UNIVERSITY OF TRIESTE

Master in

ELECTRONICS AND COMPUTER ENGINEERING

COMPUTER APPLICATIONS

LEARNING OF ROOM AUDIO RESPONSE EMPLOYING SOUND DIRECTIONAL INFORMATION

GIOVANNI ZANIN

PROF. SYLVIO BARBON JUNIOR
Supervisor

DR. PIETRO MORERIO
Co-supervisor

PROF. VITTORIO MURINO
Co-supervisor

PROF. ALBERTO CARINI
Co-supervisor

March, 2025

Abstract

The main goal of virtual reality and mixed reality systems is to convey a sense of immersion to the user. To do this, in addition to the visual rendering, also the audio transmitted into the headphones should be consistent with the positions and orientations of sound sources and listeners. We employed supervised learning techniques to estimate the value of acoustic parameters and be able to predict sound as it is heard in specific positions within given environments. We extended an existing method by exploiting additional information in the training data regarding the directionality of the acoustic field, made available by the usage of ambisonic microphones for the acquisitions. We then tested our extension and other variations on real recordings and proved their effectiveness.

Sommario

L'obiettivo principale dei sistemi di realtà virtuale e realtà mista è trasmettere un senso di immersione all'utente. Per fare ciò, oltre al rendering visivo, anche l'audio che viene trasmesso nelle cuffie deve essere coerente con le posizioni e gli orientamenti delle sorgenti sonore e dell'ascoltatore. Abbiamo impiegato tecniche di supervised learning per stimare il valore dei parametri acustici di una stanza ed essere in grado di prevedere il suono così come viene udito in posizioni specifiche all'interno di essa. Abbiamo esteso un metodo esistente sfruttando informazioni aggiuntive nei dati di addestramento riguardanti la direzionalità del campo acustico, rese disponibili dall'utilizzo di microfoni ambisonici per le acquisizioni. Abbiamo quindi testato la nostra estensione e altre varianti su registrazioni reali e ne abbiamo dimostrato l'efficacia.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Technologies such as Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR) in recent years are becoming more accessible and consequently more popular. Obviously, in this context a lot of attention should be given to the visual rendering, but since the main objective is typically giving the user an experience that is as immersive as possible, other senses should be taken into consideration as well. For this reason, many studies have been pursued to effectively render the audio of a scene by taking into consideration the different positions and orientations that the listener and the audio sources may have inside the scene.

1.1 Motivations

Audio rendering has importance in many fields including building design and film production, beyond being used for simulations such as in video-games or VR systems. In some cases music and sound are the main focus of these applications and therefore the importance of having accurate audio rendering becomes crucial. A MR that allows users to virtually attend music concerts is one of these cases, since the quality of the system depends a lot on the listening experience. In this work, we will present methodologies to achieve coherent audio rendering and we will discuss technologies useful to reach this goal.

Advances in MR systems of this type, nowadays, have a particular meaning regarding the topic of inclusiveness: quality solutions may give the possibility to everyone to attend concerts and other art performances in conditions in which experiences in presence would have been impossible due to costs, geographic distances or personal disabilities. Beyond pure entertainment, MR of this type can also be used for educational and cultural purposes and may also have therapeutic applications

for mental health, pain management and rehabilitation (for example, providing a distraction during medical procedures).

1.2 Room impulse response

With *Room Impulse Response* (RIR) we mean the signal that can be recorded in a room after generating an audio impulse inside it. In the simpler case, the RIR depends on:

- the characteristics of the room, including its geometry, the configuration of any furniture and the materials present;
- the position and the orientation of the audio source inside the room;
- the position and the orientation of the listener (human or microphone) inside the room.

In a typical scenario of RIR measurement, also the characteristics of the used devices should be taken into consideration, in particular:

- the response of the audio source and its orientation;
- the response of the listener and its orientation.

The RIR takes on meaning when we assume that the room in which the sound propagates is a linear time-invariant system (LTI)[3]. In this case, the RIR can be used to compute the response to any given signal, including, for example, musical audio. It is sufficient to convolve the given signal with the RIR to obtain the room response to that signal:

$$x'(t) = x(t) * RIR(t)$$

where $x'(t)$ is the room response to the signal $x(t)$.

Modeling the RIR with a LTI system is a simplification that doesn't take into consideration some aspects, such as addictive noise $n(t)$, which is always present in real scenarios:

$$x'(t) = x(t) * RIR(t) + n(t)$$

Anyway, the LTI assumption is widely used in literature as it guarantees simplicity and a good level of approximation.

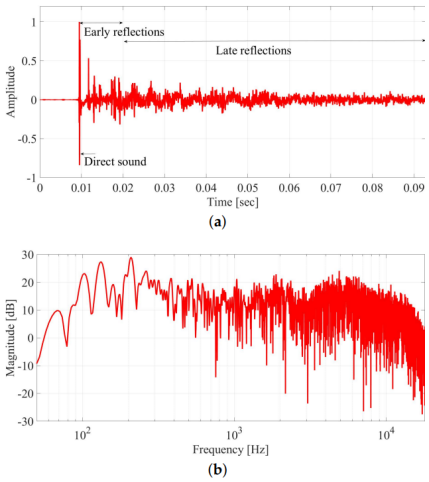


Figure 1.1: Example of RIR [7].
(a) Time domain representation. (b) Frequency domain representation.

The RIR implicitly contains the information about the acoustic characteristics of the room, including the reflectivity of its surfaces. For this reason, RIR applications go beyond audio simulation: other usages may include the analysis and the characterization of specific acoustic environments, focusing, for example, on their reverberation properties.

Although there are many variables that determine the shape of the RIR, in Figure 1.1, some of its characteristics are highlighted. As described in [7], in the time domain representation of the signal, three distinct sections can be easily identified. The first visible peak corresponds to the arrival of the direct acoustic path that departs from the source and reaches the listener. Usually, this is where the amplitude of the wave is highest. Immediately after that, the earliest reflections reach the listener, as shown by the presence of distinct secondary peaks. Later in time, the number of reflections grows and their intensity decreases (especially for the high-frequency components), causing peaks to become increasingly less distinguishable. The first two sections of the RIR, the direct sound peak and the early reflections, are heavily affected by the positions of the emitter and the listener inside the room. In contrast, late reflections depend mostly on the geometry of the room and are almost space-invariant. With respect to the frequency domain, RIRs typically have a regular behavior at low frequencies, with very distinct notches. At high frequencies, the notches become increasingly less distinguishable.

1.3 Research contribution

In this work, state-of-the-art techniques concerning audio rendering will be analyzed and discussed. Furthermore, here is a list of the main contributions to re-

search:

- Supervised learning strategies for the RIR prediction in current literature usually employ datasets for which audio samples are acquired with omnidirectional microphones. During this work, we had the opportunity to use ambisonic microphones as well, which provided us additional information about the directionality of the recorded audio. The usage of these microphones for this task will be discussed, specifically for explicit prediction methods based on acoustic reflective paths.
- Every type of microphone is characterized by a specific frequency response and a polar pattern. Existent methods for the RIR prediction task usually don't take into consideration the characteristics of the microphones used to acquire the audio samples for the dataset. In theory, this would lead to predicting RIRs and audios as they would be acquired by microphones, which is in general, different from predicting audios as they reach the listener. Since for our application we are interested in the latter, in this work, we will take into account the characteristics of the microphones used for acquisitions.
- We'll introduce a loss function to allow training with ambisonic data. Existent loss functions will also be discussed.
- Many literature works adopt a loudspeaker as the audio source. Here, some ideas will be presented to extend existing methods to real-case scenarios where sound sources are musical instruments.

1.4 Thesis outline

Our discussion will start in Chapter 2 with a state-of-the-art analysis of existing methods to solve our task or related ones. After this, in Chapter 3, some fundamentals of room acoustics will be provided. We thought this background information would have been useful to better explain our methodology, which we introduced in Chapter 4. Chapter 5 will contain a description of the datasets used to test our methods, as well as an explanation of the technologies and the strategies adopted for the acquisitions. The chosen parameters for the experiments as well as the results of the latter can be found in Chapter 6. Finally, in Chapter 7, we will discuss the obtained results, the limitations we identified in the adopted methodologies and some possible future works and extensions.

1.5 Working context

The work presented in this paper has been conducted during an internship at the *Pattern Analysis and Vision* (PAVIS) research line of the *Italian Institute of Technology* (IIT). The task was part of *Mixed Reality Environment for Immersive Experience of Art and Culture*(XTREME), a European research project having as overall objective the development of a human-centered and ethically developed MR environment to virtually experience art and music performances and concerts.

Chapter 2

Related Works

Many methods to predict audio propagation and rooms acoustic properties have been developed in the last years. In this chapter, we will discuss some strategies adopted in literature for the RIR prediction and we will present some inherent works.

Due to its relevance in audio analysis, many studies have been made in the last years concerning calculation and prediction of RIRs. The works about this topic can be divided in two macro groups: those that estimate the RIR via interpolation and those, which are rapidly growing, that employ supervised learning techniques.

2.1 RIR prediction via interpolation

This is the simplest way to estimate the RIR in a room. The required dataset should contain recordings of the RIR with fixed source location and different listener locations. With some interpolation technique on these data, a map can be obtained to associate any possible listener position in the room to an RIR signal [11, 21]. In the simplest case, this consists on taking an average of the captured RIRs weighting them according to the distance between the listener position in the acquisition and the desired listener position.

The main drawback of this type of predictions is that no real information about the geometry, interior and acoustic characteristics of the room is considered or learned. This can lead in general to worse results especially when the available datasets are small. Furthermore, interpolation based techniques are not very usable when there isn't a single static audio source. In fact, they typically require a very large number of recordings, since both multiple source locations and multiple listener locations

should be considered during acquisitions. Anyway, due to its computational simplicity, interpolation is often used as a baseline in RIR prediction.

2.2 RIR prediction via supervised learning

This approach has been studied a lot in the last few years and is proven to reach in general better results than the one employing interpolation [37, 8]. The application of machine learning or deep learning techniques for the estimation of the RIR allows to automatically learn models and parameters from data, avoiding human-made procedure that might be difficult, expensive and long to design.

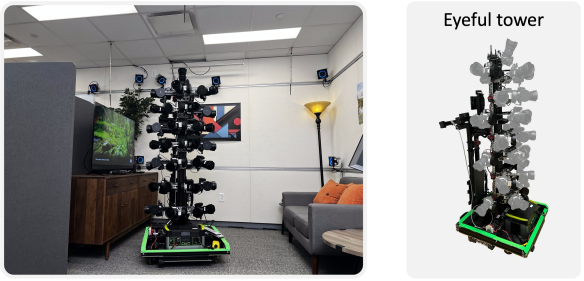
Depending on the information contained in the dataset, that will be used to train a certain model, studies can be divided into three categories:

- **Audio only dataset:** in the simplest cases the dataset contains only the recorded RIRs for different source and listener locations. No explicit information about the room is collected, it can possibly be deduced by the model during learning. In [19], a receiver-to-receiver modeling strategy is presented. In this case, the dataset consists in couples of audio recordings acquired by two distinct robots moving independently around the room. Every recording has a label indicating the positions of the two robots inside the room. A neural network is trained to map two receiver locations p_a and p_b into a warping field $W_{p_a \to p_b}$ i.e. a frequency domain filter that transforms the audio recorded in p_a into a prediction of the audio received in p_b.
- **Audio-visual dataset:** in this case, information about the room is provided through images of its interior. In [24] a few images are provided with the intention of giving some generic information about the context in which the audio is generated. On the contrary, in [23] images are collected for every recorded audio sample, from the same position where the recording microphone is located. Examples of these images, which can be acquired for both indoor and outdoor environments, are shown in Figure 2.1. A dense audio-visual dataset such as the one used in this case, however, is not easy to obtain and may require very sophisticated tools. Figures 2-2(a) and 2-2(b) show the tools used in [8] to collect audio and visual data, respectively. These tools allow dense capture with automatic annotation of source and listener positions. Many times, a reference source signal (i.e., a dry signal without the effects introduced by room characteristics) is not that easy to acquire. That's why in [36] a self-supervised method is employed to remove the room characteristics from a recorded audio. The model is trained on recorded samples together with an image of the room in which they are recorded. An off-the-shelf de-reverberator and the trained de-biaser are used to map the recorded

(a) Office (b) House (c) Apartment (d) Outdoors

Figure 2.1: Examples of images paired to the recordings in the [23] dataset.

