response as *full response*. To examine the necessary reflection order, we changed the reflection order of the new sound source as zero, one, five, and nine. In addition, the full response was added as a contrast.

Our third experiment aimed at studying reverberation time. We first computed one

room impulse response with the predefined reverberation time and a random position in the room, which is referred as the measurement point later. We assumed the measurement point is 1 m distant from the listener.

We only changed the direct path signal in room impulse responses according to the source positions. Four modified room impulse responses were used to convolve with the new source at four different positions. Two of the positions (position 1 and 2) are at the same direct path distance as the measurement point (1 m) but with two different direct path directions. One (position 3) is nearer to the listener than the measurement point (0.7 m) and one (position 4) is farther (1.4 m).

In addition, we investigated if room impulse response with correct reverberation time and incorrect room size can integrate a new sound source into an existing scenario. Hence, we changed the size of the room to $4\text{m} \times 2\text{m} \times 3\text{m}$ and to $8\text{m} \times 6\text{m} \times 3\text{m}$ and computed the corresponding room impulse responses.

5.3.3 Statistical analysis

We used the chi-square test to investigate if each test object is sufficient for integration. The full response case is our reference for integration. Since eight out of 12 people answered "yes" to this full response case, our null hypothesis is the source is considered to be integrated into the existing scenario where we expect eight out of 12 people answered "yes". The critical value is 2.706 with level of significance $\alpha = 0.10$ of a 1 degree of freedom test. When the computed value exceed the critical value, we can reject the null hypothesis. Consequently, if there are less than six out of 12 test subjects who answered "yes", we can claim that the corresponding information is not sufficient in terms of integration.

5.3.4 Experimental result and discussion

In this subsection, we present our experimental results. The experimental results of integration problem is shown in Figure 5.1, where we show the number of "yes" responses for each case and the error bar represents the Wilson scored interval for a 95% confidence interval. In addition to the integration, we are also interested in the localisation accuracy when a new source is integrated into an existing scenario. The mean absolute error is shown in Figure 5.2 and the error bar represents the standard deviation.

When we observe the experimental results of the order of ambisonics, in terms of integration, ambisonics signals of order three to nine are sufficient to reproduce the sound field. We can conclude that an ambisonics order as low as three is sufficient for integration. Ambisonics of order nine shows lower elevation localisation accuracy than ambisonics of order five and seven, which may result from that late reverberation is clearer with ninth-order ambisonics signals and it can reduce the localisation accuracy.

As for the reflection order, we conclude that if a new source is integrated into an existing scenario, reflection order nine or full response is sufficient. In addition, reflection

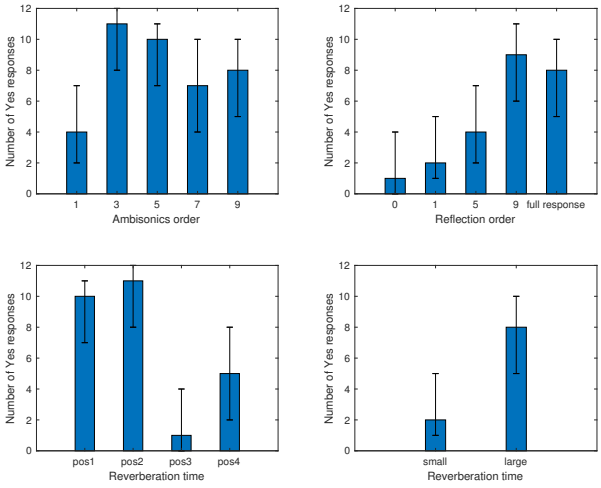


Figure 5.1: Integration experimental result.

order nine shows approximately equal localisation accuracy as full response. We found that localisation accuracy depends on source location. While we not consider this effect in the present paper, this explains the differences in the ambisonics and reflection order experiments. To conclude, a finite number of reflections can replace the full room impulse response in terms of integration.

When we observe the experimental result of reverberation time, we conclude that a room impulse response with only correct reverberation time is not sufficient to guarantee good integration. Only with the same direct path distance, the source is perceived to be in the same scenario. Similar to the reflection order experiments, we claim that listeners can approximately point out the correct direction of the integrated new source. Combining this result with the results of a preliminary suggests that when the room size is larger than the reference room but smaller than twice reference room size, listeners perceive the new sound source as integrated into the existing scenario and the localisation is also relatively accurate.

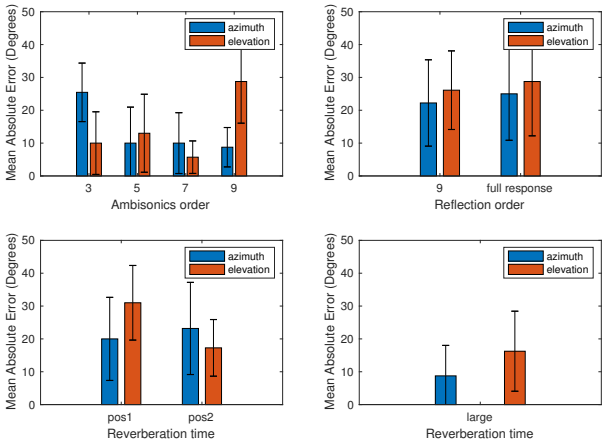


Figure 5.2: Localisation accuracy.

5.4 Conclusion

In this chapter, we used ambisonics signals to reproduce soundfield. We conducted a series of listening tests to examine the necessary information to integrate a new sound source into an existing acoustic scene and analysed the accuracy of localisation. We arrive at three conclusions. Firstly, with ambisonics signals of order three or higher, a new source can be integrated into an existing scenario. Secondly, a finite number of early reflections, for example ninth order reflection, can perform equally well in terms of integration and localisation as full room impulse responses. Finally, only using correct reverberation time to generate room impulse responses is not sufficient for integration and accurate localisation. To add a new source into an existing scenario, more information is required, such as direct path distance.

6

Ambisonics Room Impulse Response Generation from Omnidirectional Room Impulse Response Using Deep Neural Networks

Mapping a room impulse response to its ambisonics representation is not always feasible. However, by adding a weak assumption, i.e., the existence of at least two perpendicular walls in the environment, the ambisonics representation is restricted to be one of a finite set, with known transformations between the set entries. This makes the mapping of the omnidirectional response to the ambisonics response possible. We solve the mapping problem with a convolutional neural network and a multi-task variational autoencoder. Our method requires only a single room impulse response and obviates the need for specialised hardware for ambisonics measurement. The proposed method can achieve 17.62 *dB on estimated first order ARRs and* 16.15 *dB on estimated third order ARRs.*

6.1 Introduction

Augmented reality (AR) is a specific immersive audio-visual environment that provides users an interactive and enhanced experience in the real world with added artificial objects [15]. It can be used in various applications, such as education and entertainment. Spatial audio, aiming to create a 3D audio experience, is a major attribute of a believable AR system. As a consequence, the description and reproduction of the acoustical environment are of great importance. Omnidirectional room impulse responses (RIRs) are commonly used to describe the room acoustical environment. Ambisonics room impulse responses (ARRs) can provide spatial information that relates to room geometry, the positions of the sound source and the receiver, and the reflection coefficients, which are not explicitly described by RIRs. Hence AR commonly makes use of the ARR. Measurement of ARRs requires specialised equipment. The most commonly used equipment is a B-format microphone [19], which is only capable of first order ambisonics signals.

Estimation of the ambisonics representation from an impulse response is generally infeasible. For example, in a free space without floor, with the ARR coordinates centered at the receiver, the RIR is invariant with movement of the source on a sphere (surface of a ball). Hence the RIR provides no directional information that can be used to map it to an ARR. Perhaps surprisingly, with very weak prior information about the environment, i.e., the existence of at least two perpendicular walls, the problem becomes solvable. Our method requires the understanding of degeneracy. We select one particular mode to perform the estimation of the ARR directly from the omnidirectional RIR and then transform among different modes based on the available side information such as an image or the known location of a particular wall.

A RIR signal contains information about the configuration of the room acoustical environment implicitly [353]. Hence we hypothesise that it is possible to estimate the ambisonics room impulse response from an omnidirectional RIR. Since there exists a significant difference between real measured RIRs and simulated RIRs due to the scattering and diffraction [128], the analytical method based on the simulated RIRs [353] does not perform well for real world measurements. However, neural networks perform well for room measurement estimation in real-world conditions [338] and it is reasonable that this will also be the case for computing the ARR from the RIR. In addition, the commonly used B-format microphone can only capture first order signals and the spatial resolution of first order ambisonics is low. As a result, there exists work [171, 178] that upscales the ambisonics representation from first order to improve the sound quality. RIRs can be considered to be zeroth order ARRs and can be measured with a normal microphone, which has a low cost compared to a B-format microphone. The estimation of ARRs from RIRs can be interpreted as an ambisonics upscaling from zeroth order, which requires simpler input signals than [171, 178]. In this paper we show that machine learning allows us to estimate the ARR of any order directly from the omnidirectional RIR, thus obviating the need for

specialised hardware.

Ambisonics [16–18] is a soundfield reproduction technique that is suitable for AR systems as head rotations are easily modelled as the rotation of sound fields in the spherical harmonics domain. It describes the sound field by means of a small set of temporal signals. ARRs can be used to generate ambisonics signals by convolving with source signals [162, 163]. Recent work on ambisonics often uses higher order ambisonics (HOA), which is an extension of the original first-order ambisonics system developed by Gerzon [16]. Ambisonics is used for spatial audio encoding, transmission and as a basis for rendering. With an ambisonics representation of sufficient order, a high quality audio rendering system can give listeners a realistic spatial audio experience. A number of techniques can be used for binaural rendering of ambisonics [171, 178, 346].

The main contribution of this paper is the ARR estimation from RIRs using deep neural networks. As mentioned, generating an ambisonics representation from an omnidirectional signal is not always feasible. We show this mapping is possible in a room. The feasibility relies on the degeneracy of RIRs in a room. Our novel method only requires a single room impulse response without additional information if we only want to estimate ARR and reproduce the immersive environment. If we want to apply the estimated ARR in an audiovisual environment, such as AR, we need additional information, for example an image, to determine which mode it belongs to and the alignment between the coordinates of the image and the ARR. Our method is based on the image source method [13], which is sufficient for plausible augmented reality generation.

The paper is organised as follows. We review the relevant background in section II. In section III, we formulate the ambisonics room impulse response estimation problem. We then describe the ambisonics RIR estimation with convolutional neural networks in section IV. In section V, we use VAEs to generate the ambisonics RIR. The experimental results are discussed and analyzed in detail in section VI. Finally, we conclude our paper in section VII.

6.2 Problem Definition

In this section, we formulate the problem we aim to solve, i.e., ambisonics room impulse response estimation from an omnidirectional room impulse response. As noted in section II, an Ambisonics Room Impulse Response (ARR) is defined as an ambisonics representation of the corresponding room impulse response. The outcome of our work is a plausible auralization of a room with a simple measurement. We analyse the degeneracy of a RIR in the first subsection. In the second subsection, we discuss our motivation for using deep learning to solve the problem, describe how we compute the ambisonics room impulse responses, and discuss how to estimate the signals.

6.2.1 Degeneracy

Computing an ambisonics representation of an omnidirectional signal only is not always feasible. We have to add constraints to make the computation possible. We first define degeneracy.

Definition 2. *A RIR is M-fold degenerate if, given a set of coordinates, there exist M distinct ARRs that correspond to the RIR.*

The degeneracy often is finite, and the degeneracy can be removed by information from other modalities such as cameras or radar. We assume that the walls are either parallel or perpendicular; one side of a single wall defines the considered space and parallel walls enclose the considered space. Without loss of generality, the axes are assumed to be parallel to existing walls and the receiver is assumed to be located at the point of origin.

We start with discussing the degeneracy of impulse responses under different acoustic scenarios:

1. *Free space without walls, ceilings, or floors*: The impulse response is composed of a single delta pulse of the direct path. As long as the distance between the source and the receiver is the same, the RIR is the same. So there exists an uncountably infinite degeneracy for ARRs in this case.

2. *One wall, i.e., free space with a floor, or a pair of parallel walls*: A rotation of the source with respect to the receiver along the axis orthogonal to the wall or walls does not affect the RIR and hence corresponds to an infinite-fold degeneracy. Mirroring of the room introduces another two-fold degeneracy. In addition, if we exchange source and receiver, the RIR does not change, which introduces another two-fold degeneracy. For clarity, there exist infinite×4-fold degeneracy for ARRs in this case. If the direct path is parallel to the wall, the corresponding two-fold degeneracy collapses.

3. *Two perpendicular walls or two pairs of parallel walls, i.e.,* 2*D room case*: Mirroring of the room gives a four-fold degeneracy. We can additionally exchange the dimensions of the room, which introduces another two-fold degeneracy. The exchange of source and receiver gives another two-fold degeneracy. There is a 16-fold degeneracy for RIRs in total in this case. If the direct path is parallel to a wall, the corresponding degeneracy collapses. Similarly, if the room is square, the degeneracy of the x-y axis choice collapses. If the room is symmetric around the source, the degeneracy is identical to the one wall case.

4. *Three perpendicular walls or three pairs of parallel walls, i.e.,* 3*D room case*: Mirroring of the room gives an eight-fold degeneracy. The permutation of room dimensions introduces another six-fold degeneracy. Considering the exchange of source and receiver gives another two-fold degeneracy. There is 96-fold degeneracy for RIRs in

total in this case. A different level collapse of degeneracy happens when the direct path is parallel to a wall or the length of two perpendicular walls is identical.

From the analysis of the degeneracy of RIRs, we can conclude that by adding at least two perpendicular walls in the acoustic space, the problem is suddenly solvable at a cost of degeneracy. In the context of this paper, we assume a rectangular empty room, three edges of the room are of different length and the direct path is not parallel to any wall. Although there exists a 96-fold degeneracy for an empty rectangular 3D room, we can still make the ambisonics representation feasible by choosing one default mode out of a 96-fold degeneracy to solve the problem and subsequently mapping from that mode to another. We assume the direct path is always from straight ahead since head rotations are easily modelled as the rotation of sound fields in the spherical harmonics domain. We assume we have no knowledge about the direction of arrival of reflections or the environment information, such as room geometry and reflection coefficients.

The degeneracy can also be determined by the first order reflections as described in [353] with consistent results. We derive the condition of one mode as an example and refer for further details to [353]. We first chose three plane coordinates. Let (L, W, H) denote the room geometry, (x_s, y_s, z_s) and (x_r, y_r, z_r) denote the coordinates of source and receiver. Let us assume the first reflection is the pulse that reflects on $x = 0$, which gives us a six-fold degeneracy. This is equivalent to

$$\begin{cases} \|-x_s - x_r\| < \|2L - x_s - x_r\| \\ (-x_s - x_r)^2 + (y_s - y_r)^2 < (x_s - x_r)^2 + (-y_s - y_r)^2 \\ (-x_s - x_r)^2 + (y_s - y_r)^2 < (x_s - x_r)^2 + (2W - y_s - y_r)^2 \\ (-x_s - x_r)^2 + (z_s - z_r)^2 < (x_s - x_r)^2 + (-z_s - z_r)^2 \\ (-x_s - x_r)^2 + (z_s - z_r)^2 < (x_s - x_r)^2 + (2H - z_s - z_r)^2 \end{cases} . \tag{6.1}$$

Similarly, assuming the next non-x direction first order reflection reflects on $y = 0$ gives us a four-fold degeneracy, which is

$$\begin{cases} \|-y_s - y_r\| < \|2W - y_s - y_r\| \\ (-y_s - y_r)^2 + (z_s - z_r)^2 < (y_s - y_r)^2 + (-z_s - z_r)^2 \\ (-y_s - y_r)^2 + (z_s - z_r)^2 < (y_s - y_r)^2 + (2H - z_s - z_r)^2 \end{cases} . \tag{6.2}$$

Then assuming the next non-x and non-y direction first order reflection reflects on $z = 0$ gives us a two-fold degeneracy, which is $\|-z_s - z_r\| < \|2H - z_s - z_r\|$. The exchange of the source and receiver gives us another two-fold degeneracy, where we can assume $0 < x_r < x_s$,

$0 < y_r < y_s$ and $0 < z_r < z_s$. The conditions for this mode can be summarised as

$$\begin{cases} 0 < x_r < x_s \\ 0 < y_r < y_s \\ 0 < z_r < z_s \\ x_r + x_s < L \\ y_r + y_s < W \\ z_r + z_s < H \\ x_s x_r < y_s y_r < z_s z_r \\ x_s x_r < (W - y_s)(W - y_r) \\ y_s y_r < (H - z_s)(H - z_r) \end{cases} . \tag{6.3}$$

6.2.2 ARR estimation with deep learning

The state-of-art method to acquire ambisonics signals is to use special and expensive equipment, which makes the measurement difficult. Due to limitations on the equipment, we can only acquire relative low order ambisonics, which results in a low spatial resolution. Another possible method to compute ambisonics signals is to first estimate the room acoustical parameters from given signals and then base the ambisonics computation on that. However, estimating room acoustical parameters from a single RIR is difficult especially for real-world measurements since correct reflections are hard to detect. This motivates us to design a method to compute an ambisonics representation of the RIR from only an omnidirectional RIR using deep learning since it does not require special equipment or the estimation of room acoustical parameters. Since the first channel of ARR signal corresponds to zero-th order ambisonics, it is a scaled version of the omnidirectional RIR which contains no directional information explicitly. Hence our problem can also be viewed as an ambisonics upscaling problem which upscales ARRs from zero-th order to an arbitrary order.

For our ARR estimation problem, as discussed in Section 6.2.1, the degeneracy of RIR makes it hard to learn with a deep neural network. As a result, we first choose one default mode, i.e., define a one-to-one relationship between ARRs and RIRs. The coordinate system of our default mode is determined based on the first order reflections as in our previous paper [353]. We discuss how to map from one mode to another in Section 6.3.3.

Our computation of ARRs is based on the image source method [11–13] since we can compute the directions of reflections with the image source method. Using the image source method, an ARR can be viewed as a composition of real sound source and image sources. Each image source can be viewed as a separate source. An ARR signal $B_n^m(t)$ can

be computed with (4.4) and (2.42):

$$B_n^m(t) = \sum_q Y_n^m(\theta_q, \phi_q) \times$$

$$\sum_{\mathbf{p},\mathbf{m}} \beta_{x_1}^{|m_x - q|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y - j|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z - l|} \beta_{z_2}^{|m_z|} \frac{\delta(t - \tau_{\mathbf{p},\mathbf{m}})}{d_{\mathbf{p},\mathbf{m}}}. \tag{6.4}$$

This allows us to generate a large scale database of arbitrary order with RIR-ARR pairs for deep learning.

6.3 Ambisonics Room Impulse Response Estimation Using Deep Learning

In this section, we describe deep learning based ambisonics room impulse response estimation. We first compute an ARR from an omnidirectional RIR under the default mode out of 96-fold degeneracy with CNN and VAE respectively in the first two subsections. After that, we discuss the transformation matrix of ARRs among different modes of RIRs. At the end of this subsection, we describe how we can apply our ARR estimation methods for real-world applications.

6.3.1 ARR estimation with convolutional neural network

The ARR estimation problem can be viewed as a regression problem. Let the pair of random vectors (X,Y) denote the input and output signals of a neural network. These two signals are of the same length. Specifically, in this paper, X is an $\mathbb{R}^d$-valued random variable that represents a RIR where d denotes the length of each RIR signal vector, and Y is an $\mathbb{R}^d$-valued random variable that represents the corresponding ARR of one channel under default mode. The learned continuous deterministic function h is defined as $\hat{y} = h(x)$ where $\hat{\cdot}$ labels an estimate and $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$ is a realisation of the random variable pair (X,Y). The loss function $l : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ measures the mapping error of h. We can then define the risk R of the model as

$$R = \mathbb{E}[l(h(X),Y)], \tag{6.5}$$

where the expectation $\mathbb{E}$ is calculated with respect to the joint distribution $f_{XY}(X,Y)$. Since the joint distribution $f_{XY}(X,Y)$ is unknown, we approximate the risk R of the model with the empirical risk R_{emp} on the training set:

$$R_{\text{emp}} = \frac{1}{m} \sum_{i=1}^{m} l(h(x_i), y_i), \tag{6.6}$$

where m denotes the size of training dataset and each (x_i, y_i) pair is one realisation of $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$ in the training dataset. In the context of our problem, we use the mean

square error (MSE) as the empirical risk since MSE is differentiable and measures the squared Euclidean distance between the estimated outputs and corresponding ground truth. The objective function to train our neural network is then defined as

$$l(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^{m} \| y_i - \hat{y}_i \|_2^2, \tag{6.7}$$

where $\| \cdot \|_2^2$ is the squared l^2-norm, m denotes the size of training dataset, $y \in \mathbb{R}^{m \times d}$ denotes the true ARR and $\hat{y} \in \mathbb{R}^{m \times d}$ denotes the corresponding estimated ARR of one channel.

A straightforward solution to the ARR estimation problem uses a feedforward neural network. We hypothesise the ARR can be estimated from an omnidirectional RIR without any additional information. We make this hypothesis because a RIR signal contains the room acoustical parameters [353], which are sufficient to estimate corresponding ARRs. Here we use a CNN with omnidirectional RIRs as input and the estimated ARR as output. Since our signals are in the time domain, all layers are one-dimensional. Our CNN is composed of convolutional layers and transposed convolutional layers, each followed by a batch normalisation layer and an activation function, except the last layer. The number of channels increases with depth in the convolutional layers and decreases with depth in the transposed convolutional layers. Instead of learning all channels with a single neural network, we learn one ARR channel each time.

6.3.2 ARR Estimation with a Variational Autoencoder

An implicit assumption for ARR estimation we made is that we are able to extract useful information from RIRs to estimate ARRs. Based on the image source method [13], as discussed in Section II and III, the RIR and ARR signal can be represented by a 15-dim feature vector that contains four features, i.e., 3-dim room geometry, 3-dim source position, 3-dim receiver position and 6-dim reflection coefficients. We expect the autoencoder can implicitly perform the image source method to estimate RIR and ARR and the inverse process to extract room acoustical parameters. However, this turns out to be a difficult task for a neural network, which will be shown in Section V. In our preliminary test, we found if we use a normal CVAE with a single decoder, it only focuses on part of the features and loses some important information required for estimating ARRs. A multi-task autoencoder can help the latent layer form a good representation [354, 355] and result in a more robust representation of the estimated ARRs. This motivates us to use a multi-task VCAE to analyse RIRs, estimate ARRs and extract these features. In addition, we are interested to see if the dimensionality of the latent layer corresponds to the known dimensionality.

An important question is how the dimensionality of the latent layer affects the performance of a VAE. In the state-of-art VAEs, there is no agreement on the optimal dimensionality of the latent space. The intrinsic dimensionality [356] of a signal refers to the minimum number of parameters necessary for generating the signal. Intrinsic dimensionality can

help with the redundancy estimation in the embedded space [357]. In our case, the intrinsic dimensionality of the RIR is 15 by definition. We will use experiments to analyse how the dimensionality of the latent layer affects the performance of different decoders under fixed information rate. U-net [358] can outperform the earlier models with connected bypass information. Similarly, a latent layer of our VAE, which is wider than 15 dimensions, can also provide some bypass information. Inspired by U-net, we hypothesise a VAE with wider latent layer improves the performance.

We use one encoder which takes RIRs as input. We use multiple decoders to perform different tasks. We have four decoders for estimating the four room acoustical parameters respectively. As discussed, the dimensionality of these four features is 15 in total. These four decoders are connected with the first 15 neurons of $\mu_\psi(x)$ to ensure that all the information is available. Empirically we found that it is difficult to extract RIRs and ARRs with high accuracy from the first 15 latent neurons alone. Hence the decoders for the RIRs and the ARRs use additional latent neurons that encode information that the RIR and ARR decoders find difficult to extract from the first 15 latent neurons alone. This is consistent with the notion that the decoders find it difficult to mimic the image source method and need additional redundancy in the latent layer to perform well.

6.3.3 Transformations among modes of RIRs

The degeneracy of RIRs implies that a different mode results in a different ARR. Hence it is of great importance that we are able to transform ARRs from one mode to another. The transformations between the modes are linear transforms. From a transformation point of view, as discussed in Section 6.2.1, the relationship among different modes can be classified into mirroring, rotation (i.e., the permutation of room dimensions), and exchange of source and receiver. We deal with each case separately.

The mirroring refers to mirroring with respect to $x = 0$, $y = 0$, and $z = 0$. To facilitate the mirroring transformation, we first write the spherical harmonics as direction cosines [359]

$$Y_l = k_l \cdot f_l(u_x, u_y, u_z) \cdot g_l(u_x^2, u_y^2, u_z^2), \tag{6.8}$$

where l is the ambisonics channel number (ACN) of spherical harmonics and can link to (n, m) in (2.34) as $l = n(n+1)+m$,

$$\begin{cases} u_x = \cos(\theta)\cos(\phi) \\ u_y = \sin(\theta)\cos(\phi) \\ u_z = \sin(\phi) \end{cases},$$

k_l is numerical, f_l takes the form of $u_x^a \cdot u_y^b \cdot u_z^c$ where a, b, c is either 0 or 1, g_l is a polynomial of u_x^2, u_y^2, u_z^2. The mirroring can be realised as below [359]. If we mirror the soundfield with respect to $x = 0$, then all terms with u_x are negated. Similarly, If we mirror the soundfield

with respect to $y = 0$, then all terms with u_y are negated and if we mirror the soundfield with respect to $z = 0$, then all terms with u_z are negated.

Rotation is implemented by multiplying the ARRs of all channels with a rotation matrix $\boldsymbol{Q}$. For simplicity, here we show only the rotation matrix for a first order ARR rotation around z xis. Rotation matrices for higher-order ambisonics and for rotation around the x and y axis can be found in [360]. Each element of the matrix $\boldsymbol{Q}$ is denoted as $Q_{n',n}^{m',m}$ and the first order rotation matrix takes on the form [360]

$$\boldsymbol{Q} = \begin{bmatrix} Q_{0,0}^{0,0} & Q_{0,1}^{0,-1} & Q_{0,1}^{0,0} & Q_{0,1}^{0,1} \\ Q_{1,0}^{-1,0} & Q_{1,1}^{-1,-1} & Q_{1,1}^{-1,0} & Q_{1,1}^{-1,1} \\ Q_{1,0}^{0,0} & Q_{1,1}^{0,-1} & Q_{1,1}^{0,0} & Q_{1,1}^{0,1} \\ Q_{1,0}^{1,0} & Q_{1,1}^{1,-1} & Q_{1,1}^{1,0} & Q_{1,1}^{1,1} \end{bmatrix}. \tag{6.9}$$

Rotating around z axis by an angle α corresponds to rotation matrix $\boldsymbol{Q}_Y(\alpha)$ and the element can be calculated as [360]

$$Q_{n',n}^{m',m}(\alpha) = \begin{cases} \cos(m\alpha) & \text{if } n = n' \text{ and } m = m', \\ \sin(m\alpha) & \text{if } n = n' \text{ and } m = -m', \\ 0 & \text{otherwise.} \end{cases} \tag{6.10}$$

Exchange of the source and receiver positions can not be achieved by a transformation. Consequently, given no prior information, we train two separate neural networks with different source-receiver position layout. That is, given an arbitrary input, we compute each ARR channel with two neural networks. We can then apply the above transformations to both ARRs to get all ambisonics under different modes of RIRs. If there exists prior information, we can use additional information to decide on which mode is the target one.