(a) Audio capture tools.



(b) Visual capture tool.

Figure 2.2: Tools for audio-visual data capture used in [8].
(a) Movable height-adjustable loudspeaker and movable microphone tower equipped with 36 omnidirectional microphones. (b) Movable multi-camera rig.

signal into a dry version of it.

The audio-visual dataset category also includes the many cases in which the prediction of the audio response is combined with the visual prediction of what is seen from a specific listener-viewer position. In [5], audio prediction employs the output of the visual prediction: a *Neural Radiance Field* (NeRF) is used to map coordinates and visual orientations to density and color. The output of the NeRF is used both to visually render the room and to build a 3D model of it. A *Residual Neural Network* (ResNet) is then used to extract from the 3D model features such as distances and materials inside the room. This features are employed as room descriptors by a *Neural Acoustic Field*

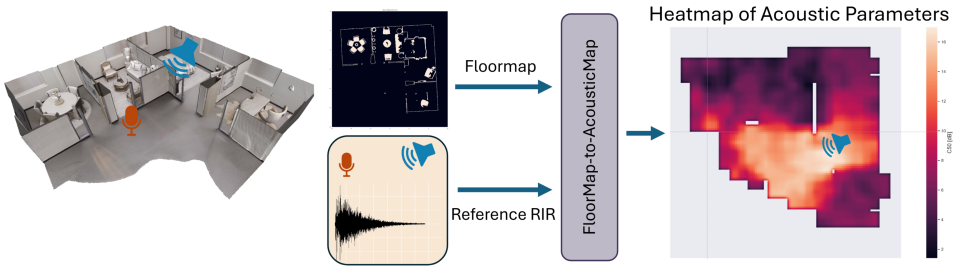


Figure 2.3: Pipeline adopted in [15] to predict the 2D heatmap of an acoustic parameter (C50).

(NAcF) to map couples of source and listener locations into RIRs.

- **Audio dataset and geometry of the room:** Instead of using images, information about the room can be included in the dataset through the explicit description of some of its characteristics that are considered useful for the correct prediction of the RIR. For example, in addition to audio recordings, [41] provides to the predicting model the equations of the (main) surfaces contained in the room (including walls, ceiling and floor). In this case surfaces positions are crucial to esteem the reflections that the audio undergoes and to subsequently compute the RIR. Similarly, in [9], the reflectivity of the surfaces in the room is approximated. In this case, the optimization employs a genetic algorithm where the fitness function is determined by comparisons between the values of simulated and measured acoustic parameters. In [15], a 2D floormap of the room is used together with a RIR recorded for a chosen couple of source and listener positions. The model predicts two heatmaps to describe how two acoustic parameters, the energy decay rate (EDR) and the early-to-late energy ratio (C50), vary inside the room. This procedure is shown in Figure 2.3.

A further distinction of the learning-based existent works is based on the approach used to predict the RIR, which can be implicit or explicit.

- **Implicit learning:** the RIR is predicted using implicit methods such as neural networks (NN). The main pro of this method is its simplicity: no particular knowledge or studies about acoustics and sound diffusion are required. The learned parameters are the ones of the NN, so the models are in general lacking in explainability. Furthermore, this methods do not allow transferability: to predict the RIR in a new environment a whole new learning is required, even if the room has some similarities with another in which training has already been carried out (for example same walls but different furniture). Implicit RIR predictions may be differentiated depending on what is given as input for the model: for example, in [25] the input consists basically in

the positions of the audio source and the listener inside the room and the orientation of the latter, in [37], features such as relative positions between points (source location, reflection points and listener location) are extracted, encoded and given as inputs for the model, using three specific modules.

- **Explicit learning:** the RIR is predicted applying the laws of acoustic phenomena to create an explicit formula. One of the most famous techniques to model the RIR as a function of the room characteristics is the image method [2], based on the wave properties of sound and still used in many acoustic simulations. Procedures like this often lead to elevated computation complexity, that's why many methods are based on the assumptions of geometrical acoustics, which provide a simplification of the sound propagation that neglects all the wave properties and adopts reflective rays [34]. Reflection is, in fact, the phenomenon that mostly affects the listening of a sound in a room and in many cases is the only one worthy of being modeled. In [41], the RIR is rendered using a model that involves the computation of the contributions of every reflective path, and optionally considers paths with transmission too (but the authors agree that in most standard environments modeling other minor acoustic effects other than reflection doesn't really positively affect the performances). If the final application consists in many environments (i.e. many rooms) you may consider modeling other physical phenomena such as diffraction. This is the case of [33].

  Once an explicit model is created, supervised learning can be used to optimize the parameters of the formula, including for example the reflectivity coefficients to associate to each surface or the coefficients describing the dissipation of the sound in the air. The explainability of explicit methods also makes them transferable. If a reflective surface changes position inside a room (for example a table or a panel are moved), there is no need to train again the model, the previously learned reflectivity coefficients of the surface can still be used adopting the new position for the computation of the reflective paths.

Chapter 3

Room acoustics fundamentals

In this chapter some necessary background topics will be discussed. Specifically, we will get in order through all the elements that influence the RIR.

The RIR is the filter that maps a source signal, diffusing from a specific source location, into an acquired signal, as recorded in a specific listener location. Let's go through the elements that determine this filter, in order, from audio production to its acquisition.

3.1 Source characteristics

If the source signal is given as input to a loudspeaker to be reproduced inside a room, the first thing that should be taken in consideration is the characteristic of the speaker. In fact, every speaker is characterized by a specific frequency response that describes how faithfully the signal given as input is reproduced. In the ideal case the response of a speaker is flat in the range of frequencies audible to humans, but in reality some of them are always either attenuated or amplified and this characteristic can slightly change also between different speakers of the same model. Another thing that should be taken into consideration is the directionality of the speaker, that can either be omnidirectional or not. Even in the omnidirectional case, the frequency response won't remain exactly the same for every outgoing direction of the sound from the speaker, so the best way to describe how the source signal is reproduced is to associate a frequency response to every angle of the speaker. For this reason, the room response should be computed knowing both the source location and its orientation.

The source might also be a music instrument instead of a loudspeaker. In this case the source signal can be considered the one that comes out of the instrument and

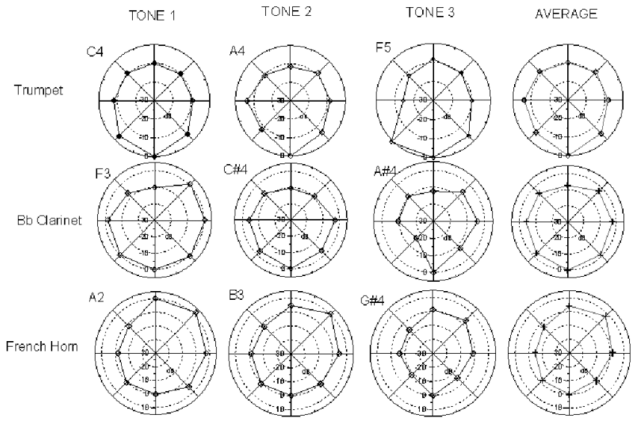


Figure 3.1: Directivity in the horizontal axis of different tones played by different music instruments [29].

directly diffuses in the room, without the need to consider source intrinsic characteristics. The directionality of the source, that in this case is determined by the orientation of the musician, should still be taken into account, since the sound propagation is not always homogeneous [29]. Figure 3.1 shows the directivity on the horizontal axis of the sound produced by some wind instruments, when playing different tones (i.e. at different frequencies).

3.2 Room characteristics

Many characteristics of the room affect its audio response, here is some background information about the ones needed for our analysis.

- **Room's geometry:** of course, the geometry of the room affects its acoustic response. With a sufficiently detailed description, the physical phenomena involved can be correctly estimated.
- **Surface reflectivity:** since reflection is the effect that influences the RIR the most, it's important to know how it occurs inside a room. When an acoustic ray reaches a surface part of the energy is absorbed and part is reflected creating a new ray, that is an attenuated copy of the original one. The attenuation coefficient depends on the material of the surface and can be frequency-variant, so a frequency response should be associated to every surface to describe the relation between the incident and the reflected signal. In practice, it's important to have a frequency response only for the largest surfaces within a room, since they are probably the ones that reflect most acoustic paths and that have a relevant role in reflection.

- **Surface transmissivity:** when an acoustic ray reaches a surface there is also a part of the energy that is transmitted trough the material. Similarly to the reflection, to correctly model transmission, a frequency response should be associated to every surface to describe how the frequencies are attenuated in the transmitted signal. However, transmission seems to be not so relevant in RIR computation [41] and in most cases it can be neglected since the materials that are usually present in a room are not so transmissive.
- **Time-of-arrival delay:** every acoustic path that reaches the listener will have a different time-of-arrival, depending on its length and on the speed of sound in the medium. This interval should be used to correctly synchronize the contribution of the paths to the resultant RIR.
- **Propagation absorption:** the longer the path the larger the area in which the audio is propagating, so the energy of the signal decreases with the square of the distance. This means that for the amplitude of the signal an attenuation factor inversely proportional to the path length should be taken into consideration.
- **Air absorption:** the time of arrival of an acoustic path can be used also to determine the attenuation of the signal due to air absorption. In this work, we will model the absorption factor with α^{t_p}, where t_p is the time-of-arrival of the path and $\alpha \in [0, 1]$ is the air absorption coefficient in the room.
- **Residual effects:** high-order reflections (i.e. paths that reflect on many surfaces) result in very attenuated signals difficult to distinguish. Their effect, together with the one of other minor physical phenomena such as diffuse reflections, diffraction and refraction, can be assumed homogeneous and isotropic inside a room [28, 30] so their overall effect can be modeled as a space-invariant contribution to the RIR.

Figure 3.2, summarizes the phenomena mentioned above.

3.3 Microphone characteristics

Typically, a microphone is used to acquire the audio from the room. For this reason, the RIR depends also on the specific characteristics of the microphone. One of the main characteristics is the range of frequencies that a microphone acquires. Usually, the response is as homogeneous as possible for frequencies in the human audible range (between 20 Hz and 20 kHz), with attenuations at the ends (see for example Figure 3.3). Even in microphones of the same model the exact curve that represents the frequency response can slightly vary (usually the upper bound of the deviation is specified in the equipment data sheets and has a value of about ± 2 dB). Because

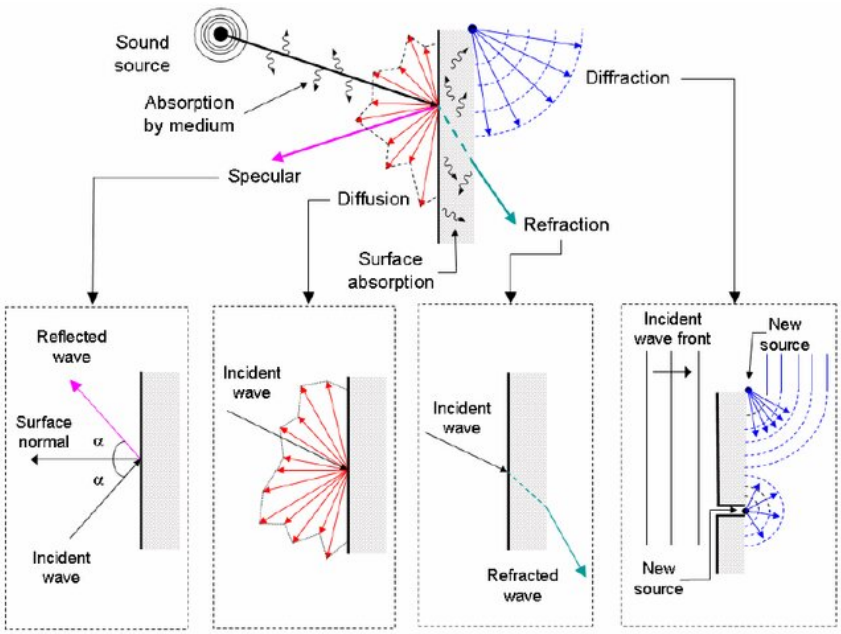


Figure 3.2: Acoustic phenomena [22].