6.3.4 Practical Application

Augmented reality is one of the important applications of the ARR estimation problem. Due to the existence of degeneracy of the RIR in a rectangular empty room, in an audio-only environment, one RIR corresponds to the multiple ARRs given alignment of all coordinates with a wall orientation. In an augmented reality environment, if we want to add a virtual object at a position whose RIR is given, it is important to determine the one correct ARR that gives the user an immersive experience. Different methods can be used to determine the correct ARR and we will discuss them below. Different methods can combine together to increase the accuracy.

One method is to use sensors to estimate distances. We can choose from different kinds of sensors based on the resolution requirement and cost, such as radar sensors, LiDAR (light detection and ranging) sensors, ultrasonic sensors, and Bluetooth sensors. The basic underlying principle is similar, i.e., estimating the distance between the user to each wall

using return time. Knowing the distance to each wall or one wall from each pair of walls, and the relative position of the source, we can choose the correct mode easily and compute the correct ARR.

We can also use image analysis to determine the degeneracy. Image analysis can be used to determine the relative positions of the walls, the source, and the image. Visual Simultaneous Localisation and Mapping (vSLAM) [361] is one set of methods to locate the user with images only. It includes feature-based, direct, and RGB-D camera-based approaches. [362] proposed a pseudo-LiDAR representation that mimics the LiDAR signal but is converted from image based depth maps. This method avoids the usage of expensive LiDAR sensors and improved the state-of-art image based method significantly. [363] trained a machine learning model which takes the captured images as input and outputs the distance between the objects and the vehicle. After localisation from the images, we can determine one mode with the method with sensors.

6.4 Experiments

We present our experiments in this section. In the first subsection, we describe the setup of our experiments. We then discuss the experiments on ARR estimation from RIRs with CNNs and CVAEs in the second and third subsection. Finally, we discuss and compare different methods to estimate ARRs from RIRs.

6.4.1 Experimental Setup

In the following, we first discuss the database we used to train and test our model. After that, we describe the configuration of our neural networks and how we trained and tested them.

Database

To build the dataset, we used the ISM to simulate RIRs [76] and the methods described in section III to compute the corresponding ARRs. We refer to this dataset as a clean RIR-ARR dataset of empty rooms. The shape of the rooms is rectangular and the rooms are empty. The speed of sound was set to $c = 340$ m/s. The sampling frequency was set to 8000 Hz. The length of each RIR was truncated at 1024 because we expect an approximate 0.25 s RIR contains the direct path signal and early reflections in an indoor environment and some of the late reverberation. Each dimension of the room geometry, i.e., length $L\times$ width $W\times$ height H, was assumed to be iid between $6\times4\times2$ m and $8\times6\times4$ m, which covers moderate and small rooms. The reflection coefficient of each wall was simulated as iid between 0 and 1. We randomly placed one source and one receiver in each room under the constraint (6.3) that guarantees that there exist a one-to-one mapping function between RIRs and a default ARR. This prevents the possibility of a one-to-multi relationship that can not be learned by a neural network. In our experiments, the number of image-source method simulated

RIR-ARR pairs was 400000, which was divided into a training dataset, a validation dataset, and a test dataset with the ratio $7:2:1$.

Neural network description

In this part, we focus on the configuration of our neural networks for different objectives and the training and testing of our neural networks. We did an ablation study on network architecture, optimisation method, and hyperparameter tuning with a grid search as a preliminary experiment to choose suitable network architectures and hyperparameters. When some properties of the database changed, an ablation study was performed again on network architecture and hyperparameters with a grid search.

We used a GPU to train our neural network to estimate ARRs from RIRs. We chose the Adam optimizer [320]. Its learning rate was set to 0.001 and the coefficients used for computing running averages of the gradient and its square were set to $(0.9, 0.999)$. We set the maximum iteration epochs to 5000 and applied early stopping as regularisation in our model [321] to prevent overfitting and and to limit the computational effort. The MSE loss is recorded per epoch on the training set under training mode and the validation set under evaluation mode and early stopping was performed when the validation error increased in 100 successive epochs. In addition, mini-batch based training was used to increase computational efficiency [322]. The batch size was set to 100. After training, we set the model to evaluation mode and computed the MSE in the test set.

Network architecture of CNN Table 6.1 shows our CNN architecture and the corresponding parameters for the ARR estimation from RIRs, where b denotes the batch size. After each (transposed) convolutional layer, there are always a batch normalisation layer and a Leaky ReLU layer [323] as the activation function, which we do not list in the Table 6.1 since the output size is not affected.

Network architecture of CVAE For the multi-task learning with CVAE, we used the architecture of the encoder and the decoder and the hyperparameters that are presented in Table 6.2, Table 6.3, Table 6.4, where b denotes the batch size and v equals to the dimensionality of latent layer, w denotes the length of room acoustical parameters. Similarly, after each (transposed) convolutional layer, there are always a batch normalisation layer and a Leaky ReLU layer [323] as the activation function. For the multi-task learning with CVAE, λ and v in (2.55) were set to be 0.1 and the latent dimensionality respectively.

6.4.2 Experiments on ARR estimation from RIRs with CNN

In this subsection, we present experiments on first-order and third-order ARR estimation from RIRs based on a feedforward neural network.

Table 6.1: CNN architecture of ARR estimation from RIRs

Operation	Kernel Size	Stride	# Channels	Output Size
Input				$(b, 1024)$
Reshape				$(b, 1, 1024)$
Conv1D	16	2	32	$(b, 32, 503)$
Conv1D	4	1	128	$(b, 128, 500)$
Conv1D	6	2	512	$(b, 512, 248)$
Conv1D	8	3	512	$(b, 512, 81)$
Conv1D	6	1	1024	$(b, 1024, 76)$
Conv1D	6	2	4096	$(b, 4096, 36)$
Conv1D	1	1	4096	$(b, 4096, 36)$
ConvTranspose1d	5	2	1024	$(b, 1024, 75)$
ConvTranspose1d	4	1	512	$(b, 512, 78)$
ConvTranspose1d	6	2	128	$(b, 128, 160)$
ConvTranspose1d	7	1	64	$(b, 64, 166)$
ConvTranspose1d	3	3	16	$(b, 16, 498)$
ConvTranspose1d	16	2	4	$(b, 4, 1010)$
ConvTranspose1d	15	1	1	$(b, 1, 1024)$
Reshape				$(b, 1024)$

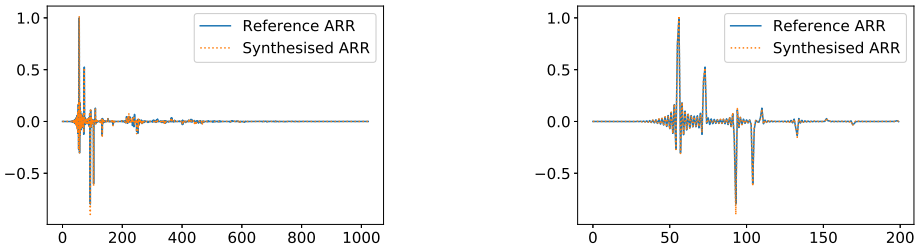
As the first experiment, we predicted first-order and third-order ARRs from RIRs. We evaluated the estimation performance with SNR and AMBIQUAL [364]. The SNR was measured on the ARRs directly. AMBIQUAL is an objective quality metric (range between 0 and 1 where 1 means a perfect match) proposed for ambisonic spatial audio, which estimates listening quality and localization quality from ambisonics. AMBIQUAL metric was shown by experiments to be strongly correlated to the subjective listening tests [364]. In the context of this paper, since B_0^0 is only a scaled version of the omnidirectional room impulse response, we are only interested in the localization quality. To obtain AMBIQUAL scores, the ARRs were convolved with ten anechoic recordings, which include six speech utterances from the TSP speech database [365] sound and four audio sound signals from the Audio/Video Anechoic Database [366]. We set the Intensity Binary Mask threshold of AMBIQUAL equal to -50 dB.

Following the convention of AMBIQUAL, we divide the ARRs into vertical channel, including B_1^0, B_2^0, and B_3^0, and horizontal channels. We average over vertical and horizontal channels respectively as the result for the vertical and horizontal channel. The experimental results are shown in Table 6.5, where the first-order ARRs outperform the third-order ARRs and horizontal channels outperform vertical channels. In addition, in Figure 6.1 and Figure 6.2, we show a channel of estimated first-order ARR with average SNR (17.3 dB) and a

Table 6.2: Network architecture of Encoder part of CVAE

Operation	Kernel Size	Stride	# Channels	Output Size
Input				$(b, 1024)$
Reshape				$(b, 1, 1024)$
Conv1D	16	2	32	$(b, 32, 503)$
Conv1D	4	1	128	$(b, 128, 500)$
Conv1D	6	2	512	$(b, 512, 248)$
Conv1D	8	3	512	$(b, 512, 81)$
Conv1D	6	1	1024	$(b, 1024, 76)$
Conv1D	6	2	4096	$(b, 4096, 36)$
Conv1D	1	1	4096	$(b, 4096, 36)$
Conv1D	1	1	128	$(b, 128, 36)$
Reshape				$(b, 128 * 36)$
Fully connected				(b, v)

channel of estimated third-order ARR with average SNR (13.6 dB) as a representative example for a visual impression of the signal quality, which is consistent with the SNR results in Table 6.5. The SNR and the figure show that the estimated ARR are reasonable. The AMBIQUAL score confirms the estimated ARR performs well in terms of localisation accuracy for an indoor environment with reverberation.



(a) Full signal.

(b) The first 200 samples.

Figure 6.1: An estimated first order ARR example with feedforward mapping.

6.4.3 Experiments on multitask-CVAE based ARR estimation

In this subsection, we present our experiments on CVAE based ARR estimation as described in section IV. We show the performance of the different decoders under different

Table 6.3: Network architecture of decoder part of CVAE (RIR reconstruction and ARR estimation)

Operation	Kernel Size	Stride	# Channels	Output Size
Input				(b, v)
Fully connected				$(b, 128 * 36)$
Reshape				$(b, 128, 36)$
ConvTranspose1d	1	1	4096	$(b, 4096, 36)$
ConvTranspose1d	5	2	1024	$(b, 1024, 75)$
ConvTranspose1d	4	1	512	$(b, 512, 78)$
ConvTranspose1d	6	2	128	$(b, 128, 160)$
ConvTranspose1d	7	1	64	$(b, 64, 166)$
ConvTranspose1d	3	3	16	$(b, 16, 498)$
ConvTranspose1d	16	2	4	$(b, 4, 1010)$
ConvTranspose1d	15	1	1	$(b, 1, 1024)$
Reshape				$(b, 1024)$

Table 6.4: Network architecture of decoder part of CVAE (room acoustical parameters)

Operation	Output Size
Input	(b, v)
Fully connected	$(b, 40)$
Fully connected	(b, w)

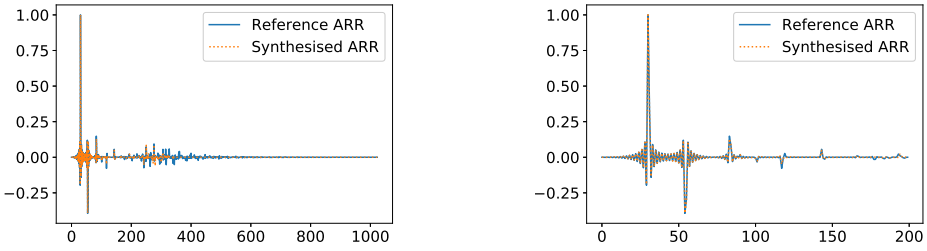
dimensionality of the latent space. In addition, we compare the performance of CVAE with feedforward mapping.

We performed the experiments on different dimensionality under the same information rate. The reference dimensionality was set to be 15 since we pre-assumed the features of a RIR can be described by a 15-dim vector as described in section IV. We also set the dimensionality of the latent space to 10, 30, 60, 80, 100, 200, and 400 for comparison. Our experiments indicate that, as long as each neuron of the latent layer can be allocated with more than one bit information rate on average, a higher information rate does not improve the experimental results. As a result, we set the information rate to 600 bits for training to make sure our multitask CVAE of different dimensionality have enough information rate. Although we use a multi-task autoencoder, we aim at synthesizing the ARRs. Consequently, we compared the different models based on the performance for the ARRs. Since the different ARR channels have similar performance, we used only channel B_1^0 to compare the different latent dimensionalities.

The relationship between latent dimensionality and the performance for the estimated ARR is shown in Figure 6.3. It shows that the model with latent dimensionality 200

Table 6.5: Experimental results of ARR estimation with CNN.

Signal	Channel	Test SNR (dB)	AMBIQUAL
First-order ARR	Horizontal	18.48	0.86
First-order ARR	Vertical	14.95	0.80
Third-order ARR	Horizontal	14.25	0.76
Third-order ARR	Vertical	10.79	0.69



(a) Full signal.

(b) The first 200 samples.

Figure 6.2: An estimated third order ARR example with feedforward mapping.

performed best on estimated ARR. It proved that a wider latent layer (before reaching the plateau) improved the results although the signal can be described by a 15-dimensional latent layer. This is consistent with the experiment in [367] where complex-valued and the magnitude and Instantaneous Frequency of the Short Time Fourier Transform result in a better performance than the time-domain waveform. Both our experiments and the results of [367] show that neural networks have difficulty learning some classes of complex relations.

The previous experiment show the model with latent dimensionality 200 is the best model. We presented experimental results of horizontal and vertical ARR channels with SNR and also performed the AMBIQUAL. Since we reconstructed RIR and estimated room acoustical parameters when we estimate each channel of ARR, we average over these estimates and compare with the ground truth in terms of SNR.

The experimental results on estimated ARRs are shown in Table 6.6. The method performs better on first-order ARR estimation than on the more difficult third-order ARR estimation. Horizontal channels outperform vertical channels, which like is related to the vertical room dimension being smaller. We also present the experimental results on reconstructed RIRs and estimated room acoustical parameters in Table 6.7. The multitask autoencoder structure also shows reasonable performance on these bypass tasks. In addi-

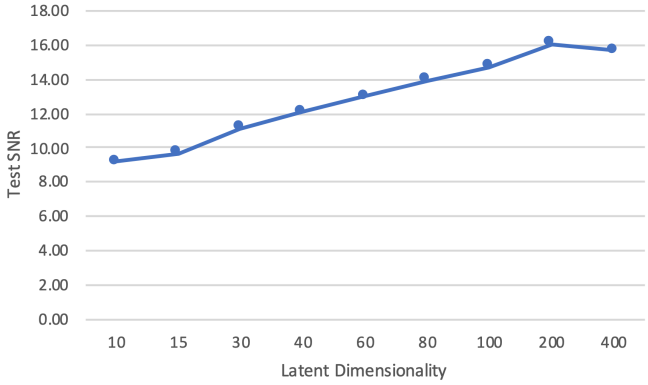


Figure 6.3: SNR of estimated ARR under different dimensionality.

Table 6.6: Experimental results of ARR estimation with CVAE.

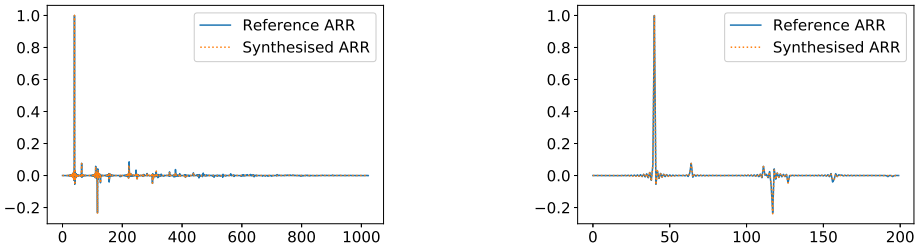
Signal	Channel	Test SNR (dB)	AMBIQUAL
First-order ARR	Horizontal	18.40	0.87
First-order ARR	Vertical	16.05	0.84
Third-order ARR	Horizontal	16.67	0.83
Third-order ARR	Vertical	14.11	0.80

tion, we plotted an example channel of estimated first order ARR with average SNR (17.62 dB) and third order ARR with average SNR (16.15 dB) as examples in Figure 6.4 and Figure 6.5 for a visual impression on the signal quality. From the SNR, the AMBIQUAL score and the figures, we can conclude that the performance of CVAE-based estimated ARRs is good.

At the end of this section, we compare the performance with the CNN in Table 6.5 and CVAE in Table 6.6. The SNR and AMBIQUAL both confirm that the CVAE-based method outperforms the CNN-based method, especially for the third order ARRs. As discussed in section 6.3.2, with our multi-task CVAE, we force it to focus on different features of RIRs and ARRs, then all important features are passed to the main decoder for the ARR estimation task. This helps to formulate a good representation of the latent layer and results in a more robust estimation of ARRs. Consequently, the CVAE-based method shows a better performance of ARR estimation than the CNN-based method.

Table 6.7: Experimental results of RIR reconstruction and the estimation of room acoustical parameters with CVAE.

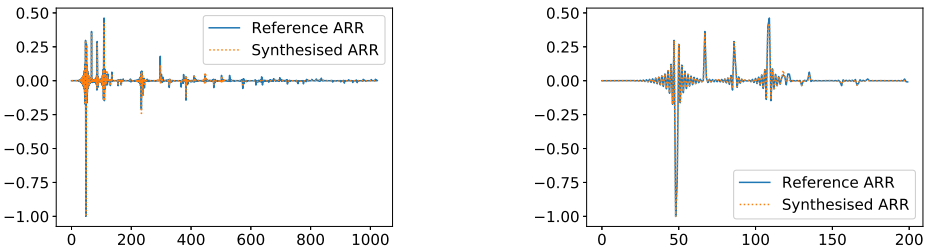
	Test SNR (dB)
RIR	22.5773
Receiver position	37.40
Source position	39.10
Room geometry	43.40
Reflection coefficients	21.36



(a) Full signal.

(b) The first 200 samples.

Figure 6.4: An estimated first order ARR example with CVAE.



(a) Full signal.

(b) The first 200 samples.

Figure 6.5: An estimated third order ARR example with CVAE.

6.5 Discussion and Conclusion

We introduced a method to estimate ARRs that performs well. We first note that the method can be made more robust if additional sensors are available.

If additional sensors can provide wall distances and relative source and receiver positions, we can compute the room geometry and the source and receiver positions. We can then use the image source method to generate a RIR and compare with the groundtruth RIR to determine the reflection coefficients. Finally, we can compute the ARR with these estimated features. The approach can be combined with the method presented in this paper to improve robustness if additional sensors are available.

In this paper, we showed it is possible to estimate ARRs from omnidirectional RIRs under the assumption that there exist at least two perpendicular walls within a finite set of degeneracy. The proposed method only requires a single RIR between a source and a receiver as input, which obviates the need of special measurement equipment. We used two methods to achieve this mapping, a feedforward mapping with CNN and a multi-task variance constrained autoencoder. We showed with experiments that the multi-task variance constrained autoencoder performs better than the feedforward mapping, especially for higher order ARRs, since the structure is more suitable for the estimation of ARRs. Future work focuses on the generalisation to the real-world measurements.

7

Conclusions and Future Work

This chapter concludes this dissertation. In addition, we address future research directions based on this dissertation.

7.1 Conclusions

Room impulse responses characterize sound propagation in an enclosed space. The room acoustic attributes lie in room impulse responses implicitly. One focus of this dissertation is the analysis of room impulse responses and the estimation of these room acoustic parameters from room impulse responses. Omnidirectional room impulse responses are insufficient to provide explicit spatial information for room acoustic applications, which is important for a realistic listening experience. Higher order ambisonics describes the sound field around the listener by a small set of temporal signals. It shows several advantages, such as, easy modeling of the head rotation and 360 degree capturing of sound field. Hence, ambisonics is also one focus of this dissertation. We conclude this dissertation by answering the research questions raised in Chapter 1.

Question 1. *How can we extract room acoustic parameters from a room impulse response? Can we analyze it using an analytical method or a deep learning based method? What are the differences between these two kinds of methods?*

Although an omnidirectional room impulse response can not always provide enough explicit spatial information, we found that it is possible to estimate room acoustic parameters from a single room impulse response without additional information. We used a deep learning based method and an analytical method to estimate room acoustic parameters from a room impulse response, respectively. In contrast to existing methods, we do not require knowledge of the relative positions of sources and receivers in the room. The proposed methods can be used with only a single RIR between one source and one receiver.

Our proposed deep learning based method can estimate room geometry and reflection coefficients from an omnidirectional room impulse response. Specifically, we used convolutional neural networks to estimate the room geometry and multilayer perceptrons to estimate the reflection coefficients. In addition, a convolutional neural network was used to link the reflection coefficients to the room geometry. The baseline method of room geometry estimation used a convolutional neural network, which takes a RIR as input and the room geometry as output. Two methods were proposed to improve the estimation accuracy. One improved method averaged over estimates from multiple RIRs. Another improved method restricted the relative position between the source and the receiver. In addition, the room geometry estimation method can be generalized to real-world measured RIRs. Due to the limited amount of real-measured data, we started with training the neural network with the RIRs simulated by the image source method. Since the image source method is an idealized model, we then augmented the simulated data by adding distortions to make the simulated data close to the real-world data. We applied transfer learning twice in total. We first apply transfer learning from the image source method simulated data to the distorted data. Next, we used transfer learning from the distorted data to the real-world measured data. The baseline method can achieve 0.038 m accuracy for each dimension on simulated data. For simulated data, the averaging method can achieve 0.025 m accuracy for

each dimension with 16 RIRs, and the semi-blind method can achieve 0.018 m accuracy for each dimension. For the real-world environments, the room geometry estimation achieved an average of 0.065 m accuracy for each dimension. We applied multiplayer perceptrons to estimate frequency dependent reflection coefficients on simulated RIRs and achieved an average of 0.09 accuracy. In addition, a convolutional neural network was used to link the reflection coefficients to the room geometry.

We proposed a two-step method to investigate how room acoustic parameters are estimated from an omnidirectional room impulse response. The first part is to train a transformer to estimate TOAs. The phase distortion can blur or bias the real TOAs of the peaks. In addition, when a peak is detected, it is difficult to determine whether it belongs to a specular reflection or other non-linear effects such as scattering. Consequently, we aim to estimate TOAs that work best for the room acoustic parameter estimation algorithm. Since the analytical method is based on the image source method and requires TOAs up to the second order, we estimated the TOAs of the direct path and specular reflections up to the second order that match the image source method. The proposed analytical method is based on the symmetry analysis of RIRs. The proposed analytical method is robust to erroneous pulses, non-specular reflections, and an unknown offset. The estimation on different dimensions is independent. For room geometry, source position, and receiver position, we achieved the RMSE of 0.0524 m, 0.0516 m, and 0.0641 m, respectively, with a failed portion of 25.9% of a real measured RIR database. The failed cases can be recognized by empty output or obvious bias of estimation, which can be reduced by repeated measurments.

We compare the two proposed methods based on the room geometry estimation since both methods estimate room geometry. The CNN based method estimates the room geometry directly from the room impulse responses. In contrast, the analytical method uses the estimated TOAs as inputs to estimate the room geometry. For the same real measured RIR database, the CNN based method achieved an RMSE of 0.065 m while the analytical method achieved an RMSE of 0.0524 m but with a failed portion of 25.9%. They shared similar accuracy on the real-world database. We hypothesize the small difference might result from the CNN based method randomly assigning reasonable estimates to the failed cases. The two methods prove that the room acoustic parameters lie implicitly in a single RIR.

Question 2. *What attributes are required for a new virtual acoustic source to be consistent with a pre-defined physical context?*

We investigated how one can integrate a new source into an existing immersive environment with finite information about the environment, aiming to let listeners perceive the new source as a natural component of the acoustic scene and in the correct direction. We used higher order ambisonics as our sound reproduction system to demonstrate our work. We assume the head is at a fixed location and focus on the acoustics-only scenario. There exists a set of possibly relevant perceived attributes of a sound source in a room. In particular, we considered the effect of the reflection order, the order of ambisonics, and the

reverberation time. A series of listening tests was conducted, and the chi-square test was used to determine whether each test object was sufficient for integration into the acoustic scene. We drew three conclusions from the listening tests. Firstly, ambisonics of order three or higher are required to integrate a new source into an existing scenario. Secondly, a finite order of early reflections can perform as well as full RIRs regarding integration and localization. Finally, it is insufficient to integrate a new source with correct reverberation time only.

Question 3. *Is it possible to estimate an ambisonics room impulse response from a single omnidirectional room impulse response?*

Although omnidirectional room impulse responses are insufficient for providing explicit spatial information for room acoustic applications, we show this spatial information is contained in RIRs. As indicated by research question 1, we can estimate room acoustic parameters from RIRs, and we should be able to estimate an ambisonics RIR from omnidirectional RIRs. Mapping from omnidirectional RIRs to ambisonics RIRs is not always feasible. The feasibility depends on the degeneracy of RIRs in an enclosed space. By adding a weak assumption, we restrict the ambisonics representations corresponding to a particular omnidirectional RIR to be a finite set with known transformations between the set entries. This allows us to map from omnidirectional RIRs to ambisonics RIRs. Two methods exist to estimate ambisonics RIRs, which will be summarized below.

The first method to estimate room acoustic parameters is analytical. It can be used to estimate an ambisonics room impulse response. Given an omnidirectional room impulse response, we first estimate the room geometry, source, and receiver position analytically. We can then use these parameters and the image source method to calculate the ambisonics RIRs. Similarly, if we can utilize additional sensors to measure the room geometry and source/receiver position, similar methods can be applied to estimate the ambisonics RIRs.

The second method is based on deep learning for room acoustic parameter estimation. We apply two kinds of deep learning model, i.e., convolutional neural networks and multi-task variational autoencoder. The CNN based method is the baseline method, which takes omnidirectional RIRs as input and estimated ARRs as output. Instead of learning all ARR channels together, each channel is learned by a separate CNN. We are interested in the intrinsic dimensionality of RIRs and ARRs. Since RIRs implicitly contain room acoustic parameters and can be determined by these parameters, we assume the RIR and ARR signal can be represented by a finite dimensional vector corresponding to the room acoustic parameters. We used a multi-task variational autoencoder to help the latent layer form a good representation instead of only focusing on the part of the features. The omnidirectional RIRs were taken as input of the encoder. We had six decoders in total, four of which connected only to the first 15 neurons of the latent layer to estimate room acoustic parameters and two of which connected to all latent neurons to reconstruct RIRs and estimate ARRs. This is because the dimensionality of room acoustic features is 15. Still, it is insufficient to output RIRs and ARRs of high accuracy with only these 15 neurons,

which implies it is difficult for decoders to mimic the image source method, and additional redundancy is required for good performance. We used experiments to investigate the optimal latent dimensionality, and experimental results showed latent dimensionality 200 performed best on estimated ARRs. In addition, the multi-task variational autoencoder based method performs better than the CNN based method, which can achieve a signal to distortion ratio of 17.62 dB on first order ARRs and 16.15 dB on third order ARRs.