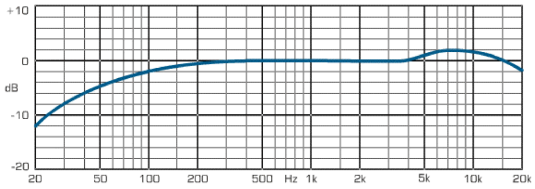


Figure 3.3: Frequency response of the Neumann KM184 microphone.

of this, in case multiple microphones should be used, ideally, a calibration phase is needed, consisting in a preliminary recording with every microphone followed by gain compensations at every frequency to obtain the same frequency response curve for all of them.

Another thing that should be taken into consideration is the polar characteristic of the microphones that are used, representing the directions for which the incoming sound is acquired. Here are some distinctions:

3.3.1 Omnidirectional microphones

Omnidirectional microphones are meant to acquire the sound incoming from every direction. In practice, the omnidirectionality is lost when increasing the frequency. As can be seen in Figure 3.4, the response of omnidirectional microphones is perfectly omnidirectional only at low frequencies, while the sound is attenuated at high frequencies for off-axis incoming directions. The reflective paths that reach the microphone from these directions will be attenuated, and then contribute less to

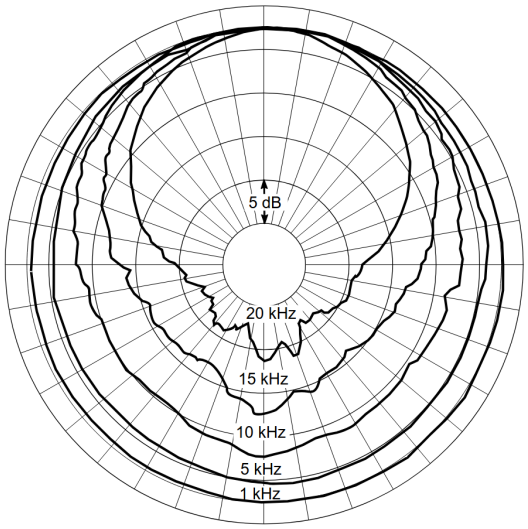


Figure 3.4: Polar characteristic of the DPA4006 omnidirectional microphone (with the grid supplied as standard).

the recorded signal. Then, for accurate analysis, the orientation of omnidirectional microphones should be taken into consideration.

3.3.2 Cardioid microphones

Cardioid microphones are directional microphones designed to acquire audio only from some specific directions, with a polar characteristic similar to a cardioid. The unitary cardioid can be described with:

$$C(\alpha) = \frac{1 + cos\alpha}{2}$$

where α is the angle between the considered direction and the reference direction (i.e. the one the microphone is oriented) [39]. However, in real cases the shape of the cardioid of a microphone varies with the frequency, leading to higher directivity at higher frequencies. Figure 3.5 shows how the directivity changes with frequency for the Neumann KM184 microphone. To model this changing we can adapt the previous formulation with:

$$C(\alpha) = \left(\frac{1 + cos\alpha}{2}\right)^k$$

where k is a number associated to every frequency describing the directivity of the microphone. The higher k the more the cardioid becomes tight and only the most frontal frequencies are acquired, so k grows with the frequency (typical values are

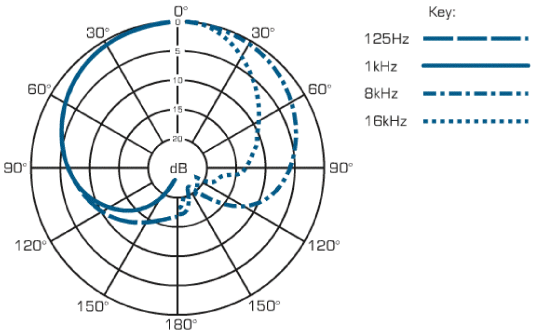


Figure 3.5: Polar characteristic of the Neumann KM184 cardioid microphone.

Figure 3.6: Neumann KU100 "dummy head" binaural microphone.

between 0.4 and 1.2).

3.3.3 Binaural microphones

Binaural microphones are designed to acquire sound simulating human hearing. They record a stereo signal where the two channels correspond to the human hears. Binaural microphones are supposed to include also the so called *Head Related Impulse Response* (HRIR), that represent the filter that map the sound that reaches the head to the one that actually enters the ears, that includes the reflections, the transmissions and the other acoustic phenomena that involve the ears and the head. To do this, binaural microphones are often shaped to recreate human anatomy, as can be seen in Figure 3.6.

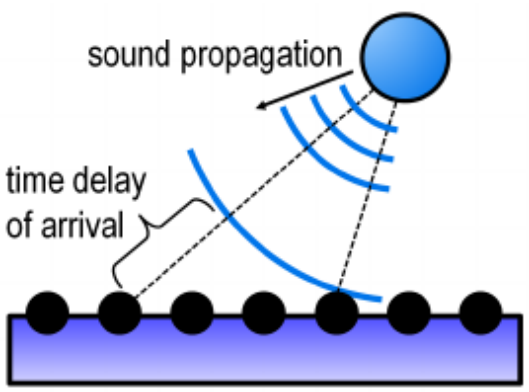


Figure 3.7: Microphone dependent time delays introduced by the arrival of sound waves at a microphone array [16].

3.3.4 Ambisonic microphones

Ambisonic microphones are microphone arrays capable of recording audio coming from every direction and subsequently obtain directional information from the recordings. They are composed of a certain number of capsules arranged with a specific and known geometry.

Every capsule is an omnidirectional microphone that will record the signal from a specific spatial position. The recordings will differ mainly in the delay time, as the time-of-arrival of the signal will be different for every capsule (see Figure 3.7). Since the time-of-arrival depends on the direction where the sound comes from, these recordings, combined, implicitly contain information about the signal directionality .

The *A-format* signal, that is composed by the recordings of the capsules of the microphone, is usually processed to obtain a signal in the *B-format* (or *Ambisonic format*) that represents the tridimensional acoustic field as a sum of harmonic spherical components [45, 44]. Figure 3.8 shows the spherical harmonics up to the 3rd order. In this case, the B-format signal would be composed of 16 channels, each of them corresponding to one spherical harmonic. This signal can then be processed to isolate sound coming from one specific direction through a procedure called *beamforming*.

In a real case, the output of the beamforming isn't determined exclusively by the sound waves coming from the desired direction. In fact, every direction will have a contribution to the output, even if minimal. The contributions of waves coming from each direction are described by a *beampattern*, whose shape can vary depending on the technique used for the beamforming. The simplest technique is the so called *Plane Wave Decomposition beamforming* (PWD beamforming), which creates a maximum-directivity beampattern that can be approximated with a hyper-

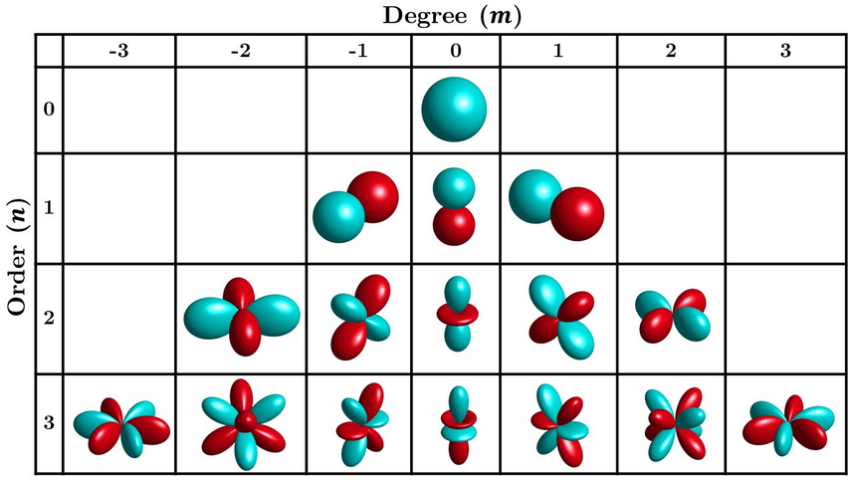


Figure 3.8: Spherical harmonics up to 3rd order [42]. Light blue lobes represent positive polarity and red lobes represent negative polarity.

cardioid [38, 40]. Figure 3.9 shows a beampattern of this kind. Here, the amplitude of the hyper-cardioid represents the attenuation applied to sound waves when trying to isolate one specific direction (in this case, the zenithal direction). The hyper-cardioid in the picture is of the 4th order as evidenced by the presence of 5 lobes (an hyper-cardioid of order n has $n+1$ lobes, including the main one and the rear one).

The order of the hyper-cardioid beampattern depends on the characteristics of the ambisonic microphones, in particular on the number of capsules. An ambisonic microphone with m capsules is supposed to enable beamforming up to the order n, where n is the maximal integer such that $(n+1)^2 \leq m$ [43]. The higher the order of the beampattern, the more directive the beamforming will be (i.e., the more isolated the desired direction will be), as shown in Figure 3.10.

To completely understand the behavior of beamforming we should take into consideration the fact that ambisonic microphones do not always guarantee the highest possible level of directionality. In fact, at low frequencies, sound waves coming from different directions are harder to distinguish. This is equivalent to saying that at low frequencies the beampattern may have an order lower than the maximum possible (i.e., the order granted by the number of capsules of the microphone). This happens because audio waves at low-frequency, which have large wavelength, present very slow variations. Phase variations between the positions of different capsules may then be very small or even insignificant, making it hard for ambisonic microphones to do spatial distinctions. This is also the reason why this directivity limitation is more accentuated in small-sized microphone arrays.

Zylia ZM-1 (Figure 3.11) is an example of ambisonic microphone. It is composed by 19 capsules that allow a beamforming up to the 3rd order ($(3+1)^2 = 16 \leq 19$).

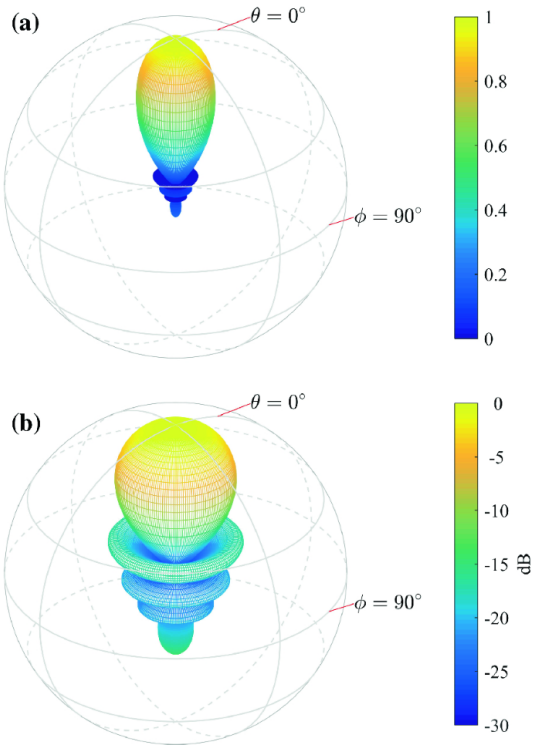


Figure 3.9: 4th order hyper-cardioid beampattern with zenithal orientation. (a) Beampattern in linear domain. (b) Beampattern in the logarithmic domain.

However, for the reasons mentioned above, the ZM-1 microphone can guarantee its nominal beamforming order only for waves above a certain frequency, that given its size (4.9 cm radius) is about 3.3 kHz [31]. At lower frequencies the beampattern will be 2nd order or less.

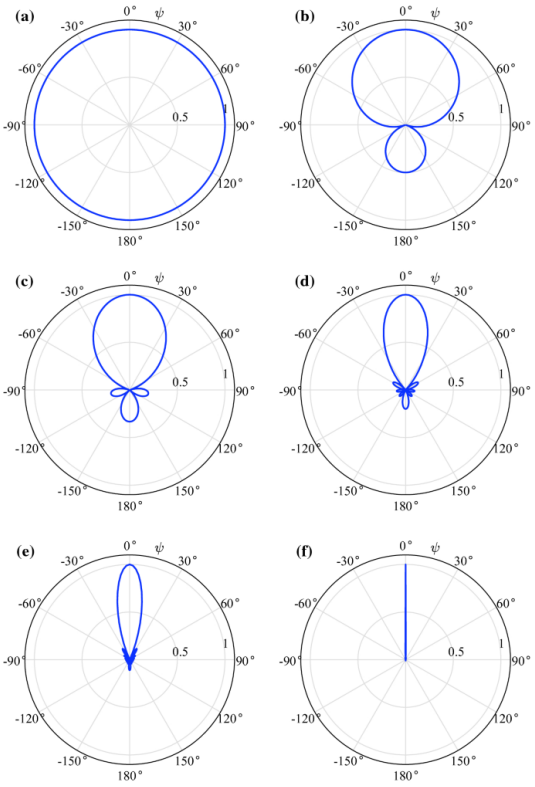


Figure 3.10: 2D projections of hyper-cardioid beampatterns of different orders. (a-e) Beampatterns of order 0 to 4. (f) Ideal beampattern.

Figure 3.11: Zylia ZM-1 ambisonic microphone.

Chapter 4

Methods

In this chapter the methodology used for the RIR prediction will be described. We will discuss the basic idea, taken from another existing work, and subsequently the extensions we implemented.

The method we used for the RIR prediction is based on the one in [41], which employs an explicit model trained on an audio-only dataset and provided with an explicit description of the room.

4.1 The model

The proposed RIR prediction procedure starts with the computation of the acoustic paths that depart from the source position (S_{xyz}) and reaches the listener position (L_{xyz}). The considered paths are the direct ones and the ones that reach the listener after a certain number of reflections on the surfaces of the room (optionally, also transmission through the surfaces can be taken into account). The reflective paths are computed only up to a certain order (5, for example), where the order of path is the number of surfaces the sound is reflected on before reaching the listener. For every multiset of surfaces up to the maximum order, it is verified if the corresponding acoustic path exists. Higher order reflections result in more attenuated signals, whose effect is modeled together with other late response minor phenomena, as described later in this section. If inside a room, two parallel surfaces are close to each other, multiple audio reflections between them may lead to resonance and amplification of some frequencies. This is, for example, the case of the walls in a narrow corridor. Sometimes, this resonance is relevant for the resultant RIR and is worth being modeled, so for reflections on pairs of surfaces of this kind, the highest order considered may be set to a larger value (10, for example).

An initial frequency response is associated with every acoustic path, corresponding to the response of the source for the specific direction the ray is emitted (d_p). This frequency response includes information only about the module, not the phase, of the filter that has to be applied to each frequency. Every path's response is then multiplied by the frequency response associated with every surface s the path reflects on (the path may also reflect on the same surface more than once). Still, only the modules of the frequency responses are taken into consideration. Then, every frequency response is anti-transformed, and the obtained signal in the time domain is delayed according to the time-of-arrival of the path (t_p). The adopted anti-transformation technique uses the Hilbert relation method to compute the minimum-phase of the Fourier transform [35]. Assuming the phase to be minimal is a method justified by acoustic research literature [18, 20] and physically is equivalent to considering every reflection as "instantaneous". To reduce border discontinuities and artifacts, the anti-transforms are also windowed using a Hamming function. The last step to obtain the contribution of every acoustic path to the RIR consists of attenuating each ray depending on air absorption and propagation. The contribution K of each acoustic path can then be expressed as:

$$K(d_p, S_p, t_p) = \frac{\alpha^{t_p}}{\rho}\tau \left[M \left(D(d_p) \cdot \prod_{s \in S_p} R_s \right), t_p \right]$$

where S_p is the multiset of the surfaces on which path p reflects, α is the air absorption coefficient, ρ is the length of the path in meters (useful to model the propagation attenuation), $\tau(x, t)$ is an operator that delays the signal x by a certain interval t, M is the minimum-phase Fourier anti-transformation, $D(d)$ is the function that maps every direction d exiting the source into the correspondent frequency response and R_s is the frequency response associated with the reflection on surface s.

Once the contributions of all paths are computed, they are all summed together and the resultant time-domain signal is convolved with a filter describing the source impulse response (IR_s). While $D(d_p)$ is a frequency-domain filter supposed to model the directionality of the source, IR_s is a time-domain filter which is intended to describe the overall response of the source, that might also depend on some user-settable features but is independent of the outgoing direction. The obtained signal represents the *early response* since it takes into account the direct path and the low order reflections that are supposed to define completely the first part of the RIR signal.

Finally, the early response is combined with the *late response* (r), a space-invariant signal in the time domain that represents the contribution of minor effects (high order reflections, diffraction, etc.). The early and the late responses are combined

in a weighted average where the weights are given by the values of a temporal spline γ. This last step can be expressed as:

$$RIR(S_{xyz}, L_{xyz}) = (1-\gamma)\left[IR_s * \sum_{p \in P} K(d_p, S_p, t_p)\right] + \gamma r$$

where P is the set of all the considered acoustic paths.

Once the RIR is predicted it can be compared with the ground-truth one, a signal recorded with a microphone in L_{xyz} and a speaker in S_{xyz}. The parameters of the model can then be optimized. In [41], the model is trained on a dataset of 12 audio samples recorded with a fixed S_{xyz}. The optimized parameters, that we also adopted, were:

- **Energy vector (R):** the frequency responses associated to reflection on every surface in the room. For every surface, the gain is learned only for a few frequencies, and the whole response is then obtained through interpolation. When initializing the model, all these frequency responses are assumed to be flat.
- **Source impulse response (IR_s):** a time-domain signal that contains source characteristic. This parameter is initialized as an impulse, that is equivalent to assume the transfer function to be an identity.
- **Directivity sphere (D):** a function that maps an outgoing direction from the source into a frequency response. The sphere is sampled and the response is learned only for some directions (128, for example). The gains are learned only for a few frequencies, so the complete function is obtained by interpolating both on the outgoing directions and on the frequencies. For all directions, frequency responses are assumed to be flat at initialization time.
- **Decay (α):** the air absorption coefficient. Since these coefficients should have a value between 0 and 1, at rendering time, the learned parameter is normalized with a sigmoid to obtain a usable value for α. The initial value was set to 5.
- **RIR residual (r):** a time-domain signal that represents the late response. Initially, this signal is considered null.
- **Spline values (γ):** time-dependent relative weights of the contributions of the early and the late response. Only a few values are learned and the complete spline is obtained via temporal interpolation and sigmoid normalization. This parameter is initialized as an ascending sequence, since the weight of the late response is supposed to grow with time.

With these training parameters the model is differentiable and can then be optimized with gradient descent techniques.

4.2 Loss function

As basic training loss function, the one described in [41] was adopted. This function calculates a distance between the *short-time Fourier transform* (STFT) of the predicted signal and the one of the ground-truth signal. The distance is the sum of the mean distance $L1$ between the windows in the linear domain and in the logarithmic domain (the $L1$ distance between two signals in the logarithmic domain, which will be used later also for evaluation, is equivalent to the absolute value of the logarithm of the rate between the two signals). This comparison is repeated adopting different temporal window sizes for the STFT, in a way to extract different information from the signals (smaller windows give more information about the characteristics of the signal in the time domain, while larger windows give more information about its characteristics in the frequency domain). Specifically, we compared 4 different STFTs adopting Hann windows of size $s = 512, 1024, 2048, 4096$ samples, spaced with a hop length $\frac{s}{4}$. The time-domain characteristic of the initial part of the response is considered very important as it contains information about the time-of-arrival of the direct acoustic path from the source to the listener. To ensure that this starting section is accurately predicted, a *tiny-hop loss* is added to the loss described above. The tiny-hop loss is characterized by STFT with window size $s = 256$ and very small hop length (1, typically). This loss involves only the first segment of the predicted and the ground-truth signal.

Experiments were also conducted with a slight variation of the training loss, similar to the one described in [26]. The idea is to assign, during training, more importance to the correct prediction of the decay of the RIR. The decay of a signal, that is closely related to reverberation, is supposed to be very important to get a feeling of realism in an audio simulation. We obtained the decay curve D for the n^{th} time window (of size 512) of a signal with the formula:

$$D[n] = 1 + \frac{E[n]}{\sum_{j=n+1}^{N-1} E[j]}$$

where $E[n]$ is the energy of the n^{th} window of the STFT, and N is the total number of windows. The new loss function that we introduced can then be formulated as:

$$L_{new}(x, y) = L(x, y) + \lambda L_D(x, y)$$

where L is the standard loss function, L_D is a decay loss measuring the average $L1$ distance between the decay curves of the predicted and ground truth RIRs in the logarithmic domain and λ is the weight assigned to L_D.

Taking inspiration from [41], we also adopted *pink noise supervision* during training. This technique involves summing two loss functions: one of them is calculated between the predicted and the ground-truth RIR, the other between a convolution of these two with a generated pink noise. The latter is a signal characterized by power inversely proportional to frequency, resulting in a frequency spectrum similar to the one of music pieces. With pink noise supervision, predicted and ground-truth RIRs are compared not only in a direct way, but also in a similar-to-real-case scenario.

4.3 Directional extension

The model described above assumes that the ground-truth samples are recorded by perfectly omnidirectional microphones, and the same weight is assigned to every incoming path, disregarding its direction of arrival.

During this work, we had the opportunity to collect data also with ambisonic microphones, which provided us additional information about the directionality of the incoming sound. The procedure discussed above, based mainly on the computation of acoustic paths, makes it easy to predict the isolated audio signal that reaches the microphone from one specific direction. This signal can then be compared with a ground-truth obtained by doing the beamforming in the same direction using the data acquired with an ambisonic microphone. This comparison can then be repeated in many directions.

The consequence is a more sophisticated comparison between the predicted signal and the ground-truth that should favor the correct learning of room parameters by the model. The predictions would be obtained considering more correct contributions from every direction, leading, in principle, to a better identification of the most reflective surfaces and subsequently to better results.

Ambisonic acquisitions allow us to introduce three new loss functions:

- The first and most immediate loss function consists of the summation, for every considered microphone incoming direction, of the basic loss function described in [41] and in Section 4.2:

$$L_{amb} = \sum_{dir \in D} L(x_{dir}, y_{dir})$$

where x_{dir} is the RIR predicted in the direction dir, y_{dir} is the ground-truth signal coming from direction dir (obtained via beamforming) and D is the set of directions considered. The number of these directions should depend on the angular resolution the ambisonic microphone allows for the beamforming, in general 6 is a good number (in this way directions can be used to describe signal coming from right, left, up, down, front and back). This function was used as a basic loss function for the ambisonic case.

- To favor the correct prediction of reverberation in the room, the decay loss can be taken into consideration for every direction:

$$L_{amb} = \sum_{dir \in D} L(x_{dir}, y_{dir}) + \lambda L_D(x_{dir}, y_{dir})$$

where L_D is the decay loss described in Section 4.2 and λ is the associated weight.

- Accurate beamforming is not easy to implement and this fact often leads to noisy signals when trying to isolate a specific direction. For this reason, it may not be appropriate to make a time domain comparison between the predicted signal that reaches the listener from a certain direction and the ground-truth obtained by orientating the beampattern in the same direction. To make the training more robust to inaccurate beamforming, directional information can be employed to simply compute the ratio between the average power of the audio signals coming from opposite directions. Starting from this idea we introduced the following loss function for ambisonic data:

$$\begin{aligned} L_{amb} = &|R_{DxSx}(x) - R_{DxSx}(y)| + \\ &+ |R_{UpDown}(x) - R_{UpDown}(y)| + \\ &+ |R_{FrontBack}(x) - R_{FrontBack}(y)| \end{aligned}$$

where $R_{DxSx}(z)$ is the rate $\frac{RMS_{Dx}(z)}{RMS_{Sx}(z)}$ between the root mean squared signal coming from right and the one coming from left, $R_{UpDown}(z)$ and $R_{FrontBack}(z)$ are the same rate but considering Up-Down and Front-Back directions.

4.4 Microphone response

During this work, we wanted to take into consideration the real characteristics of the microphones that acquired the datasets. In principle, the model described in [41] can predict the RIR as the signal that reaches the microphone, which is slightly

different from the signal that is recorded by the microphone, since the latter introduces both fixed and direction-dependent frequency gains (described respectively by the frequency response and by the polar characteristic of the microphone, see Section 3.3).