7.2 Future Work

We have already answered the three research questions in this dissertation. This section suggests a few potential research directions based on this dissertation from different perspectives.

Modeling of Room Impulse Responses The room impulse response characterizes the sound propagation from a source to a receiver in an enclosed space, which is important in many room acoustic applications. Measuring real-world room impulse responses is time-consuming and requires specific hardware and procedures. As a result, a limited amount of real-measured room impulse responses exist, which can only cover a limited variability. The rapid progress of deep learning based research requires a large scale database, but the real-world measured data is insufficient. Several RIR simulation methods exist, which can be roughly categorized into wave-based methods and geometrical acoustic based methods. Wave based methods can model RIRs with high accuracy but face computational problems, especially for high frequencies. Hence the wave-based methods are not appropriate for generating a large scale database. Geometrical acoustic based methods are relatively computationally efficient, but many approximations are made, for example, the sound propagates as rays which neglects the wave properties. We use the image source method as an example to illustrate the limitation of the geometric based method. The image source method is most widely used for RIR simulation. However, it cannot simulate frequency dependent components, scattering, and diffraction. In addition, it cannot handle non-smooth surfaces and assumes empty rectangular rooms. Consequently, one potential research direction is to propose a new room impulse response simulation method that can accurately and efficiently simulate RIRs.

Modeling of Ambisonics Room Impulse Responses As discussed above, several RIR simulation methods exist, although each method has drawbacks. When we worked on the estimation of ARR, we were unaware of the existing method to simulate ARRs. On the one hand, ambisonics is essential to describe a 3D sound field. With the development of AR and VR, ambisonics plays an increasingly important role in spatial audio. On the other hand, measuring ambisonics signals is more difficult than omnidirectional signals. Spatial microphones, such as B-format microphones

and Eigenmike, are more expensive than normal microphones. In addition, spatial microphones face some problems, for example, the hard sphere affects the actual sound field. There exist very few ARR databases available online. Consequently, one potential research direction is a new method for ARR modeling. The ARR modeling method may, for example, be adapted from existing RIR modeling methods by including additional directional information.

Improvement on Room Acoustic Parameter Estimation In this dissertation, we used an analytical method and a deep learning based method to estimate room acoustic parameters, respectively. In addition, when we used a multi-task variational autoencoder to estimate ARRs, we also estimated room acoustic parameters as a byproduct. The room acoustics parameter estimation can be improved in different aspects. The deep learning based method can be improved by training on a database generated by a hybrid method instead of the image source method only. In addition, more recent deep learning models, for example, transformers and ResNets, can be used to replace the CNNs and improve the estimation accuracy of room acoustic parameters. It can also be improved to estimate room acoustic parameters in non-rectangular rooms, for example, an L-shape room. The analytical method was relatively sensitive to the time of arrival errors. This can be improved in different aspects. Including higher order reflections can improve the estimation accuracy if the TOAs of higher order reflections can be estimated. It is possible to integrate the room acoustic parameter estimation algorithm into the TOA estimation method and optimize the transformer using the estimated room acoustic parameters. Formulating the analytical method as an optimisation algorithm by including the multi-directional second order reflections may also benefit the room acoustic parameter estimation.

Ambisonics Room Impulse Response Estimation We estimated ARRs from omnidirectional RIRs using convolutional neural networks and multi-task variational autoencoders. We demonstrated our work in a rectangular room and did not consider many factors, such as scattering and frequency dependent reflective surfaces. A possible research direction is to investigate how imperfect room affects the ARR signals and how the scattering effect behaves in ARRs. These can provide a training database closer to real-world data and generalize our experiments. In addition, listening experiments can be conducted to verify the quality of estimated ARRs using the image source method only in a rectangular room.

Acknowledgments

Time flies, and my Ph.D. journey comes to its end. Looking back over the past several years, I want to express my appreciation to a number of people.

First of all, I would like to offer my deepest appreciation to my promotors Prof. dr. W. B. Kleijn and Prof. dr. ir. R. Heusdens. I would like to thank Prof. dr. ir. R. Heusdens for providing me the opportunity to the Ph.D. journey and bringing me into the world of room acoustics. Thank you for your encouragement and support when I faced difficulties. I would like to thank Prof. dr. W. B. Kleijn for teaching me how to do research, what is critical and analytical thinking, how to write research articles, how to face disappointments in life, etc. You showed me the possible directions and supported me when I could not see the future. What I learned from you will benefit my entire life. Thank you for your guidance, suggestions, criticism, patience, and encouragement.

I want to thank my office mates, Thomas Sherson, Pim van der Meulen, Elvin Isufi, Jie Zhang, Jamal Amini, Andreas Koutrouvelis, Aydin Rajabzadeh, and Jiani Liu. Thank you for your company and support in the past few years. I will never forget our Friday cakes, the jokes and laughs, and the discussions with you. You made the office feel like another home. I would also like to thank all the colleagues in the Signal Processing Systems group. I enjoyed the time with you during lunches, coffee times, and outings. I would like to thank Prof. Alle-Jan van der Veen for leading the group. I would like to thank Minaksie Ramsoekh, Irma Zomerdijk, Laura Bruns, Rosario Salazar Lozano, and Antoon Frehe for all the support. All members of the Signal Processing Systems group are such nice people who gave me enjoyable memories to my time in Delft.

I would like to thank all my friends, no matter where you are. Thank you for being there, although we could not always meet up. You always listen to me and cheer me up when I am frustrated. Our shared wonderful moments are the best candy in my life. Thank you for all the moments I share with you. You decorate my life and lighten up my journey.

Last but not least, I want to express my sincerest appreciation to my family. Many thanks to my parents. You always support me when I feel disappointed and always give me suggestions when I lose my direction. I learned so much from you, and you make me who I am. I am sorry that I have been away from you for so long. Thank you for flying ten hours to accompany me during your vacations. I really miss you every moment and hope we can get together in the near future. I would like to thank my grandparents. Although I could not be with you so much during the past few years, I still feel surrounded by your unconditional love. Ultimately, I want to thank my dear husband, Chongze, and my lovely

son, Yuhao. It is incredible to get married, be pregnant, and have a baby during the Ph.D. journey. I even could not imagine this myself when I started my Ph.D., but I am so grateful all this happened. You are my family in the Netherlands, and I never feel lonely because of you. Thank you for the life I share with you, whether it is bitter or sweet. I love you so much and look forward to the future with you.

Wangyang
's-Gravenhage, August 2023

Bibliography

References

[1] R. Aralikatti, A. Ratnarajah, Z. Tang, and D. Manocha, "Improving reverberant speech separation with synthetic room impulse responses," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 900–906.

[2] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.

[3] Wikipedia, "Reflection (physics)," 2022.

[4] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[5] D. Florencio and Z. Zhang, "Maximum a posteriori estimation of room impulse responses," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 728–732.

[6] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio engineering society*, vol. 50, no. 4, pp. 249–262, 2002.

[7] F. Seron, F. Sanz, M. Kindelan, and J. Badal, "Finite-element method for elastic wave propagation," *Communications in applied numerical methods*, vol. 6, no. 5, pp. 359–368, 1990.

[8] J. D. De Basabe and M. K. Sen, "Grid dispersion and stability criteria of some common finite-element methods for acoustic and elastic wave equations," *Geophysics*, vol. 72, no. 6, pp. T81–T95, 2007.

[9] L. L. Thompson, "A review of finite-element methods for time-harmonic acoustics," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1315–1330, 2006.

[10] H. Kuttruff, *Room acoustics*. New York: CRC Press, 2014.

[11] D. R. Begault and L. J. Trejo, "3-D sound for virtual reality and multimedia," 2000.

[12] S. G. McGovern, "Fast image method for impulse response calculations of box-shaped rooms," *Applied Acoustics*, vol. 70, no. 1, pp. 182 – 189, 2009.

[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[14] Wikipedia, "Immersion (virtual reality)," 2019.

[15] Wikipedia, "Augmented reality," 2019.

[16] F. Hollerweger, "An introduction to higher order ambisonic," 2013.

[17] D. Jerome and M. Sebastien, "Further study of sound field coding with higher order ambisonics," in *Audio Engineering Society Convention 116*, May 2004.

[18] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of Audio Engineering Society*, pp. 1004–1025, 2005.

[19] Wikipedia, "Soundfield microphone," 2021.

[20] R. Rabenstein and S. Spors, *Sound Field Reproduction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1095–1114.

[21] N. Ueno, S. Koyama, and H. Saruwatari, "Three-dimensional sound field reproduction based on weighted mode-matching method," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1852–1867, 2019.

[22] S. Koyama, K. Kimura, and N. Ueno, "Weighted pressure and mode matching for sound field reproduction: Theoretical and experimental comparisons," *J. Audio Eng. Soc*, vol. 71, no. 4, pp. 173–185, 2023.

[23] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," *Journal of the Audio Engineering Society*, may 2003.

[24] S. Spors, V. Kuscher, and J. Ahrens, "Efficient realization of model-based rendering for 2.5-dimensional near-field compensated higher order ambisonics," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 61–64.

[25] P. N. Samarasinghe, M. Poletti, S. M. A. Salehin, T. D. Abhayapala, and F. M. Fazi, "3D soundfield reproduction using higher order loudspeakers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 306–310.

[26] A. Omoto, S. Ise, Y. Ikeda, K. Ueno, S. Enomoto, and M. Kobayashi, "Sound field reproduction and sharing system based on the boundary surface control principle," *Acoustical Science and Technology*, vol. 36, no. 1, pp. 1–11, 2015.

[27] A. Berkhout, "A holographic approach to acoustic control," *Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, December 1988.

[28] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 05 1993.

[29] Wikipedia, "Wave field synthesis," 2022.

[30] J. Ahrens, R. Rabenstein, and S. Spors, "The theory of wave field synthesis revisited," *Journal of the Audio Engineering Society*, May 2008.

[31] J. Ahrens, *Analytic Methods of Sound Field Synthesis*. Berlin: Springer Science & Business Media, 2012.

[32] S. Spors, H. Wierstorf, M. Geier, and J. Ahrens, "Physical and perceptual properties of focused virtual sources in wave field synthesis," in *Audio Engineering Society Convention 127*, Oct 2009.

[33] J. Ahrens and S. Spors, "Local sound field synthesis by virtual secondary sources," *Journal of the Audio Engineering Society*, October 2010.

[34] N. Hahn, F. Winter, and S. Spors, "Local wave field synthesis by spatial band-limitation in the circular/spherical harmonics domain," *Journal of the Audio Engineering Society*, May 2016.

[35] B. Pueo, J. J. López, J. Escolano, and L. Hörchens, "Multiactuator panels for wave field synthesis: Evolution and present developments," *Journal of the Audio Engineering Society*, vol. 58, no. 12, pp. 1045–1063, December 2011.

[36] P.-A. Gauthier, A. Berry, and W. Woszczyk, "Sound-field reproduction in-room using optimal control techniques: Simulations in the frequency domain," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 662–678, 2005.

[37] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2992–3000, 1993.

[38] O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Local sound field reproduction using digital signal processing," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1584–1593, 1996.

[39] Y. J. Wu and T. D. Abhayapala, "Theory and design of soundfield reproduction using continuous loudspeaker concept," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 107–116, 2009.

[40] G. N. Lilis, D. Angelosante, and G. B. Giannakis, "Sound field reproduction using the Lasso," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1902–1912, 2010.

[41] Q. Feng, F. Yang, and J. Yang, "Time-domain sound field reproduction using the group Lasso," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. EL55–EL60, 2018.

[42] H. Chen and T. Abhayapala, "Spatial soundfield reproduction using deep neural networks," in *Proceedings of the 23rd International Congress on Acoustics*, 09 2019.

[43] S. Spors, H. Buchner, R. Rabenstein, and W. Herbordt, "Active listening room compensation for massive multichannel sound reproduction systems using wave-domain adaptive filtering," *The Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 354–369, 2007.

[44] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, 2005.

[45] R. Rabenstein, M. Renk, and S. Spors, "Limiting effects of active room compensation using wave field synthesis," *Journal of the Audio Engineering Society*, May 2005.

[46] L. Fuster, A. González, J. J. Lopez, and P. Zuccarello, "Room compensation using multichannel inverse filters for wave field synthesis systems," *Journal of the Audio Engineering Society*, May 2005.

[47] M. Poletti, F. M. Fazi, and P. A. Nelson, "Sound-field reproduction systems using fixed-directivity loudspeakers," *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3590–3601, 2010.

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

[49] T. Rajapaksha, X. Qiu, E. Cheng, and I. Burnett, "Geometrical room geometry estimation from room impulse responses," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 331–335.

[50] M. Coutino, M. B. Møller, J. K. Nielsen, and R. Heusdens, "Greedy alternative for room geometry estimation from acoustic echoes: A subspace-based method," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 366–370.

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

[52] Y. E. Baba, A. Walther, and E. A. P. Habets, "3D room geometry inference based on room impulse response stacks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 857–872, May 2018.