In the final application case, that is rendering a RIR and using it to predict via convolution a signal to be listened by the user, the sound should be used as it is when it reaches the listener and microphone characteristics should not be taken into consideration. On the other side, if the predicted RIR should be compared to a ground-truth one during training or validation phases, frequency modifications introduced by the used microphones should be taken into consideration. Otherwise, the model would wrongly associate this attenuations and amplifications to room characteristics, causing artifacts in the learned parameters. To take into account microphone characteristics, the predicted RIR should be properly processed before being compared to the recorded one, according to:

- **Microphone gain:** in case precise calibration is not possible and the microphones used for recording have slightly different gains, the predicted RIR should be amplified/attenuated according to the gain of the corresponding microphone. In principle, this step can be substituted by proper preprocessing of the ground-truth acquisitions.
- **Microphone frequency response:** some frequencies of the predicted RIR should be slightly amplified/attenuated depending on the response of the microphone. Also, this step can be, in principle, avoided if the acquisitions are preprocessed.
- **Microphone polar characteristic:** to obtain a more accurate prediction of the recorded signal, the contribution of each acoustic path that reaches the microphone should be weighted according to the polar characteristic of the latter. The exact shape of this characteristic is frequency-dependent (see Figures 3.4 and 3.5) so a different frequency-domain filter should be applied to every path, depending on its incoming direction in relation to the orientation of the microphone. This correction has slight effect on omnidirectional microphones since they're designed to minimize off-axis attenuations. In this case, the gain on each direction can be assumed to linearly decrease moving away from the microphone orientation direction, with frequency-dependent decreasing speed. On the other side, this idea can also be applied to cardioid microphones, making it now possible, in principle, to use them to train and validate the model. In practice, this compensation consists of a correction of the contribution K associated to each path and described in Section 4.1:

$$K(d_p, S_p, t_p, d'_p) = \frac{\alpha^{t_p}}{\rho}\tau \left[M\left(D(d_p) \cdot \prod_{s \in S_p} R_s \cdot L(d'_p) \right), t_p \right]$$

where d'_p is the direction from which the path p reaches the listener, and $L(d)$ is the function that maps every direction d entering the microphone into the corresponding frequency response of the sensor. For every path, we computed few samples of the frequency response $L(d'_p)$, obtaining the others via frequency interpolation. This computation was based on the nominal characteristics of the microphones declared in the datasheets and on the gains measured with the calibrator. Concerning the late response, which we assume to be isotropic (see Section 3.2), we attenuated it according to the portion of the surrounding sphere that is captured by the microphone. For example, cardioid microphones capture approximately half of this sphere, then the residual r should be attenuated by a factor 0.5.

A similar procedure is used for ambisonic microphones, for which we need to render the signal that comes from one specific direction. This rendering should be done according to the beamforming technique that was used to obtain the ground-truth: instead of simply taking into account the paths coming from the desired direction, all paths should be considered and properly weighted according to the shape of the beampattern. During our experiments, we also took into account the loss of directionality that beamforming presents at low-frequencies, using lower-order beampatterns to weight the paths (see Figure 3.10).

To compare the predicted RIR in one specific listener position to the BRIR acquired by a binaural microphone in the same position we adopted the binauralization techniques used in [41], based on dataset [4]. This technique can also be used in the final application case to map the monophonic predicted signal into a binaural signal to be transmitted into the user's headphones, recreating more in detail the sound that reaches each of the listener's ears and giving a more authentic feeling of immersion. Anyway, in this work, binauralization and binaural data were used only during validation.

Chapter 5

Dataset

In this chapter the set up for the experiments will be described. Special attention will be given to the the acquisitions of the datasets and the technologies employed.

Two datasets were used to train and test the model, both of them acquired using different types of microphone. One is the public dataset [27], acquired in the *Arni* variable acoustics room at the Acoustics Lab in Espoo, Finland. The second dataset was acquired by other partners of the XTREME project in the *VIP studio*, a room made available by the University of Nottingham (UN). In addition to RIR measurements, the Nottingham dataset also contains recordings of a snippet of a classical music piece (from [13]) reproduced by a speaker and of live performance of two groups of artists: a quartet from the *Irish Chamber Orchestra* (ICO, composed by two violinists, one violist and one cellist) and a trio of artists from the *Irish World Academy of Music and Dance* (IWAMD) (composed by one violinist, one singer/flutist and one dancer).

Training our model on data acquired in a room would cause the learning of parameters useful only for one specific set of surfaces, but it's still a good way to test the behavior of the method and its extensions.

5.1 The rooms

For both rooms, the approximate positions of the main surfaces should be given as input to the model to calculate the reflective paths. Every surface was described either with a triangle or a parallelogram.

The Arni room is a small place measuring approximately 8m $\times$ 6m $\times$ 3m. The few objects that were inside it during the acquisitions were considered irrelevant as

Figure 5.1: Set up used in [27] for the acquisitions with the Zylia ZM-1 microphones.

regards reflections (see Figure 5.1), so the surfaces taken into account were only the walls, the floor and the ceiling, which positions were known. On the other side, the VIP studio is a much larger space (17m $\times$ 13m $\times$ 7m). Starting from a point cloud scan describing the room, we used off-the-shelf methods based on *Random Sample Consensus* (RANSAC) techniques to localize the main surfaces. Figure 5.2 shows all these considered surfaces, which include walls, floor, ceiling, tables, doors and panels. During audio acquisitions, musicians, loudspeaker and microphones were not in a central position but closer to a corner of the room (corner [0,0,0] in Figure 5.2). Figure 5.3 shows this corner after the set up for the ICO quartet recording.

5.2 RIR measurement

As already said, the RIR is the signal that can be recorded in a room after an acoustic impulse is reproduced. In practice, this signal is usually not recorded directly but is extracted from other recordings, since the propagation of an ideal acoustic impulse is very sensitive to background noise [30]. A technique often used in this kind of measurements consists in acquiring the response to another chosen reference signal and then obtaining the RIR by deconvolving the original reproduced signal from the recorded one. In both datasets, an exponential sine sweep was used as reference signal, which is a solution widely adopted in literature due to its robustness to both ambient noise and nonlinearities in the measurement system [6]. A sine sweep is

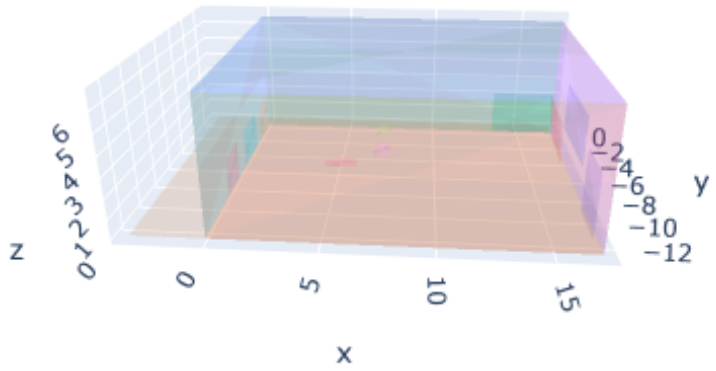


Figure 5.2: Surfaces considered for the room where the acquisitions were taken in Nottingham.

Figure 5.3: Set up before the recording of the ICO performance (kind permission of Christian Sivertsen, ITU).

a sinusoid which frequency increases over time, making it possible to cover completely a wide frequency range. This range should be approximately 20Hz-20kHz, to completely cover the human audible. If the increment is exponential, frequency will sound as linearly growing to the human ear, that perceives these variations in a logarithmic scale. Some software can automate the process of reproducing a reference signal, recording the response and applying the deconvolution, giving the measured RIR as output.

For both datasets, the measured RIR were subsequently time-synchronized according to the distance of the microphones from the audio source.

5.3 Equipment

The Espoo dataset was acquired employing the following equipment:

- **Zylia ZM-1 ambisonic microphones ($\times$ 7)**, that can perform 3$^{\text{rd}}$ order beamforming.
- **mh Acoustics Eigenmike em32 ambisonic microphones ($\times$ 7)**, that can perform 4$^{\text{th}}$ order beamforming.
- **Genelec 8331A coaxial loudspeakers ($\times$ 3)**, to reproduce the sine sweeps.

In this case, during the experiments, the ambisonic microphones were used both as actual ambisonic microphones, i.e. doing the beamforming and exploiting the directional information, and as omnidirectional microphones. In the latter case, the 0$^{\text{th}}$ order spherical harmonic of the B-format was used, as it represents the omnidirectional component (see Figure 3.8). In other words, omnidirectional signal was obtained by applying 0$^{\text{th}}$ order beamforming.

Concerning the Nottingham acquisitions, the equipment included:

- **DPA4006 omnidirectional microphones ($\times$ 8)**, used to record train and validation data.
- **Neumann KM 184 cardioid microphones ($\times$ 4)**, used to record reflections only (they were pointed towards the walls).
- **Neumann KU100 binaural microphone ($\times$ 1)**, dummy head used to acquire binaural data.
- **Zylia ZM-1 ambisonic microphones ($\times$ 2)**, used to record ambisonic data.
- **Condenser microphones (one for each instrument)**, positioned close to each artist, to record the dry sound (i.e.: the sound without the effects introduced by room characteristics). The dry sound can be convolved with a

RIR to predict how an instrument is heard in a specific position in the room. The dry sound was collected also for the shoes of the IWAMD dancer, modeling them as an audio source just like the other instruments. Different models of microphones were used depending on the instrument they had to record.

- **Sound Pressure Level calibrator ($\times$ 1)**, used to measure the exact gain of each omnidirectional microphone response at 1kHz. This value could be used to get the amplification/attenuation implicitly introduced by microphones in every recording.
- **Yamaha HS5 loudspeaker ($\times$ 1)**, used to reproduce the sine sweeps and the music snippet.

Furthermore, in the Nottingham VIP studio, a 4m $\times$ 5m $\times$ 2m cubic scaffolding was available, which allowed to position the omnidirectional microphones around the musicians/loudspeaker.

5.4 Acquisitions

5.4.1 Espoo acquisitions

Espoo dataset contains RIRs in the A-format (raw recordings of the microphone capsules) and B-format (in the spherical harmonics domain). Figure 5.4 shows the configuration during the recordings.

Ambisonic RIRs were acquired with both Eigenmike em32 and Zylia ZM-1 microphones in 7 listener positions while reproducing sine sweeps from 3 different source positions. Measurements were repeated with 5 different reverberation levels, obtained applying different coverings on the walls.

5.4.2 Nottingham acquisitions

In Nottingham, acquisitions were conducted over a 4 day period with different dispositions of the microphones and the audio source. The recordings include:

- **RIRs**, which were used to train and evaluate the model.
- **Sine sweeps**, from which the RIRs were computed.
- **Classical music snippet**, 30 seconds from an Italian Opera piece (“Come Paride vezzoso” from "Elisir d’amore" by G. Donizetti) reproduced by a loudspeaker. This recordings, together with the reference anechoic audio (the audio given as input to the speaker) could be used to evaluate the model in a real case scenario, by comparing the predicted room response (obtained by

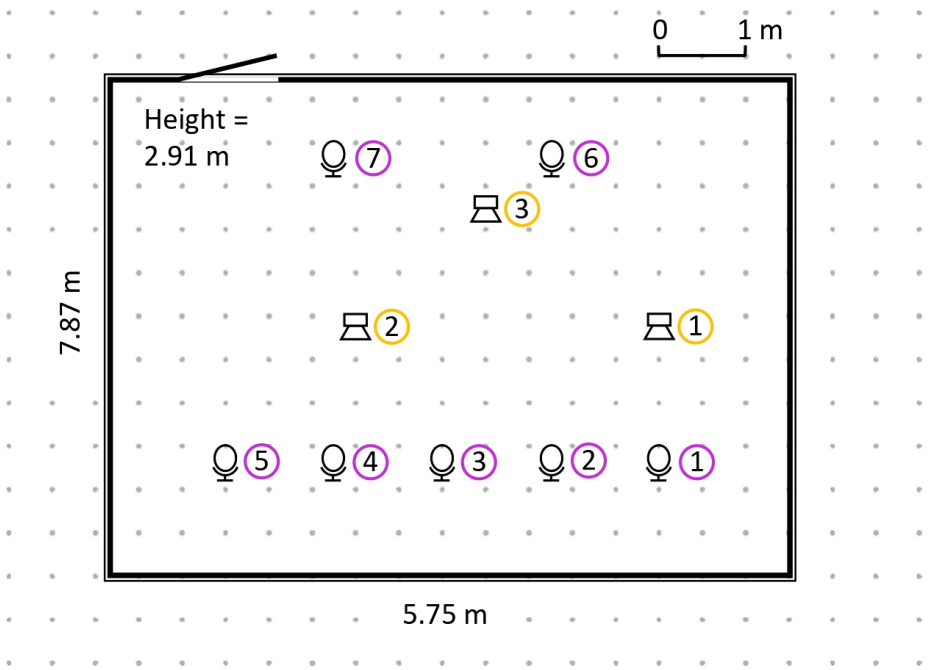


Figure 5.4: Espoo recording configuration [27].

convolving the predicted RIR with the reference audio) with the recorded ground-truth. This evaluation technique is used also in [41].