[53] M. Crocco, A. Trucco, and A. Del Bue, "Room reflectors estimation from sound by greedy iterative approach," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6877–6881.

[54] M. Crocco, A. Trucco, and A. Del Bue, "Uncalibrated 3D room geometry estimation from sound impulse responses," *Journal of the Franklin Institute*, vol. 354, 11 2017.

[55] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2683–2695, 2012.

[56] A. H. Moore, M. Brookes, and P. A. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *21st European Signal Processing Conference (EUSIPCO 2013)*, Sep. 2013, pp. 1–5.

[57] I. Dokmanić, Y. M. Lu, and M. Vetterli, "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)*, 2011, pp. 321–324.

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

[59] R. Parhizkar, I. Dokmanić, and M. Vetterli, "Single-channel indoor microphone localization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1434–1438.

[60] I. Dokmanić, L. Daudet, and M. Vetterli, "How to localize ten microphones in one finger snap," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2275–2279.

[61] S. Pasha and C. Ritz, "Informed source location and DOA estimation using acoustic room impulse response parameters," in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015, pp. 139–144.

[62] L. Remaggi, P. Jackson, and P. Coleman, "Source, sensor and reflector position estimation from acoustical room impulse responses," 2015.

[63] L. Kinsler, A. Frey, A. Coppens, and J. Sanders, *FUNDAMENTALS OF ACOUSTICS, 4TH ED.* Wiley India Pvt. Limited, 2009.

[64] M. R. Bai, J.-G. Ih, and J. Benesty, *Appendix: Acoustic Boundary Element Method*, 2013, pp. 501–511.

[65] T. Wu, "Boundary element acoustics fundamentals and computer codes," 2002.

[66] A. H.-D. Cheng and D. T. Cheng, "Heritage and early history of the boundary element method," *Engineering Analysis with Boundary Elements*, vol. 29, no. 3, pp. 268–302, 2005.

[67] N. Mohanan, R. Velmurugan, and P. Rao, "Speech dereverberation using nmf with regularized room impulse response," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4955–4959.

[68] M. Joorabchi, S. Ghorshi, and A. Sarafnia, "Single-channel speech dereverberation in acoustical environments," in *Proceedings ELMAR-2014*, 2014, pp. 1–4.

[69] M. Ochmann, "Exact solution for the acoustical impulse response of a line source above an absorbing plane," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1539–1549, 2018.

[70] N. Raghuvanshi, R. Narain, and M. C. Lin, "Efficient and accurate sound propagation using adaptive rectangular decomposition," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 789–801, Sep. 2009.

[71] J. Shragge and B. Tapley, "Solving the tensorial 3D acoustic wave equation: A mimetic finite-difference time-domain approach," *Geophysics*, vol. 82, no. 4, pp. T183–T196, 2017.

[72] S.-M. Sadrpour, V. Nayyeri, M. Soleimani, and O. M. Ramahi, "A new efficient unconditionally stable finite-difference time-domain solution of the wave equation," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 6, pp. 3114–3121, 2017.

[73] D. Murphy, *Digital Waveguide Mesh Topologies in Room Acoustics Modelling*. University of York, 2000.

[74] J. van Mourik, "Higher-order finite difference time domain algorithms for room acoustic modelling," November 2016.

[75] A. G. Prinn, "A review of finite element methods for room acoustics," *Acoustics*, vol. 5, no. 2, pp. 367–395, 2023.

[76] I. A. L. Erlangen, "RIR generator," 2014.

[77] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*. Society for Industrial and Applied Mathematics, 2011.

[78] A. G. Prinn, "On computing impulse responses from frequency-domain finite element solutions," *Journal of Theoretical and Computational Acoustics*, vol. 29, no. 01, p. 2050024, 2021.

[79] T. A. Davis, *Direct Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, 2006.

[80] N. Papadakis and G. E. Stavroulakis, "Time domain finite element method for the calculation of impulse response of enclosed spaces. Room acoustics application," *AIP Conference Proceedings*, vol. 1703, no. 1, 12 2015, 100002.

[81] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Society for Industrial and Applied Mathematics, 1994.

[82] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed. Society for Industrial and Applied Mathematics, 2003.

[83] S. Marburg and T.-W. Wu, *Treating the Phenomenon of Irregular Frequencies*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 411–434.

[84] S. Kirkup, "The boundary element method in acoustics: A survey," *Applied Sciences*, vol. 9, no. 8, 2019.

[85] F. C. Araújo, K. I. Silva, and J. C. F. Telles, "Generic domain decomposition and iterative solvers for 3d bem problems," *International Journal for Numerical Methods in Engineering*, vol. 68, no. 4, pp. 448–472, 2006.

[86] M. Bonnet, G. Maier, and C. Polizzotto, "Symmetric Galerkin boundary element method." *Appl. Mech. Rev.*, vol. 51, pp. 669–704, 1998.

[87] Y. Li, J. Meyer, T. Lokki, J. Cuenca, O. Atak, and W. Desmet, "Benchmarking of finite-difference time-domain method and fast multipole boundary element method for room acoustics," *Applied Acoustics*, vol. 191, p. 108662, 2022.

[88] T. Sakuma, S. Schneider, and Y. Yasuda, *Fast Solution Methods*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 333–366.

[89] I. Harari and T. J. Hughes, "A cost comparison of boundary element and finite element methods for problems of time-harmonic acoustics," *Computer Methods in Applied Mechanics and Engineering*, vol. 97, no. 1, pp. 77–102, 1992.

[90] S. Marburg and S. Schneider, "Performance of iterative solvers for acoustic problems. part i. solvers and effect of diagonal preconditioning," *Engineering Analysis with Boundary Elements*, vol. 27, no. 7, pp. 727–750, 2003, special issue on Acoustics.

[91] K. Kowalczyk and M. van Walstijn, "Room acoustics simulation using 3-d compact explicit fdtd schemes," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 34–46, 2011.

[92] S. Bilbao, "Modeling of complex geometries and boundary conditions in finite difference/finite volume time domain room acoustics simulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1524–1533, 2013.

[93] K. Kowalczyk and M. v. Walstijn, "Modeling frequency-dependent boundaries as digital impedance filters in FDTD and K-DWM room acoustics simulations," *J. Audio Eng. Soc*, vol. 56, no. 7/8, pp. 569–583, 2008.

[94] B. Hamilton and C. Webb, "Room acoustics modelling using GPU-accelerated finite difference and finite volume methods on a face-centered cubic grid," in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx)*, 2013, 16th International Conference on Digital Audio Effects Conference (DAFx-13) ; Conference date: 02-09-2013 Through 05-09-2013.

[95] B. Hamilton, C. Webb, A. Gray, and S. Bilbao, "Large stencil operations for GPU-based 3-D acoustics simulations," in *Proceedings of the 18th International Conference on Digital Audio Effects*. Norwegian University of Science and Technology, Nov. 2015.

[96] L. Savioja, "Real-time 3D finite-difference time-domain simulation of low- and mid-frequency room acoustics," in *DAFX the 13th Int. Conference on Digital Audio Effects (DAFx-10), Graz, Austria, Sept., 2010*, 2010, vK: Savioja.

[97] C. J. Webb and S. Bilbao, "Computing room acoustics with CUDA - 3D FDTD schemes with boundary losses and viscosity," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 317–320.

[98] K. Yee, "Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media," *IEEE Transactions on Antennas and Propagation*, vol. 14, no. 3, pp. 302–307, 1966.

[99] J. Saarelma, "Finite-difference time-domain solver for room acoustics using graphics processing units," Master's thesis, Aalto University, Nov. 2013.

[100] L. Gilles, S. Hagness, and L. Vázquez, "Comparison between staggered and unstaggered finite-difference time-domain grids for few-cycle temporal optical soliton propagation," *Journal of Computational Physics*, vol. 161, no. 2, pp. 379–400, 2000.

[101] J. He and M. Zhu, "Simulation of combined head and room impulse response based on sound ray tracing in frequency domain," in *IET International Conference on Smart and Sustainable City 2013 (ICSSC 2013)*, 2013, pp. 361–365.

[102] A. Alpkocak and M. K. Sis, "Computing impulse response of room acoustics using the ray-tracing method in time domain," *Archives of Acoustics*, vol. 35, pp. 505–519, 2010.

[103] C. Gu, M. Zhu, H. Lu, and B. Beckers, "Room impulse response simulation based on equal-area ray tracing," in *2014 International Conference on Audio, Language and Image Processing*, 2014, pp. 832–836.

[104] A. Alpkocak and M. Sis, "Computing impulse response of room acoustics using the ray-tracing method in time domain," *Archives of Acoustics*, vol. 35, no. 4, p. 505, 12 2010, copyright - Copyright Versita Dec 2010; Last updated - 2013-03-28.

[105] A. Krokstad, S. Strom, and S. SÃžrsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[106] J. P. Walsh and N. Dadoun, "The design and development of godot: A system for computer?aided room acoustics modeling and simulation," *The Journal of the Acoustical Society of America*, vol. 69, no. S1, pp. S36–S36, 1981.

[107] N. Dadoun, D. G. Kirkpatrick, and J. P. Walsh, "The geometry of beam tracing," in *Proceedings of the First Annual Symposium on Computational Geometry*, ser. SCG '85. New York, NY, USA: Association for Computing Machinery, 1985, p. 55?61.

[108] D. van Maercke and J. Martin, "The prediction of echograms and impulse responses within the epidaure software," *Applied Acoustics*, vol. 38, no. 2, pp. 93–114, 1993.

[109] T. Lewers, "A combined beam tracing and radiatn exchange computer model of room acoustics," *Applied Acoustics*, vol. 38, no. 2, pp. 161–178, 1993.

[110] A. Farina, "RAMSETE-a new pyramid tracer for medium and large scale acoustic problems," vol. 95. Proc. of Euro-Noise, 1995.

[111] A. Wareing and M. Hodgson, "Beam-tracing model for predicting sound fields in rooms with multilayer bounding surfaces," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2321–2331, 2005.

[112] I. A. Drumm and Y. W. Lam, "The adaptive beam-tracing algorithm," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1405–1412, 2000.

[113] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West, "A beam tracing approach to acoustic modeling for interactive virtual environments," *Proceedings of SIGGRAPH 98*, pp. 21–32, Jul. 1998.

[114] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, and A. Ngan, "A beam tracing method for interactive architectural acoustics," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 739–756, 2004.

[115] F. Antonacci, M. Foco, A. Sarti, and S. Tubaro, "Fast tracing of acoustic beams and paths through visibility lookup," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 812–824, 2008.

[116] H. Lehnert, "Systematic errors of the ray-tracing algorithm," *Applied Acoustics*, vol. 38, no. 2, pp. 207–221, 1993.

[117] A. Southern, S. Siltanen, D. T. Murphy, and L. Savioja, "Room impulse response synthesis and validation using a hybrid acoustic model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1940–1952, 2013.

[118] R. Tenenbaum, T. Camilo, J. C. Torres, and S. Gerges, "Hybrid method for numerical simulation of room acoustics with auralization: Part 1 -theoretical and numerical aspects," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 29, 06 2007.

[119] I. Drumm, "A hybrid finite element / finite difference time domain technique for modelling the acoustics of surfaces within a medium," *Acta Acustica united with Acustica*, vol. 93, no. 5, pp. 804–809, September 2007.

[120] M. Aretz, R. Nöthen, M. Vorlaender, and D. Schröder, "Combined broadband impulse responses using fem and hybrid ray-based methods," in *EAA Auralization Symposium 2009, Espoo Finland*, 06 2009.

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

[122] E. Deines, F. Michel, M. Bertram, J. Mohring, and H. Hagen, "Simulation, visualization, and virtual reality based modeling of room acoustics," 2007.

[123] J. PONGSIRI, P. AMIN, and C. Thompson, "Modeling the acoustic transfer function of a room," 01 1999.

[124] A. Ratnarajah, Z. Tang, and D. Manocha, "IR-GAN: room impulse response generator for speech augmentation," *CoRR*, vol. abs/2010.13219, 2020.

[125] M. A. Biot, "Generalized boundary condition for multiple scatter in acoustic reflection," *The Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1616–1622, 1968.

[126] M. VorlÃ€nder and E. Mommertz, "Definition and measurement of random-incidence scattering coefficients," *Applied Acoustics*, vol. 60, no. 2, pp. 187–199, 2000.

[127] Wendt, Florian and Höldrich, Robert, "Precedence effect for specular and diffuse reflections," *Acta Acust.*, vol. 5, p. 1, 2021.

[128] T. J. Cox, B.-I. Dalenbäck, P. D'Antonio, J.-J. Embrechts, J. Y. Jeon, E. Mommertz, and M. Vorländer, "A tutorial on scattering and diffusion coefficients for room acoustic surfaces," *Acta Acustica United With Acustica*, vol. 92, pp. 1–15, 2006.

[129] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 187–187, 2007.

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

[131] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja, "The room acoustic rendering equation," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1624–1635, 2007.

[132] H. Bai, G. Richard, and L. Daudet, "Modeling early reflections of room impulse responses using a radiance transfer method," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[133] M. A. Biot and I. Tolstoy, "Formulation of wave propagation in infinite media by normal coordinates with an application to diffraction," *The Journal of the Acoustical Society of America*, vol. 29, no. 3, pp. 381–391, 1957.

[134] R. R. Torres, U. P. Svensson, and M. Kleiner, "Computation of edge diffraction for more accurate room acoustics auralization," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 600–610, 2001.

[135] J. B. Keller, "Geometrical theory of diffraction*," *J. Opt. Soc. Am.*, vol. 52, no. 2, pp. 116–130, Feb 1962.