- **Live performances**, the direct captures of the instruments could be used to render the performance as it was heard in a certain listener position, by convolving them with the corresponding RIR (multiple audio sources in multiple positions should be modeled to render the entire orchestra). In theory, this renderings could also be compared to the ground-truth acquired in some positions inside the room and provide a further method to evaluate model.

In order to obtain variety in the dataset, different configurations were used for the recordings:

Main configuration (Source in S1)

This configuration is shown in Figure 5.5. In the figure, positions O1-O8 are the ones of the omnidirectional microphones (when two microphones have the same positions in the picture, it means they were placed at different heights), A1-A4 indicate cardioid microphones (that were directed towards the walls to acquire reflections), Z1-Z2 indicate the positions of the Zylia ZM-1 ambisonic microphones and H1 is the position of the dummy head binaural microphone. The recordings with the microphones in this position include the RIRs (only for omnidirectional and ambisonic microphones), the classical music snippet, and the performances of

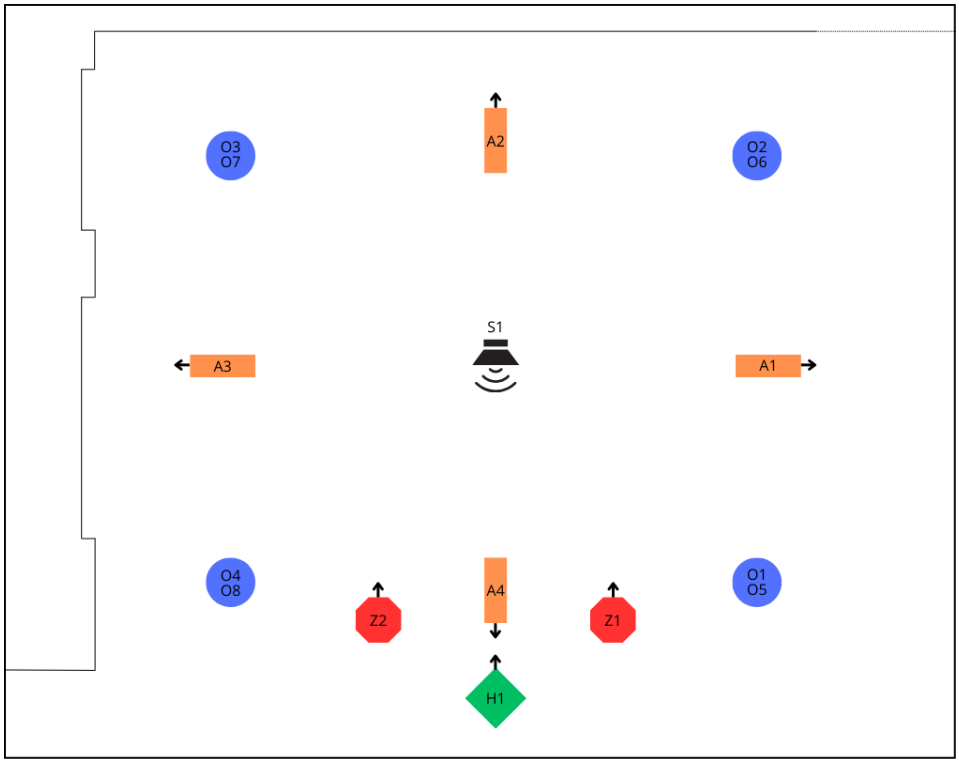


Figure 5.5: Main capturing configuration.
Kind permission of Caroline Gaudeoso, ITU.

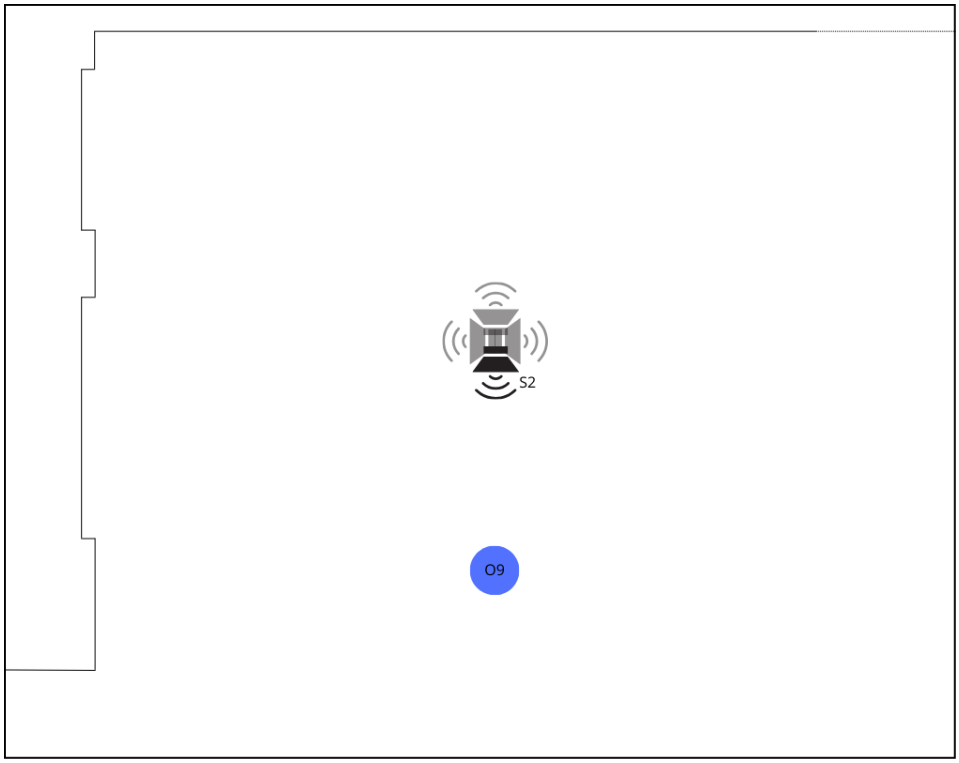
both groups. In the first two cases, the loudspeaker was positioned in S1. For live performances, the musicians were distributed around the same position.

Data acquired with this configuration can be used for training (in particular ambisonic and omnidirectional RIRs) and for evaluating the model with the audio source in a fixed position.

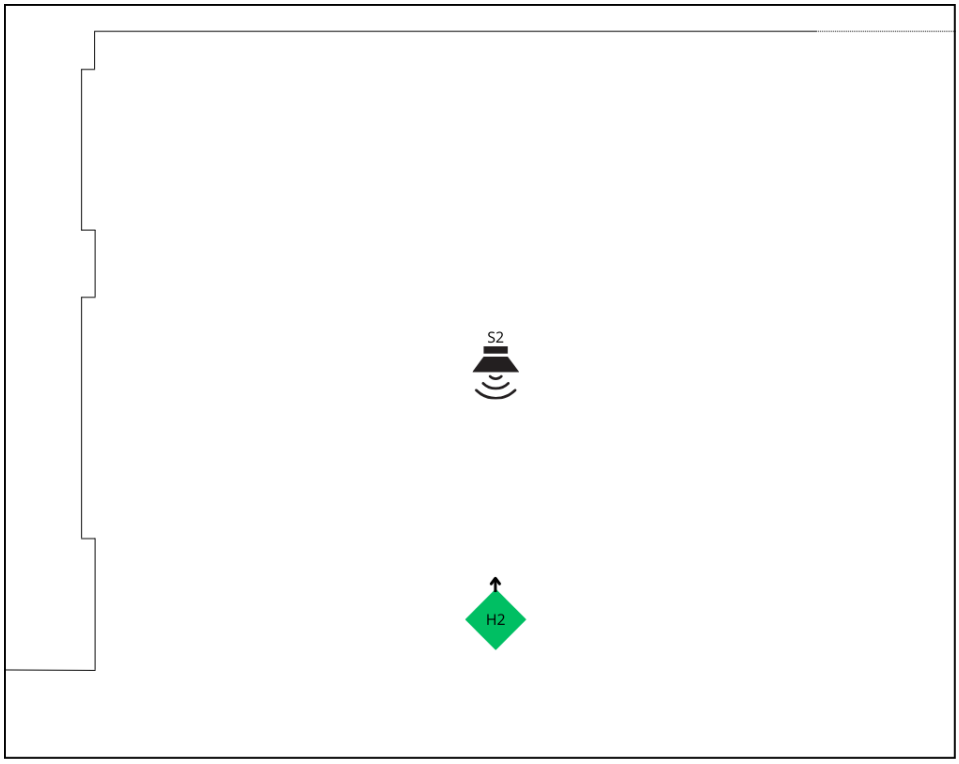
Source rotation (Source in S2)

Later, the speaker was raised from the height of 1.07m to the height of 1.40m, in a new position that we will call S2. 4 RIRs were measured by an omnidirectional microphone in position O9, with 4 different orientations of the speaker, differing by 90 degrees each, as shown in Figure 5-6(a). These acquisitions can be used to validate the model. Once the parameters describing the characteristic of the source have been learned, it's easy to rotate the ones that describe its orientation to simulate the rotation of the speaker. This simulations can then be compared to the ground-truths acquired by recording with the source pointing in different directions.

Keeping the loudspeaker in position S2, a binaural RIR (BRIR) was also acquired, with the dummy head in position H2 (see Figure 5-6(b)). The BRIR can be used to test the model and the binauralization method.



(a) Source rotation and omnidirectional capture.

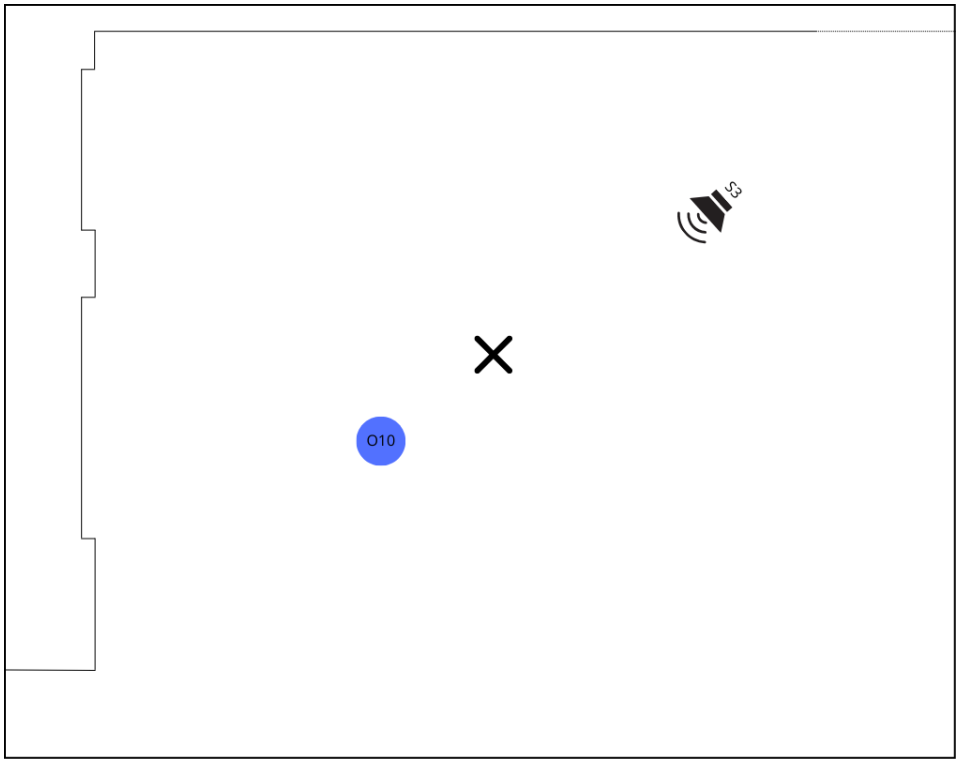


(b) Binaural capture.