[136] P. H. Pathak, G. Carluccio, and M. Albani, "The uniform geometrical theory of diffraction and some of its applications," *IEEE Antennas and Propagation Magazine*, vol. 55, no. 4, pp. 41–69, 2013.

[137] R. Kouyoumjian and P. Pathak, "A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface," *Proceedings of the IEEE*, vol. 62, no. 11, pp. 1448–1461, 1974.

[138] R. Paknys, *Uniform Theory of Diffraction*, 2016, pp. 268–316.

[139] Y. Furue, "Sound propagation from the inside to the outside of a room through an aperture," *Applied Acoustics*, vol. 31, no. 1, pp. 133–146, 1990.

[140] N. TSINGOS, "Fast rendering of sound occlusion and diffraction effects for virtual acoustic environments," *Proc. Audio Engineering Society 104 Conv., May 1998*, 1998.

[141] N. Tsingos, C. Dachsbacher, S. Lefebvre, and M. Dellepiane, "Instant Sound Scattering," in *Rendering Techniques*, J. Kautz and S. Pattanaik, Eds. The Eurographics Association, 2007.

[142] J. Rindel, "Attenuation of sound reflections due to diffraction," in *Proceedings of the Nordic Acoustical Meeting*, 01 1986, pp. 257–260.

[143] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection localization from room impulse responses," *Acta Acustica united with Acustica*, vol. 98, no. 3, pp. 418–440, 2012.

[144] B. L. Sturm and G. DeFrance, "Detection and estimation of arrivals in room impulse responses by greedy sparse approximation," in *2010 18th European Signal Processing Conference*, 2010, pp. 1934–1938.

[145] I. J. Kelly and F. M. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1139–1147, 2014.

[146] L. Hirvisalo, S. Vesa, and T. Lokki, "Segmentation and analysis of early reflections from a binaural room impulse response," WorkingPaper TKK-ME-R-1, 2009.

[147] S. Vesa and T. Lokki, "Detection of room reflections from a binaural room impulse response," in *The 9th International Conference on Digital Audio Effects (DAFx'06), Montreal, Canada, September 18-20, 2006*, V. Verfaille, Ed. Canada: McGill University, 2006, pp. 215–220.

[148] R. C. Gonzales and R. E. Woods, *Digital Image Processing, 4th edition.* Pearson, 2018.

[149] D. M. Ristić, M. Pavlović, D. v. Pavlović, and I. Reljin, "Detection of early reflections using multifractals," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. EL235–EL241, 03 2013.

[150] D. Ristić, M. Pavlovic, M. Mijic, and I. Reljin, "Improvement of the multifractal method for detection of early reflections," *Serbian Journal of Electrical Engineering*, vol. 11, pp. 11–24, 2014.

[151] A. Prodeus and M. Didkovska, "Detection of early reflections in the room impulse response by estimating the excess coefficient at short time intervals," in *2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology*, 2021, pp. 1–6.

[152] N. R. Shabtai, Y. Zigel, and B. Rafaely, "Room volume classification from room impulse response using statistical pattern recognition and feature selection," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1155–1162, 2010.

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

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

[155] N. Antonello, M. Moonen, and P. A. Naylor, "Evaluation of a numerical method for identifying surface acoustic impedances in a reverberant room," in *Proc. of the 10th European Congress and Exposition on Noise Control Engineering*, 2015.

[156] A. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, 06 1989, vol. 34.

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

[158] M. Karjalainen, P. Antsalo, and T. Peltonen, "Estimation of modal decay parameters from noisy response measurements," *Journal of the Audio Engineering Society*, 2002.

[159] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Room boundary estimation from acoustic room impulse responses," in *2014 Sensor Signal Processing for Defence (SSPD)*, 2014, pp. 1–5.

[160] H. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, ser. Detection, Estimation, and Modulation Theory. John Wiley & Sons, Ltd, 2004.

[161] L. Remaggi, P. J. B. Jackson, W. Wang, and J. A. Chambers, "A 3d model for room boundary estimation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 514–518.

[162] K. Müller and F. Zotter, "Auralization based on multi-perspective ambisonic room impulse responses," *Acta Acust.*, vol. 4, no. 6, p. 25, 2020.

[163] P. Massé, T. Carpentier, O. Warusfel, and M. Noisternig, "A robust denoising process for spatial room impulse responses with diffuse reverberation tails," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2250–2260, 2020.

[164] Wikipedia, "Spherical harmonics," 2021.

[165] Wikipedia, "Ambisonic data exchange formats," 2015.

[166] M. Chapman, W. Ritsch, T. Musil, J. Zmölnig, H. Pomberger, F. Zotter, and A. Sontacchi, "A standard for interchange of ambisonic signal sets. including a file standard with metadata," in *Proc. of the Ambisonics Symposium, Graz, Austria*, Jan 2009, pp. 25–27.

[167] N. Epain and C. T. Jin, "Spherical harmonic signal covariance and sound field diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1796–1807, 2016.

[168] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2542–2556, June 2007.

[169] E. G. Williams, "Fourier acoustics: sound radiation and nearfield acoustical holography," London, 1999.

[170] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2002, pp. II–1781–II–1784.

[171] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 385–388.

[172] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 1–4.

[173] M. Samarawickrama, N. Epain, and C. Jin, "Super-resolution acoustic imaging using non-uniform spatial dictionaries," in *2014 International Conference on Audio, Language and Image Processing*, 2014, pp. 973–977.

[174] G. Routray, S. K. Sahu, and R. M. Hegde, "Upscaling hoa signals using order recursive matching pursuit in spherical harmonics domain," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2022, pp. 1–5.

[175] G. Routray and R. M. Hegde, "Sparse plane-wave decomposition for upscaling ambisonic signals," in *2020 International Conference on Signal Processing and Communications (SPCOM)*, 2020, pp. 1–5.

[176] N. Epain and C. Jin, "Super-resolution sound field imaging with sub-space preprocessing," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 350–354.

[177] G. Routray, P. Dwivedi, and R. M. Hegde, "Binaural reproduction of hoa signal using sparse multiple measurement vector projections," in *2021 National Conference on Communications (NCC)*, 2021, pp. 1–6.

[178] W. B. Kleijn, "Directional emphasis in ambisonics," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 1079–1083, 2018.

[179] M. Frank and F. Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," in *DAGA, Aachen*, 03 2016.

[180] L. Gölles and F. Zotter, "Directional enhancement of first-order ambisonic room impulse responses by the 2+2 directional signal estimator," in *Proceedings of the 15th International Audio Mostly Conference*, ser. AM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 38–45.

[181] G. Routray, S. Basu, P. Baldev, and R. M. Hegde, "Deep-sound field analysis for upscaling ambisonic signals," in *EAA Spatial Audio Signal Processing Symposium*, Paris, France, Sep. 2019, pp. 1–6.

[182] M. Frank, F. Zotter, and A. Sontacchi, "Producing 3D audio in ambisonics," *Journal of the Audio Engineering Society*, March 2015.

[183] M. A. Poletti, "A unified theory of horizontal holographic sound systems," *Journal of the Audio Engineering Society*, vol. 48, no. 12, pp. 1155–1182, December 2000.

[184] M. Frank, "How to make ambisonics sound good," in *Forum acusticum*, 09 2014.

[185] F. Zotter, H. Pomberger, and M. Noisternig, "Energy-preserving ambisonic decoding," *Acta Acustica United With Acustica*, vol. 98, pp. 37–47, 2012.

[186] D. Murillo Gomez, F. Fazi, and M. Shin, "Evaluation of ambisonics decoding methods with experimental measurements," in *Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany*, 04 2014.

[187] j. daniel, j.-b. rault, and j.-d. polack, "Ambisonics encoding of other audio formats for multiple listening conditions," *Journal of the audio engineering society*, September 1998.

[188] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807–820, October 2012.

[189] J. Trevino, T. Okamoto, Y. Iwaya, and Y. Suzuki, "High order ambisonic decoding method for irregular loudspeaker arrays," in *20th International Congress on Acoustics 2010, ICA 2010 - Incorporating Proceedings of the 2010 Annual Conference of the Australian Acoustical Society*, 2010, pp. 1050–1057.

[190] T. Qu, Z. Huang, Y. Qiao, and X. Wu, "Matching projection decoding method for ambisonics system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 561–565.

[191] L. S. Davis, R. Duraiswami, E. Grassi, N. A. Gumerov, Z. Li, and D. N. Zotkin, "High order spatial audio capture and its binaural head-tracked playback over headphones with hrtf cues," in *Audio Engineering Society Convention 119*, 2005/// 2005.

[192] G. Enzner, M. Weinert, S. Abeling, J.-M. Batke, and P. Jax, "Advanced system options for binaural rendering of ambisonic format," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 251–255.

[193] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3D ambisonic based binaural sound reproduction system," *journal of the audio engineering society*, June 2003.

[194] T. McKenzie, D. T. Murphy, and G. Kearney, "Diffuse-field equalisation of binaural ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, 2018.

[195] J. G. Tylka and E. Choueiri, "Comparison of techniques for binaural navigation of higher-order ambisonic soundfields," in *Audio Engineering Society Convention 139*, Oct 2015.

[196] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 823–828, 02 2010.

[197] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[198] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 05 2013.

[199] C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Improving binaural ambisonics decoding by spherical harmonics domain tapering and coloration compensation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 261–265.

[200] M. Kuster, "Multichannel room impulse response rendering on the basis of underdetermined data," *J. Audio Eng. Soc*, vol. 57, no. 6, pp. 403–412, 2009.

[201] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 982–993, 08 2008.

[202] C. Pörschmann and S. Wiefling, "Perceptual aspects of dynamic binaural synthesis based on measured omnidirectional room impulse responses," ser. ICSA 2015: 3rd International Conference on Spatial Audio, Graz, 18. - 20. Sep., 2015.

[203] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of omnidirectional room impulse responses-algorithm and technical evaluation," in *Proceedings of the DAFx*, 2017, pp. 345–352.

[204] C. Pörschmann and P. Stade, "Auralizing listener position shifts of measured room impulse responses," in *Proceedings of the DAGA*, 2016, pp. 1308–1311.

[205] C. Pörschmann, P. Stade, and J. M. Arend, "Binaural auralization of proposed room modifications based on measured omnidirectional room impulse responses," *Proceedings of Meetings on Acoustics*, vol. 30, no. 1, p. 015012, 2017.

[206] J. M. Arend, S. V. A. Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response," *J. Audio Eng. Soc*, vol. 69, no. 7/8, pp. 557–575, 2021.

[207] F. Menzer, C. Faller, and H. Lissek, "Obtaining binaural room impulse responses from B-format impulse responses using frequency-dependent coherence matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 396–405, Feb 2011.

[208] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," in *Audio Engineering Society Convention 144*, May 2018.

[209] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3d source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 18th European Signal Processing Conference*, 2010, pp. 442–446.

[210] P. Stade, J. Arend, and C. Pörschmann, "A parametric model for the synthesis of binaural room impulse responses," *Proceedings of Meetings on Acoustics*, vol. 30, no. 1, p. 015006, 2017.

[211] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, pp. 65–386, 1958.

[212] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, Dec 1989.

[213] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6231–6239.

[214] Y.-S. Park and S. Lek, "Chapter 7 - artificial neural networks: Multilayer perceptron for ecological modeling," in *Ecological Model Types*, ser. Developments in Environmental Modelling, S. E. Jørgensen, Ed. Elsevier, 2016, vol. 28, pp. 123 – 140.

[215] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 39, no. 1, pp. 43 – 62, 1997.

[216] R. Pearson, T. Dawson, P. Berry, and P. Harrison, "Species: A spatial evaluation of climate impact on the envelope of species," *Ecological Modelling*, vol. 154, no. 3, pp. 289 – 300, 2002.

[217] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5, pp. 183 – 197, 1991.

[218] M. H. Esfe, M. Afrand, S. Wongwises, A. Naderi, A. Asadi, S. Rostami, and M. Akbari, "Applications of feedforward multilayer perceptron artificial neural networks and empirical correlation for prediction of thermal conductivity of mg(oh)2–eg using experimental data," *International Communications in Heat and Mass Transfer*, vol. 67, pp. 46 – 50, 2015.

[219] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Networks*, vol. 1, no. 2, pp. 119 – 130, 1988.

[220] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[221] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.

[222] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843–4855, Oct 2017.

[223] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[224] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[225] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[226] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1725–1732.

[227] I. Aljarrah and D. Mohammad, "Video content analysis using convolutional neural networks," in *2018 9th International Conference on Information and Communication Systems (ICICS)*, April 2018, pp. 122–126.

[228] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *CoRR*, vol. abs/1503.08909, 2015.

[229] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[230] S. Jingzhou, W. Yongbin, and C. Xiaosen, "Audio segmentation and classification approach based on adaptive CNN in broadcast domain," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, 2019, pp. 1–6.

[231] T. V. Kumar, R. S. Sundar, T. Purohit, and V. Ramasubramanian, "End-to-end audio-scene classification from raw audio: Multi time-frequency resolution CNN architecture for efficient representation learning," in *2020 International Conference on Signal Processing and Communications (SPCOM)*, 2020, pp. 1–5.

[232] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4628–4632.

[233] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C.-H. Lee, "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1289–1300, 2017.

[234] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech Dereverberation Using Fully Convolutional Networks," *arXiv e-prints*, p. arXiv:1803.08243, Mar 2018.

[235] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, May 2017, pp. 1–5.

[236] N. Mamun, S. Khorram, and J. H. L. Hansen, "Convolutional neural network-based speech enhancement for cochlear implant recipients," *CoRR*, vol. abs/1907.02526, 2019.

[237] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *CoRR*, vol. abs/1609.07132, 2016.

[238] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[239] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[240] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[241] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning."

[242] L. Medsker and L. Jain, *Recurrent Neural Networks: Design and Applications*, ser. International Series on Computational Intelligence. CRC Press, 1999.

[243] A. C. Tsoi, *Recurrent neural network architectures: An overview*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 1–26.

[244] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[245] K. M.Tarwani and S. Edem, "Survey on recurrent neural network in natural language processing," *international journal of engineering trends and technology*, vol. 48, pp. 301–304, 2017.

[246] R. M. Schmidt, "Recurrent neural networks (rnns): A gentle introduction and overview," *CoRR*, vol. abs/1912.05911, 2019.

[247] H. Salehinejad, J. Baarbe, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," *CoRR*, vol. abs/1801.01078, 2018.

[248] S. Abujar, A. K. M. Masum, S. M. M. H. Chowdhury, M. Hasan, and S. A. Hossain, "Bengali text generation using bi-directional RNN," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–5.

[249] Z. Yang, K. Wang, J. Li, Y. Huang, and Y.-J. Zhang, "TS-RNN: Text steganalysis based on recurrent neural networks," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1743–1747, 2019.

[250] S. Abujar, A. K. M. Masum, M. Sanzidul Islam, F. Faisal, and S. A. Hossain, *A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN*. Singapore: Springer Singapore, 2020, pp. 509–518.

[251] J. Xiao and Z. Zhou, "Research progress of RNN language model," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2020, pp. 1285–1288.

[252] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, and S. Khudanpur, "Extensions of recurrent neural network language model," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5528–5531.

[253] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 234–239.

[254] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," *CoRR*, vol. abs/2106.11342, 2021.

[255] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *CoRR*, vol. abs/1808.03314, 2018.

[256] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[257] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[258] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[259] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[260] M. Berglund, T. Raiko, M. Honkala, L. Kärkkäinen, A. Vetek, and J. T. Karhunen, "Bidirectional recurrent neural networks as generative models," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[261] E. Arisoy, A. Sethy, B. Ramabhadran, and S. Chen, "Bidirectional recurrent neural network language models for automatic speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5421–5425.

[262] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[263] F. He, T. Liu, and D. Tao, "Why ResNet works? residuals generalize," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5349–5362, 2020.

[264] Z. Allen-Zhu and Y. Li, "What can resnet learn efficiently, going beyond kernels?" in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[265] R. U. Khan, X. Zhang, R. Kumar, and E. O. Aboagye, "Evaluating the performance of resNet model based on image recognition," in *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, ser. ICCAI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 86–90.

[266] M. F. Haque, H.-Y. Lim, and D.-S. Kang, "Object detection based on VGG with resNet network," in *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, 2019, pp. 1–3.

[267] X. Lu, X. Kang, S. Nishide, and F. Ren, "Object detection based on SSD-resNet," in *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2019, pp. 89–92.

[268] X. Ou, P. Yan, Y. Zhang, B. Tu, G. Zhang, J. Wu, and W. Li, "Moving object detection method via resNet-18 with encoder–Decoder structure in complex scenes," *IEEE Access*, vol. 7, pp. 108 152–108 160, 2019.

[269] S. Hayou, E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau, "Stable resNet," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 1324–1332.

[270] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[271] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR*, vol. abs/1312.6114, 2014.

[272] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 1278–1286.

[273] C. Doersch, "Tutorial on variational autoencoders," 2021.

[274] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[275] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 371–375.

[276] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun 2021.

[277] X. Chen, Y. Sun, M. Zhang, and D. Peng, "Evolving deep convolutional variational autoencoders for image classification," *IEEE Transactions on Evolutionary Computation*, pp. 1–1, 2020.

[278] C. Varano, "Disentangling variational autoencoders for image classification," *cs231n. stanford. edu*, 2017.

[279] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[280] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Balancing learning and inference in variational autoencoders," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5885–5892, Jul. 2019.

[281] Z. Ding, Y. Xu, W. Xu, G. Parmar, Y. Yang, M. Welling, and Z. Tu, "Guided variational autoencoder for disentanglement learning," *CoRR*, vol. abs/2004.01255, 2020.

[282] H. Kim and A. Mnih, "Disentangling by factorising," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2649–2658.

[283] Z. Li, J. V. Murkute, P. K. Gyawali, and L. Wang, "Progressive learning and disentanglement of hierarchical representations," *CoRR*, vol. abs/2002.10549, 2020.

[284] C. Shi, "PVAE: Learning disentangled representations with intrinsic dimension via approximated l0 regularization," in *NeurIPS2019 Disentanglement Challenge, Proceedings of Machine Learning Research*, 2019.

[285] D. T. Braithwaite and W. Kleijn, "Bounded information rate variational autoencoders," *ArXiv*, vol. abs/1807.07306, 2018.

[286] D. T. Braithwaite and W. Kleijn, "Speech enhancement with variance constrained autoencoders," *INTERSPEECH*, 2019.

[287] H. Huang, z. li, R. He, Z. Sun, and T. Tan, "IntroVAE: Introspective variational autoencoders for photographic image synthesis," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[288] M. Razghandi, H. Zhou, M. Erol-Kantarci, and D. Turgut, "Variational autoencoder generative adversarial network for synthetic data generation in smart home," *CoRR*, vol. abs/2201.07387, 2022.

[289] D. T. Braithwaite, M. O'Connor, and W. B. Kleijn, "Variance constrained autoencoding," *CoRR*, vol. abs/2005.03807, 2020.

[290] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[291] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, and A. Ku, "Image transformer," *CoRR*, vol. abs/1802.05751, 2018.

[292] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[293] L. Meng, H. Li, B. Chen, S. Lan, Z. Wu, Y. Jiang, and S. Lim, "Adavit: Adaptive vision transformers for efficient image recognition," *CoRR*, vol. abs/2111.15668, 2021.

[294] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," in *International Journal of Computer Vision*, vol. 131, 2023, p. 1141–1162.

[295] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," *CoRR*, vol. abs/2104.01778, 2021.

[296] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 551–561.

[297] H. Lin, X. Cheng, X. Wu, and D. Shen, "Cat: Cross attention in vision transformer," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[298] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 357–366.

[299] P. Dufter, M. Schmitt, and H. Schütze, "Position Information in Transformers: An Overview," *Computational Linguistics*, vol. 48, no. 3, pp. 733–763, 09 2022.

[300] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[301] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[302] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng, and Z. Zha, "A battle of network structures: An empirical study of CNN, transformer, and MLP," *CoRR*, vol. abs/2108.13002, 2021.

[303] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, cited By 8482.

[304] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, Oct 2014.

[305] S. Park, Y. Jeong, and H. S. Kim, "Multiresolution cnn for reverberant speech recognition," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, Nov 2017, pp. 1–4.

[306] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, Dec 2016.

[307] J. Fan, C. Ma, and Y. Zhong, "A selective overview of deep learning," *arXiv preprint arXiv:1904.05526*, 2019.

[308] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[309] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global, 2010, pp. 242–264.

[310] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, 2019, pp. 2613–2617.

[311] Y. S. Rickard, N. K. Georgieva, and Wei-Ping Huang, "Application and optimization of pml abc for the 3-d wave equation in the time domain," *IEEE Transactions on Antennas and Propagation*, vol. 51, no. 2, pp. 286–295, Feb 2003.

[312] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[313] Wikipedia, "Order statistic," 2020.

[314] D. K. Nagar and Y. A. Ramirez-Vanegas, "Distributions of sum, difference, product and quotient of independent non-central Beta type 3 variables," 2013.

[315] T. Rossing, *The Science of Sound*. Addison-Wesley Publishing Company, 1990.

[316] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. USA: California Technical Publishing, 1997.

[317] L. Paarmann, *Design and Analysis of Analog Filters: A Signal Processing Perspective*, ser. The Springer International Series in Engineering and Computer Science. Springer US, 2006.

[318] P. P. K. Normalizacyjny, *Acoustics - Measurement of room acoustic parameters - Part 2: Reverberation time in ordinary rooms (ISO 3382-2: 2008)*, 2008.

[319] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, and J. Cernocky, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, Aug 2019.

[320] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[321] L. Prechelt, *Early Stopping — But When?* Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 53–67.

[322] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: primal estimated sub-gradient solver for svm," *Mathematical Programming*, vol. 127, no. 1, pp. 3–30, 2011.

[323] A. L. Maas, "Rectifier nonlinearities improve neural network acoustic models," 2013.

[324] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814.

[325] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[326] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2013.

[327] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588 – 601, 2007.

[328] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Distant talking robust speech recognition using late reflection components of room impulse response," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4581–4584.

[329] J. Liu and G.-Z. Yang, "Robust speech recognition in reverberant environments by using an optimal synthetic room impulse response model," *Speech Communication*, vol. 67, pp. 65 – 77, 2015.

[330] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 1668–1672.

[331] Y. Li, P. F. Driessen, G. Tzanetakis, and S. Bellamy, "Spatial sound rendering using measured room impulse responses," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 432–437.

[332] J. Merimaa and V. Pulkki, "Spatial impulse response rendering i: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, December 2005.

[333] J. Sheaffer and B. Rafaely, "Equalization strategies for binaural room impulse response rendering using spherical arrays," in *2014 IEEE 28th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, 2014, pp. 1–5.

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

[335] L. C. Jain and L. R. Medsker, *Recurrent Neural Networks: Design and Applications*, 1st ed. USA: CRC Press, Inc., 1999.

[336] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 07 2019.

[337] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.

[338] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2021.

[339] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulations and array processing algorithms," *CoRR*, vol. abs/1710.04196, 2017.

[340] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.

[341] J. Yan and W. B. Kleijn, "Fast simulation method for room impulse responses based on the mirror image source assumption," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2016, pp. 1–5.

[342] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, Aug 2010.

[343] M. Lee and J. Chang, "Blind estimation of reverberation time using deep neural network," in *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, Sep. 2016, pp. 308–311.

[344] N. Faraji, S. M. Ahadi, and H. Sheikhzadeh, "Reverberation time estimation based on a model for the power spectral density of reverberant speech," in *2016 24th European Signal Processing Conference (EUSIPCO)*, Aug 2016, pp. 1453–1457.

[345] F. Lim, P. A. Naylor, M. R. P. Thomas, and I. J. Tashev, "Acoustic blur kernel with sliding window for blind estimation of reverberation time," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2015, pp. 1–5.

[346] W. B. Kleijn, A. Allen, J. Skoglund, and F. Lim, "Incoherent idempotent ambisonics rendering," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017, pp. 209–213.

[347] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 921–930, Aug 2015.

[348] A. Warzybok, J.Rennies, T. Brand, S. Doclo, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, 2013.

[349] D. R. Begault, B. U. McClain, and M. R. Anderson, "Early reflection thresholds for virtual sound sources," in *Proc. 2001 Int. Workshop on Spatial Media*, 2001.

[350] M. Karjalainen and H. Jarvelainen, "More about this reverberation science: Perceptually good late reverberation," in *Audio Engineering Society Convention 111*. Audio Engineering Society, 2001.

[351] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[352] M. M. Lab, "HRTF measurements of a KEMAR dummy-head microphone," 1994.

[353] W. Yu and W. B. Kleijn, "Estimation of source and receiver positions, room geometry and reflection coefficients from a single room impulse response," 2023.

[354] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[355] C. Fifty, E. Amid, Z. Zhao, T. Yu, R. Anil, and C. Finn, "Efficiently identifying task groupings for multi-task learning," 2021.

[356] R. S. Bennett, "Representation and analysis of signals part xxi. the intrinsic dimensionality of signal collections," John Hopkins University, Tech. Rep., 1965.

[357] S. Gong, V. N. Boddeti, and A. K. Jain, "On the intrinsic dimensionality of image representations," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3982–3991.

[358] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[359] M. Chapman, "Symmetries of spherical harmonics: applications to ambisonics," in *Ambisonics Symposium, Graz*, 2009, pp. 1–14.

[360] J. G. Tylka and E. Y. Choueiri, "Algorithms for computing ambisonics translation filters," 3D Audio and Applied Acoustics Laboratory, Princeton University, Tech. Rep., March 2019.

[361] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: a survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017.

[362] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8437–8445.

[363] J. A. Musk, S. K. Sahai, and A. K. Elluswamy, "Estimating object properties using visual image data," U.S. Patent 10 956 755, March 23, 2021.

[364] M. Narbutt, J. Skoglund, A. Allen, M. Chinen, D. Barry, and A. Hines, "Ambiqual: Towards a quality metric for headphone rendered compressed ambisonic spatial audio," *Applied Sciences*, vol. 10, no. 9, 2020.

[365] P. Kabal, "Tsp speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[366] D. Thery and B. F. Katz, "Anechoic audio and 3D-video content database of small ensemble performances for virtual concerts," in *Intl Cong on Acoustics (ICA)*, Aachen, Germany, Sep. 2019.

[367] J. Nistal, S. Lattner, and G. Richard, "Comparing representations for audio synthesis using generative adversarial networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 161–165.

Curriculum Vitæ

Wangyang Yu was born in 24th December, 1992 in Dandong, Liaoning, China. She received the Bachelor of Science (B.Sc.) degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2015. In 2017, she received the Master of Science (M.Sc.) degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands. She started her Ph.D in the Signal Processing System group in the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) of the Delft University of Technology from 2017. Her research interests include room acoustics, ambisonics, audio signal processing, and machine learning.