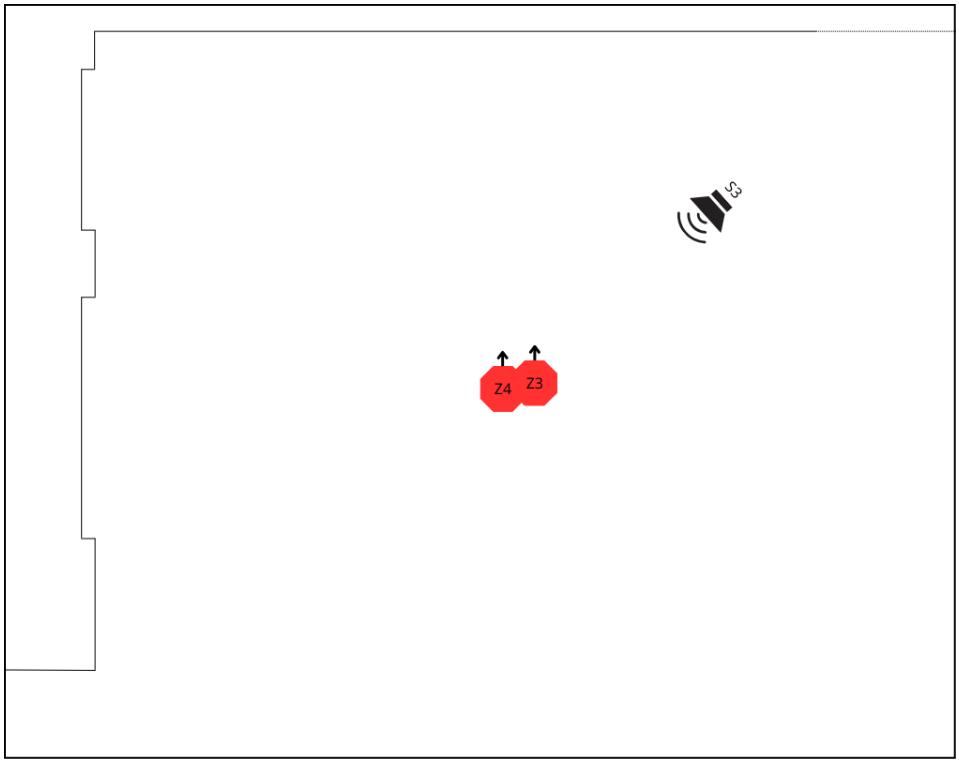
Figure 5.6: Configuration for source rotation and binaural capture. Kind permission of Caroline Gaudeoso, ITU. (a) RIR measurements with rotated audio source. (b) BRIR measurement.

New source position (Source in S3)

Position S1 and S2 differed only for the height of the speaker. To test the model for a completely unexplored source position, the speaker was moved to S3 and RIRs were acquired with an omnidirectional microphone in O10 (Figure 5-7(a)) and with two ambisonic microphones, at different heights, in Z3 and Z4 (Figure 5-7(b)).



(a) Omnidirectional capture.



(b) Ambisonic capture.

Figure 5.7: Configuration for capture with unexplored source position. Kind permission of Caroline Gaudeoso, ITU. (a) RIR measurement. (b) Ambisonic measurement.

Chapter 6

Experiments

This chapter contains the results of the conducted training experiments. At first, a description of the chosen parameters and the used evaluation metrics will be provided. Then, the effectiveness of the proposed methods will be discussed.

6.1 Training details

For all experiments, the model was trained for 500 epochs, with a learning rate equal to 0.01. The predicted RIRs were 2 seconds long and sampled with a frequency of 48kHz. The ground-truth RIRs were truncated to have the same length. The directivity sphere of the audio source (D in Section 4.1) was modeled via interpolation on 128 directions, evenly distributed on the sphere with the Fibonacci sampling technique [17]. The frequency response was learned only for these directions and was obtained via interpolation for all the others. Frequency interpolation was used to model the frequency filters, both the learned ones (source directivity sphere and surfaces reflectivities) and the fixed one (microphone characteristic). The values of the spline (γ in Section 4.1), were obtained via temporal interpolation. During training, we adopted a time-of-arrival (ToA) perturbation technique, consisting in adding Gaussian noise to the ToA delays before applying them to each path. This perturbation should make the training more robust to errors in the localization of the reflective surfaces and is proven to lead to better results [41]. Pink noise supervision (see Section 4.2) was employed starting from epoch 250. Table 6.1 summarizes the chosen values for the training and rendering parameters.

6.2 Metrics

As evaluation metrics we adopted the following:

Table 6.1: Parameters used for experiments.

Parameter	**Value**
Number of epochs	500
Learning rate	0.01
RIR length (samples)	96000
sampling rate (Hz)	48000
Length of source impulse response IR_s (samples)	1023
Length of Fourier anti-transformations M(samples)	1023
Fibonacci sampling directions	128
Frequencies for source directivity sphere D interpolation (Hz)	[32, 63, 125, 250, 500, 1000, 2000, 4000, 8000, 16000]
Frequencies for surfaces reflectivity R interpolation (Hz)	[32, 63, 125, 250, 500, 1000, 2000, 4000, 8000, 16000]
Samples for spline γ temporal interpolation (indexes)	[200, 500, 1000, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 6000, 7000, 8000, 10000, 12000, 14000, 16000, 18000, 20000, 22000, 24000, 26000, 28000, 30000, 32000, 34000, 36000, 38000, 40000, 44000, 48000, 56000, 70000, 80000]
ToA perturbation	True
ToA perturbation standard deviation (samples)	7
Pink noise supervision	True
Pink noise supervision starting epoch	250
Model transmission	False

- **Multiscale Log-Spectral L1 (Mag)**: STFT of the predicted and the ground-truth signals are computed at several time-frequency resolutions (window sizes were 64, 128, 256, 512, 1024, 2046 and 4096 samples, with hop length equal to the window size divided by 4). Then, for every resolution, the average L1 distance is calculated between the windows in the logarithmic domain and the results are summed together. This metric is used in [41] and inspired by [10] and [14].
- **Envelope Distance (ENV)**: this metric is supposed to evaluate the reverberation characteristics of the predictions [12]. It represents the L1 distance between the energy envelopes of the ground-truth and the predicted signal, in the logarithmic domain. The energy envelopes were obtained by flattening the instantaneous energy of the signal with a 32 samples wide moving average filter.

All the tables in this chapter contain the values of the Mag and the ENV metric multiplied by 10. For both metrics, the lower the better.

We evaluated our methods both by testing them for unexplored source location (i.e. different from the one used during training) and for fixed source position. In the latter case, we repeated the training keeping out one of the listeners available, to then use it for evaluation. This means that for this evaluations the model was trained with one less observation than the unexplored source position case. Despite this, it always resulted easier for the model to predict RIRs for explored source positions, as shown in the next tables.

6.3 Exploiting directional information

The first thing we wanted to evaluate was the effectiveness of the directional extension we introduced to the model. We trained on the Espoo dataset, specifically on the recordings with the ZM-1 microphones, the source in position 2 (see Figure 5.4) and the room configuration with the maximal reverberation. We chose position 2 because it is intended to recreate the position of an artist on a stage, in the center of the room surrounded by the listeners. We trained the model in two ways:

- **Omni**: without exploiting the directional information. The model was trained using the ambisonic microphones just as if they were simple omnidirectionals, using the basic loss function described in Section 4.2.
- **Ambi**, exploiting the directional information for the first 200 epochs, by employing the basic loss function for the ambisonic case described in Section 4.3. For the following 300 epochs, only the omnidirectional information associated to the microphones was used.

Table 6.2: Effectiveness of directional information exploitation. Results are multiplied by 10.

	Training source location		**Unexplored source location**	
	Mag	ENV	Mag	ENV
Omni	9.48	2.32	25.08	8.81
Ambi	**9.26**	**2.08**	**9.54**	**2.24**

To evaluate the two trainings on an unexplored source position we used data acquired with source in position 1 (see Figure 5.4).

Table 6.2 shows the results of these experiments (all results are the average of multiple training experiments). It can be seen that employing the directional information provided by the ambisonic microphones leads to better rendering results, as the values of both error metrics are lower with this training technique. The improvement is particularly evident when testing for unexplored source location, for which the two distances are more than halved by the directional extension. This proves that ambisonic training leads to a better identification of the reflectivity of surfaces and in general better estimation of the room acoustic parameters, which provides better generalization capabilities to the model.

6.4 Compensating microphone characteristics

To evaluate our microphone characteristics compensation method we used data acquired with omnidirectional microphones in Nottingham. The nominal characteristics of the DPA4006 were adopted, taking into consideration also the orientations that the microphones had during the recordings. We trained the model with S1 as source location, using S2 and S3 for unexplored source position evaluations (see Subsection 5.4.2). Table 6.3 shows the results of the experiments conducted applying and not applying the compensation for the microphone characteristics (all results are the average of multiple training experiments).

Results show that taking into account microphone characteristics during training leads to very slight improvements in the performance of the model. This means that if not explicitly attributed to the microphones, the model is able to consistently integrate this compensation into the learned room parameters.

Table 6.3: Effectiveness of microphone characteristics compensation. Results are multiplied by 10.

	Training source location		**Unexplored source location**	
	Mag	ENV	Mag	ENV
w/o Compensation	2.15	0.38	**3.60**	0.76
w/ Compensation	**2.12**	**0.36**	**3.60**	**0.75**

Table 6.4: Performance of loss with decay loss contribution. Results are multiplied by 10.

	Unexplored source location		
	Mag	ENV	RT60 error [ms]
$\lambda = 0$	**3.60**	0.75	16.43
$\lambda = 10$	4.31	0.66	10.42
$\lambda = 50$	4.31	**0.65**	**1.85**

6.5 Using different loss functions

We evaluated the performance of the training adopting as loss function the one with the decay loss contribution described in Section 4.2. We compared it to the performance with the basic loss function (that is equivalent to set the weight λ of the loss decay equal to 0). Experiments were conducted on omnidirectional data acquired in Nottingham with the source in S1 and microphone characteristics compensation was applied.

The results are showed in Table 6.4, where also the error in the *Reverberation Time 60* (RT60) is specified, in milliseconds (all results are the average of multiple training experiments). The RT60 of the RIR is a widely used index to describe the reverberation level of a room, as it represents the time for the sound to reach a 60dB attenuation in respect to the initial value. As expected, the more the decay loss is weighted, the better the prediction of the reverberation room characteristics. This is demonstrated by the decrease of ENV and RT60 errors as λ increases. On the other side, the decay loss assigns less importance to correct time domain prediction, as testified by the increment of the Mag error.

Table 6.5: Effectiveness of rates loss for ambisonic epochs. Results are multiplied by 10.

	Training source location		**Unexplored source location**	
	Mag	ENV	Mag	ENV
Basic loss	**9.26**	2.08	9.54	2.24
Rates loss	**9.26**	**2.06**	**9.50**	**2.23**

We also repeated the experiments of Section 6.3 using, during the ambisonic training epochs, the rates loss function described in Section 4.3. This function exploits the directional information only to compute the rates of power coming from opposite directions.

Results are shown in Table 6.5. They demonstrate that the rates loss function, that doesn't require a time domain comparison between ambisonic signals, is a valid alternative to the basic loss, as the performances seems to be similar or even slightly better.

Chapter 7

Conclusions

In this chapter, we will discuss the achieved results. After recognizing the limitations present in our procedures, the conclusions drawn from the experiments will be discussed. Finally, ideas for future works will be presented as well as the acknowledgements.

7.1 Limitations

Some aspects represented limitations for the performance of the proposed methods, they include:

- **The number of training samples**: in [41] the model was trained with 12 different listener locations, more than we had available in the Espoo dataset (7) and in the Nottingham dataset (8).
- **The distribution of the microphones**, especially in the Nottingham dataset, in which both training and evaluation samples were recorded by microphones placed in the immediate vicinity of the source, leaving a large part of the room unexplored. Because of this, we suppose the learned model won't have great generalization capability for very shifted positions of the source or the listener.
- **The approximate knowledge of the surfaces positions**, especially in the Nottingham dataset. During the recordings, some surfaces were not in exact same position they were when the point cloud was acquired, so their location was approximated, introducing slight inaccuracy in path computation.

The fact that transmissions weren't modeled is not a real limitation. In the vast majority of cases, the surfaces inside a room have very low transmissivity and the

effect of this phenomenon is minimal. In [41], it is proven that modeling transmissions through surfaces doesn't significantly improve performances.

7.2 Conclusions drawn from experiments

Our experiments have proven that exploiting the directional information of audio recordings acquired with ambisonic microphones, even for only 200 epochs, leads to a better prediction of room acoustic parameters and, consequentially, to better rendering results. The positive effect of the directional information exploitation is particularly evident when testing the model for an an unexplored source location, which testifies that using ambisonic microphones leads to better generalization capabilities due to better room acoustic parameters estimation. Anyway, the usage of ambisonic microphones for RIR prediction tasks should depend on the required quality for the rendering and the availability of this type of microphones, since their cost is about 10 times the one of omnidirectional microphones.

We have also proved that compensating the microphone characteristics during training can lead to slightly better results, since in this way certain frequency attenuations/amplifications are explicitly associated to the microphone and are no longer wrongly attributed to room characteristics. The importance of adopting this compensation during training necessarily depends on the properties of the microphones, specifically on how close their characteristics are to the ideal ones, i.e. frequency responses flat in the audible range and consistent in every direction.

The results show that the introduction of a decay loss can help the correct prediction of the reverberation characteristic of a room, and its effect can be controlled by modifying the weight λ.

Finally, the rates loss for ambisonic data is proven to be a valid alternative to the basic ambisonic loss function. The former doesn't require accurate time-domain computation of the predicted and the ground-truth signal, which makes the model less dependent on the accuracy of the beamforming.

7.3 Future works

In this section, we will discuss some possible future works and further extensions of the methods proposed.

7.3.1 Model improvements

- **Exploitation of materials a priori knowledge**: in many cases, information about the materials the surfaces are made of is available, at a certain level of detail, or is easy to obtain. From this, the surfaces that are likely to be the most reflective can be identified and the energy vector (R in Section 4.1) can be initialized with more plausible values of reflectiveness for each surface.
- **Learning of the microphone characteristics**: microphone datasheets may not contain all the necessary information and in some cases the real values may deviate from the nominal ones. This is why the microphone frequency response and polar characteristic, instead of being estimated from the datasheets, can be included in the learnable parameters. The risk in this case is that the model may attribute to the microphone some attenuations/amplifications that are caused by other characteristics of the room. This would cause artifacts and imprecise predictions when the microphone response is excluded.
- **Ambisonic data distillation**, due to the big price gap between professional omnidirectional and ambisonic microphones, it may be worth to explore the usage of some trainable teacher models to map recordings acquired with omnidirectional microphones into ambisonic data.

7.3.2 Audio simulation task completion

- **Data compression**: as already said, to render a sound for a specific source-listener position pair, the corresponding RIR should be convolved. For a VR application, a dense grid of RIR should be computed to be able to respond to every possible configuration. In practice, saving every possible RIR as a time-domain signal is impossible due to computation complexity and memory occupation, then some sort of compression is needed. In [32], a parametric compression is adopted. In this case, the RIRs are characterized only by few parameters, that include delay, direction and loudness of the direct path, delay and direction of the early reflections and the RT60. This parameters are supposed to be sufficient for the rendering to sound plausible, providing a huge reduction of the data size.
- **Audience effect modeling**: in the specific application of VR to attend concerts, the rendering of the acoustic effects introduced by the audience may affect the overall sensation of immersion that the user experiences. Acquiring a dataset with the presence of an audience is not easy and neither is modeling people as a set of surfaces. Then, the main possibility is to compensate for this effects by properly modifying the model after conducting the learning

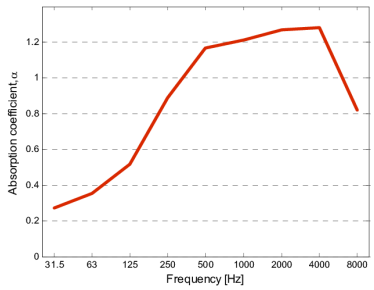


Figure 7.1: Absorption coefficients for standing audience [1].

Figure 7.2: Audience associated to one specific surface.

phase with an empty room. The first thing to consider is that human bodies absorb part of the acoustic paths, especially medium-high frequencies [1] (see Figure 7.1). With a low density of people, the effect of this absorption is negligible, and the model can be kept unchanged. In other cases, the absorption can be associated to one or more specific surfaces (an example is provided in Figure 7.2), and the model can be adapted by lowering the reflectiveness of that surfaces. Finally, if the audience is scattered but too numerous to be neglected, its effect can be compensated by applying a proper low-pass filter to the overall resultant RIR. Another aspect to consider is the audio introduced by the possible movement/chatting of the audience, that can be added to the rendered one of the musicians. In general, this noise can be modeled as the sum of a music-independent component (related to people chattering) and a music-dependent one (related to possible people humming/singing). The former can be rendered by using existing recordings of audience noise and the latter by properly filtering and processing the source signal.

Acknowledgements

This work is funded by the European Union within the Horizon Europe research and innovation programme under grant agreement No. 101136006 – XTREME project, coordinated by S.S. Brandt/IT University of Copenhagen, Denmark. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union. European Union can not be held responsible for them.

Bibliography

[1] Niels Adelman-Larsen and Eric Thompson. Variable low-frequency absorber for multi-purpose concert halls.

[2] Jont Allen and David Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65:943–950, 04 1979.

[3] Adil Alpkocak and Kemal Sis. Computing impulse response of room acoustics using the ray-tracing method in time domain. *Archives of Acoustics*, 35, 12 2010.

[4] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney. A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database. *Applied Sciences*, 8(11), 2018.

[5] Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. Neraf: 3d scene infused neural radiance and acoustic fields, 2024.

[6] Alberto Carini, Stefania Cecchi, and Simone Orcioni. Robust room impulse response measurement using perfect periodic sequences for wiener nonlinear filters. *Electronics*, 9:1793, 10 2020.

[7] Stefania Cecchi, Alberto Carini, and Sascha Spors. Room response equalization—a review. *Applied Sciences*, 8:16, 12 2017.

[8] Ziyang Chen, Israel D. Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. 2024.

[9] Claus Christensen, George Koutsouris, and Jens Rindel. Estimating absorption of materials to match room model against existing room using a genetic algorithm. 09 2014.

[10] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. Diffimpact: Differentiable rendering and identification of impact sounds. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 662–673. PMLR, 08–11 Nov 2022.

[11] Orchisama Das, Paul Calamia, and Sebastià Amengual Garí. Room impulse response interpolation from a sparse set of measurements using a modal architecture. 05 2021.

[12] Simona De Cesaris, Dario D'Orazio, Federica Morandi, and Massimo Garai. Extraction of the envelope from impulse responses using pre-processed energy detection for early decay estimation. *The Journal of the Acoustical Society of America*, 138(4):2513–2523, 10 2015.

[13] Dario D'Orazio. Anechoic recordings of italian opera played by orchestra, choir, and soloistsa). *The Journal of the Acoustical Society of America*, 147(2):EL157–EL163, 02 2020.

[14] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing, 2020.

[15] Ricardo Falcon-Perez, Ruohan Gao, Gregor Mueckl, Sebastia V. Amengual Gari, and Ishwarya Ananthabhotla. Novel view acoustic parameter estimation, 2024.

[16] Friedrich Faubel, Munir Georges, Kenichi Kumatani, Andres Bruhn, and Dietrich Klakow. Improving hands-free speech recognition in a car through audio-visual voice activity detection. pages 70 – 75, 07 2011.

[17] D. P. Hardin, T. J. Michaels, and E. B. Saff. A comparison of popular point configurations on $\mathbb{S}^2$, 2016.

[18] Sahar Hashemgeloogerdi and Mark Bocko. Invertibility of acoustic systems: An intuitive physics-based model of minimum phase behavior. volume 23, page 055002, 01 2015.

[19] Yuhang He, Shitong Xu, Jia-Xing Zhong, Sangyun Shin, Niki Trigoni, and Andrew Markham. Spear: Receiver-to-receiver acoustic neural warping field, 2024.

[20] Jun-Hyeok Heo, Deok-Ki Kim, and B. D. Lim. Application of minimum phase condition to the acoustic reflection coefficient measurement. *Transactions*

of The Korean Society for Noise and Vibration Engineering, 15:1131–1136, 2005.

[21] Csaba Huszty, Bottyán Németh, Peter Baranyi, and Fülöp Augusztinovicz. Measurement-based fuzzy interpolation of room impulse responses. *The Journal of the Acoustical Society of America*, 123:3771, 06 2008.

[22] Bill Kapralos, Michael Jenkin, and Evangelos Milios. Sonel mapping: A probabilistic acoustical modeling method. *Building Acoustics*, 15, 12 2008.

[23] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *ArXiv*, abs/2302.02088, 2023.

[24] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *ArXiv*, abs/2309.15977, 2023.

[25] Andrew Luo, Yilun Du, Michael J. Tarr, Joshua B. Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *ArXiv*, abs/2204.00628, 2022.

[26] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics, 2022.

[27] Thomas McKenzie, Leo McCormack, and Christoph Hold. Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis, 2021.

[28] Mélanie Nolan, Marco Berzborn, and Efren Fernandez-Grande. Isotropy in decaying reverberant sound fieldsa). *The Journal of the Acoustical Society of America*, 148(2):1077–1088, 08 2020.

[29] Felipe Otondo and Jens Rindel. The influence of the directivity of musical instruments in a room. *Acta Acustica united with Acustica*, 90:1178–1184, 11 2004.

[30] Beth Paxton. Room acoustics, sixth ed., heinrich kuttruff. crc press (2017). isbn: 978-1-4822-6043-4. *Applied Acoustics*, 126, 11 2017.

[31] Boaz Rafaely. *Springer Topics in Signal Processing*, volume 8. 01 2015.

[32] Nikunj Raghuvanshi and John Snyder. Parametric directional coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2018*, 37(4), September 2018.

[33] Matthew Rosen, Keith Godin, and Nikunj Raghuvanshi. Interactive sound propagation for dynamic scenes using 2d wave simulation. *Computer Graphics Forum (Symposium on Computer Animation)*, 39(8), September 2020.

[34] Lauri Savioja and U. Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 08 2015.

[35] Julius O. Smith. *Spectral Audio Signal Processing*. accessed <date>. online book, 2011 edition.

[36] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching, 2023.

[37] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. In *Neural Information Processing Systems*, 2022.

[38] Haohai Sun, Shefeng Yan, U. Peter Svensson, Audun Solvang, and Johan L. Nielsen. Achievable maximum-directivity beamforming for spherical microphone arrays with random array errors. In *2010 18th European Signal Processing Conference*, pages 1929–1933, 2010.

[39] Ivan Jelev Tashev. *Sound and Sound Capturing Devices*, pages 82–84. 2009.

[40] Lei Wang and Jie Zhu. Flexible beampattern design algorithm for spherical microphone arrays. *IEEE Access*, 7:139488–139498, 2019.

[41] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere, 2024.

[42] Ning Xiang and Christopher Landschoot. Bayesian inference for acoustic direction of arrival analysis using spherical harmonics. *Entropy*, 21:579, 06 2019.

[43] Shefeng Yan. *Modal Beamforming for Spherical Arrays*, pages 255–297. Springer Singapore, Singapore, 2019.

[44] Franz Zotter. *Analysis and Synthesis of Sound-Radiation with Spherical Arrays*. PhD thesis, 01 2009.

[45] Franz Zotter and Matthias Frank. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. 01 2019.